DISTRIBUTION STATEMENT A (Approved for Public Release, Distribution Unlimited)

SUPPORTING INFORMATION Interpretable models for extrapolation in scientific machine learning

Eric S. Muckley¹, James E. Saal¹, Bryce Meredig¹, Christopher S. Roper², John H. Martin²



¹Citrine Informatics, Redwood City, 94063, CA, USA ²HRL Laboratories, Malibu, 90265, CA, USA

Figure S1. Histograms of the target variable distribution of each test dataset, which include materials datasets featurized using Magpie (top row, green), materials datasets without Magpie features (middle row, blue), and non-materials datasets (bottom row, red). Vertical dashed lines show the median target value of each dataset along with its standard deviation.

		tra	ain	validation			test			
melting temp										
*		17 10 10 10 10 10 10 10 10 10 10 10 10 10				100 A	Sec. 1		and	
bulk modulus		* *		· · · · · · · · · · · · · · · · · · ·	*		•		**************************************	
band gap		200 B 400 B	23 5 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1	100 Ph 20 Ph	12 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5	12 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5	23	100 100 100 100 100 100 100 100 100 100	135 843 41 135 843 41 136 843 41	
heat capacity	de la compañía de la comp									
compressive stre	ength									
formation energy	3	REAL	R	****	RIN	REAL	1	REAL	R	
fish weight		8		2	2	p	2	2	2	
Tary in Ta	and in the second	Torial Maria	To the species	The state of the state	Torisk sports	Tor is good	Torial Parts	Toring a part	Torright append	
airfoil sound	NY EI	SY EP A-	ALL F	SY EP	AL EL	SY El	SALL -	AL EL	A. F.	
abalone rings										
¢ (Có	Ć	C	¢,	C)	U	C)	<i< td=""><td>Ó</td></i<>	Ó	

Figure S2. Visualization of datasets in 2-dimensional t-SNE¹ space (each row corresponds to 1 dataset) showing example partitions used for LOCO CV to test model extrapolation (each column corresponds to 1 CV split). At each CV split, train (green) points were used for training regression models, validation (purple) points were used for feature engineering, and test (orange) points were used for evaluating model performance.

¹ Van der Maaten, L. and Hinton, G., 2008. Visualizing data using t-SNE. *Journal of machine learning research*, *9*(11).

		tra	ain I	validation			test		
melting temp	- *	- 6 - 10	- 5 . 15	- 6	- 5 - 14	- 6	- 6	- 5	- 5
-	10 A	*	-	-					
bulk modulus	•	•	• •	•	•	•	**************************************		
band gap	137 - 14 18 - 14 - 14 18 - 14 - 14 18 - 14 18 - 14 19	25 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5	22 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5	1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1. 1	25 5 5 5 185 5 5 5 185 5 5 5 185 5 5 185 5	25 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5	23 8 4 185 8 19 185 8 19 185 8 19 18 18 18 18 18 18 18 18 18 18 18 18 18	25 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5 5	23 5 5 5 185 5 5 5 18 8 5 5 18 8 5 18 5 18
heat capacity									
compressive s	trength								
formation ener	gy	R. S.	R	REAL	RING	REAL	R. S.	REAL	RENT
fish weight	The state	The state	To set of the	The state of the s	The spect	The set speech	To set spece	The state of the state	The start of the second
airfoil sound	ALL!	N'EL	NY EL	ALL L	NY EL	NE L	AN EL	NY EL	AVE!
abalone rings	Ċ	Ć	Ø	Ø	Ø	Ø	C	Ø	U

Figure S3. Visualization of datasets in 2-dimensional t-SNE space (each row corresponds to 1 dataset) showing example partitions used for random CV to test model interpolation (each column corresponds to 1 CV split). At each CV split, train (green) points were used for training regression models, validation (purple) points were used for feature engineering, and test (orange) points were used for evaluating model performance.



Figure S4. Visualization of the best feature found at each CV split for extrapolation. Each row corresponds to a single dataset, and each column corresponds to a single LOCO CV split. Each panel represents the values of the target variable (vertical axis) plotted against the values of the best engineered feature found at that CV split. Engineered features were selected only on their correlation to the target values in the validation set (purple), while the train (green) and test (orange) points were used for training and testing of regression models.



Figure S5. Comparison of model extrapolation and interpolation performance, as quantified by median NDME for 3 different regression types (linear, random forest, and neural network, on the vertical axis), and 3 different featurization strategies: (1) original: the model was trained using only the original input features in the dataset, including those calculated using Magpie, (2) best engineered (BE): the model was trained using the single best engineered input feature for the given CV split as ranked by its r² correlation to the validation set, and (3) BE + original: the model was trained using all original input features and the single best engineered feature. Each panel corresponds to a single dataset, where the left-hand side of the panel shows median NDME for LOCO CV (extrapolation) and the right-hand side of the panel shows median NDME for random CV (interpolation). Error bars correspond to standard deviation in NDME across all train-test splits for a given model configuration. Solid vertical lines correspond to the lowest model error recorded for a given extrapolation/interpolation task.

DISTRIBUTION STATEMENT A (Approved for Public Release, Distribution Unlimited)