Supporting Information for "Extrapolative Bayesian Optimization with Gaussian Process and Neural Network Ensemble Surrogate Models"

Yee-Fun Lim[1], Chee Koon Ng[1], US Vaitesswar[1], and Kedar Hippalgaonkar[1]

[1]Affiliation not available

July 20, 2021

Supporting Information

Yee-Fun Lim,* Chee Koon Ng, US Vaitesswar, Kedar Hippalgaonkar*

Contents

Figures

	Cement (component 1)(kg in a m^3 mixture)	Blast Furnace Slag (component 2)(kg in a m^3 mixture)	Fly Ash (component 3)(kg in a m^3 mixture)	Water (component 4)(kg in a m^3 mixture)	Superplasticizer (component 5)(kg in a m^3 mixture)	Coarse Aggregate (component 6)(kg in a m^3 mixture)	Fine Aggregate (component 7)(kg in a m^3 mixture)	Age (day)	Concrete compressive strength(MPa, megapascals)
0	540.0	0.0	0.0	162.0	2.5	1040.0	676.0	28	79.986111
1	540.0	0.0	0.0	162.0	2.5	1055.0	676.0	28	61.887366
2	332.5	142.5	0.0	228.0	0.0	932.0	594.0	270	40.269535
3	332.5	142.5	0.0	228.0	0.0	932.0	594.0	365	41.052780
4	198.6	132.4	0.0	192.0	0.0	978.4	825.5	360	44.296075

Figure. S1. Snapshot of the concrete compressive strength dataset, showing the inputs for composition and processing parameters and the corresponding output target of the compressive strength. There are 8 input columns in total.

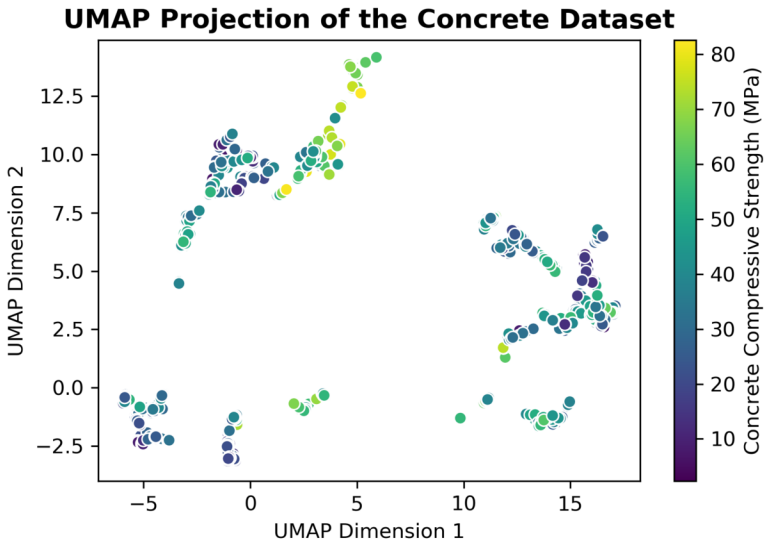

Figure. S2. UMAP projection of the concrete dataset, showing in reduced 2D space the data distribution and the spread in the concrete compressive strength values.

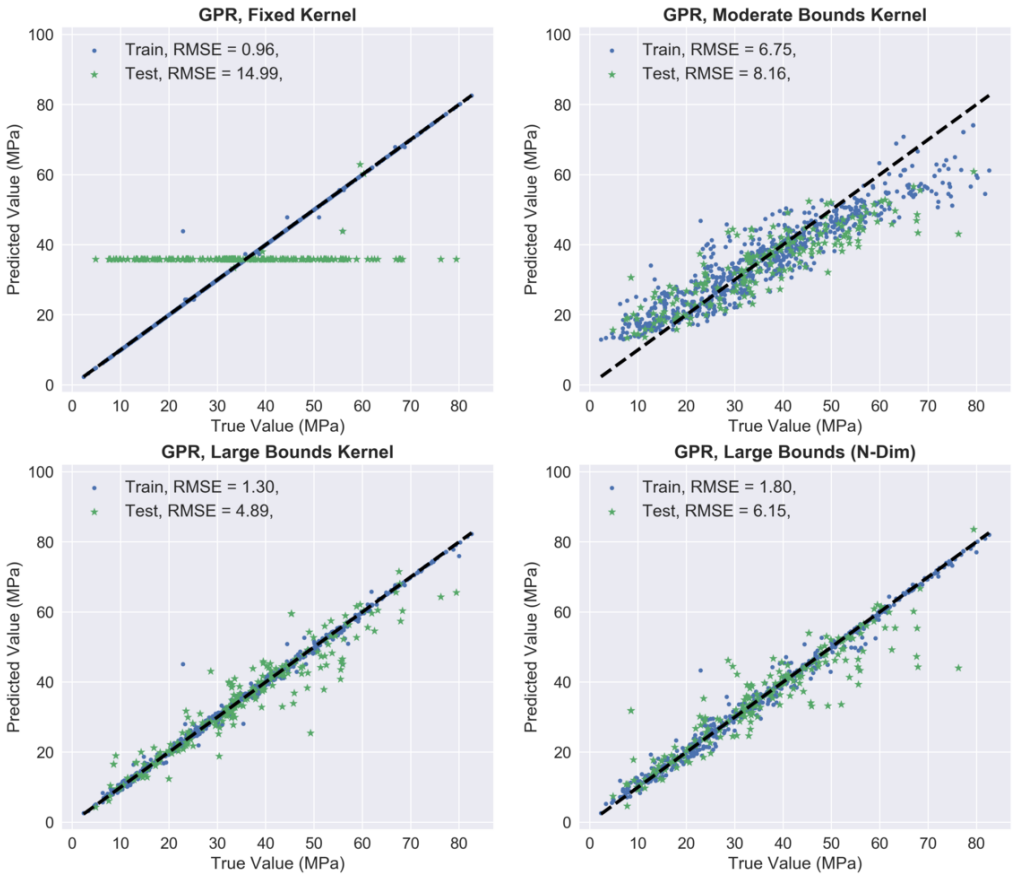

Figure. S3. Performance of the Gaussian Process Regressor on the concrete dataset with different choice of the range for the kernel bounds: fixed (top left), moderate (top right), large (bottom left), large and varying independently for each input dimension (bottom right).

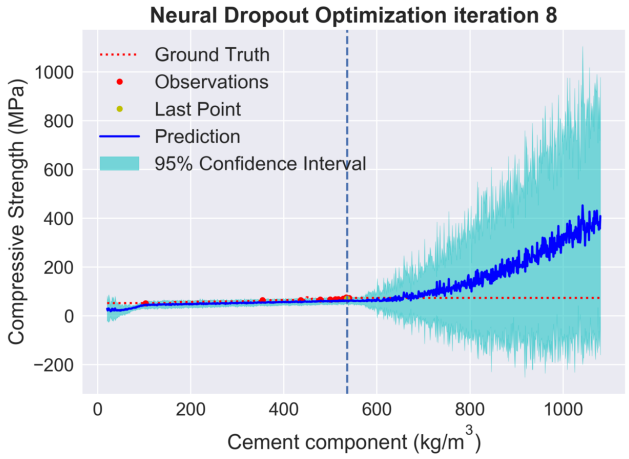

Figure. S4. Visualization of the prediction and uncertainty manifolds for the Neural Dropout surrogate model, with the range expanded beyond the bounds of the dataset. Note that within the dataset bounds (50-600), the uncertainty is relatively constant, but it starts to increase in the regions outside the bounds.

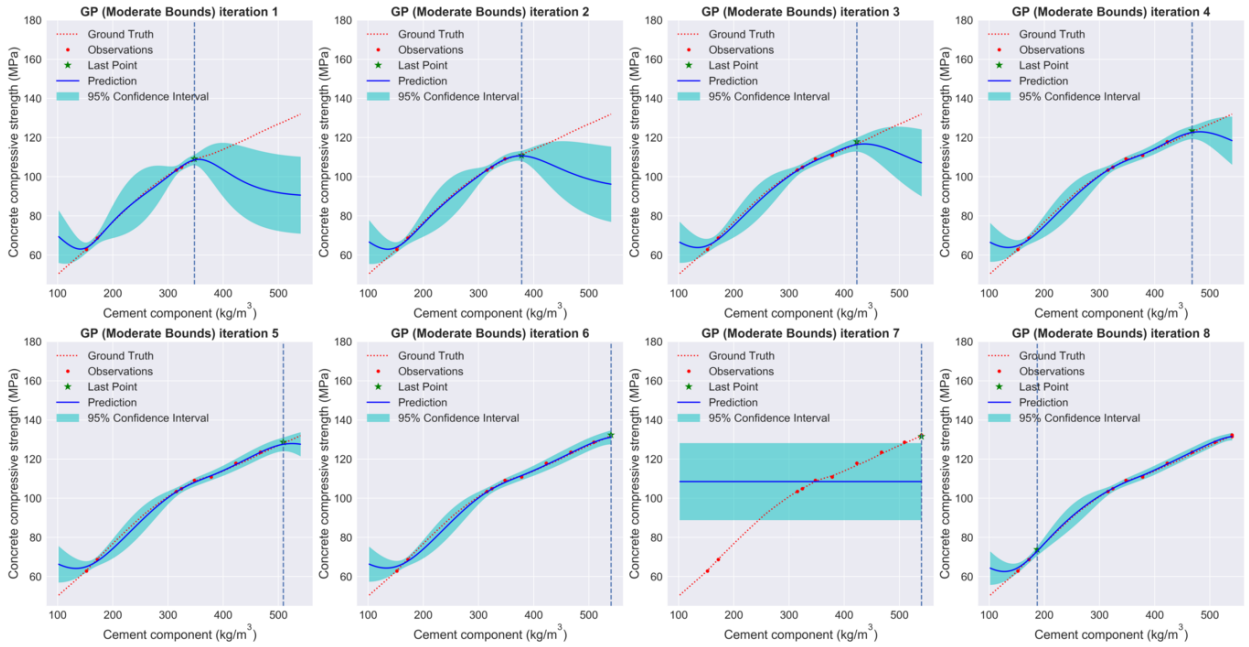

Figure. S5. Evolution of the prediction and uncertainty manifolds over the first few optimization iterations for the Gaussian Process surrogate model, with moderate bounds for the Matern kernel length scale.

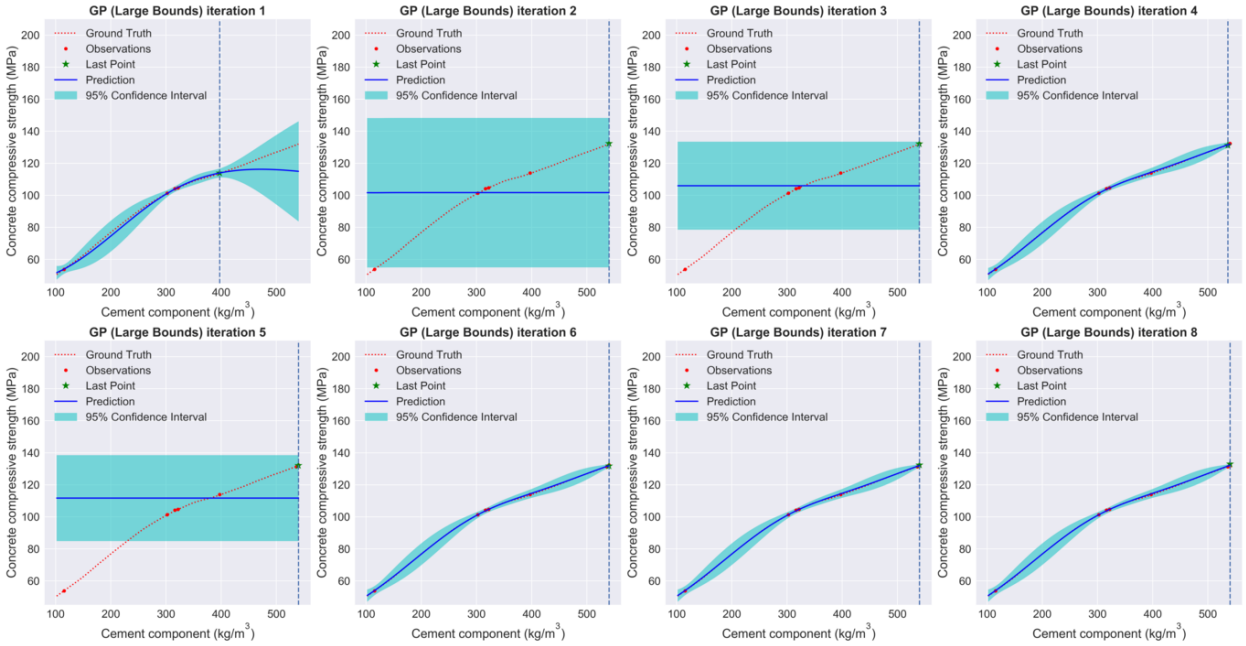

Figure. S6. Evolution of the prediction and uncertainty manifolds over the first few optimization iterations for the Gaussian Process surrogate model, with large bounds for the Matern kernel length scale.

	Ambient Temperature (°C)	Exhaust Vacuum (cm Hg)	Ambient Pressure (mbar)	Relative Humidity (%)	Energy Output (MW)
0	14.96	41.76	1024.07	73.17	463.26
1	25.18	62.96	1020.04	59.08	444.37
2	5.11	39.40	1012.16	92.14	488.56
3	20.86	57.32	1010.24	76.64	446.48
4	10.82	37.50	1009.23	96.62	473.90

Figure. S7. Snapshot of the combined cycle power plant dataset, showing the inputs for operating conditions and the corresponding output target of the energy output. This is a simple dataset with just 4 input columns.

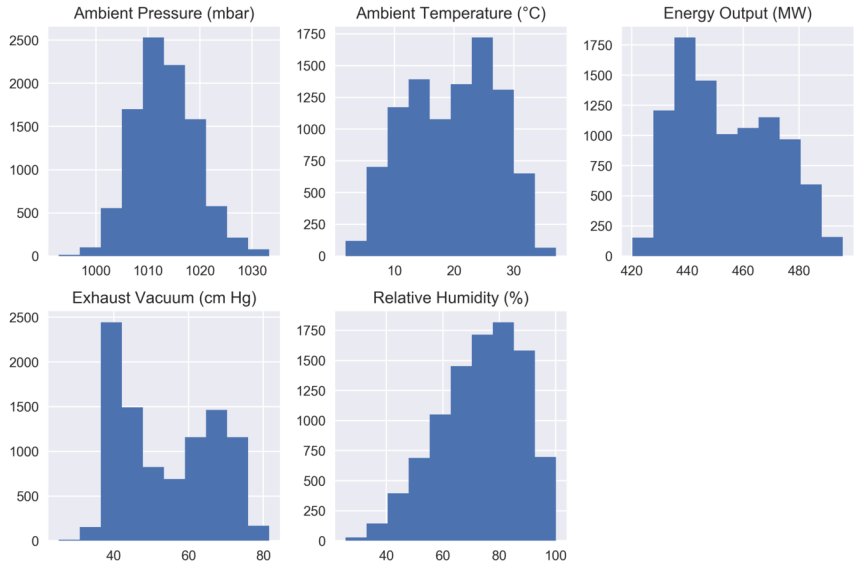

Figure. S8. Histograms illustrating the data distribution of the various inputs of the power plant dataset. In general, the data is well distributed, although the variations in the Ambient Pressure and Energy Output are small.

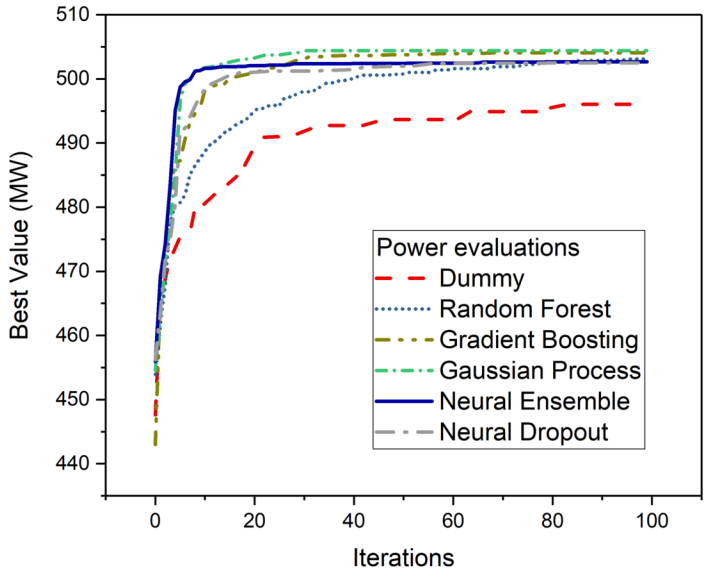

Figure. S9. Best power plant energy output evaluations averaged over 10 runs, using the trained Polynomial Regression (degree 2) algorithm as the predictive oracle and compared across different surrogate models. For this simple dataset, the Polynomial Regression turned out to be one of the better performing algorithms for extrapolative purposes. Here, there does not appear to be perceptible difference in the performance of the various models.

	range AtomicWeight	mean AtomicWeight	avg_dev AtomicWeight	range CovalentRadius	mean CovalentRadius	avg_dev CovalentRadius	range Electronegativity	mean Electronegativity
45821	136.319060	29.458364	43.147454	184	87.200000	67.440000	1.31	1.694000
65075	16.074800	15.447671	3.610808	40	78.142857	9.387755	1.14	2.737143
23106	150.064231	79.455291	49.100454	48	152.826087	14.427221	1.03	1.481739
83647	116.906052	38.726009	31.393148	178	101.666667	47.555556	2.65	2.951667
60362	48.904000	25.407667	14.773333	30	114.400000	13.226667	2.18	2.199333

	avg_dev Electronegativity	molecular_weight	Temperature	doping	s_fraction	d_fraction	p_fraction	formation_energy_per_atom
45821	0.607200	147.291820	900	1.000000e+17	0.280166	0.408264	0.311569	-0.594464
65075	0.346122	108.133700	1100	1.000000e+16	0.249139	0.750861	0.000000	-0.410709
23106	0.499055	1827.457693	300	1.000000e+17	0.272837	0.169440	0.557723	-0.271880
83647	0.720556	232.356052	800	1.000000e+18	0.119776	0.789589	0.090636	-1.206258
60362	1.024711	381.115000	500	1.000000e+18	0.278361	0.595622	0.126017	-1.893060

	e_above_hull	final_energy_per_atom	volume	density	bandgap	efermi	Power Factor	Cbrt_PF
45821	0.000000	-3.304820	64.904603	3.768359	2.2130	0.808213	0.041902	0.347331
65075	0.029398	-8.365881	298.664071	1.202423	4.2444	-1.728879	0.002759	0.140253
23106	0.000000	-3.000903	719.888592	4.215326	0.6806	3.229650	0.022410	0.281933
83647	0.011967	-4.386624	133.136311	2.898055	5.3068	-1.502903	0.139385	0.518488
60362	0.030933	-3.888959	274.947202	2.301738	0.6533	2.353798	0.044309	0.353860

Figure. S10. Snapshot of the thermoelectrics power factor dataset, showing the inputs corresponding to various material electronic properties and the output target which is the cube root of the power factor. This is a significantly more complex dataset with 22 input columns.

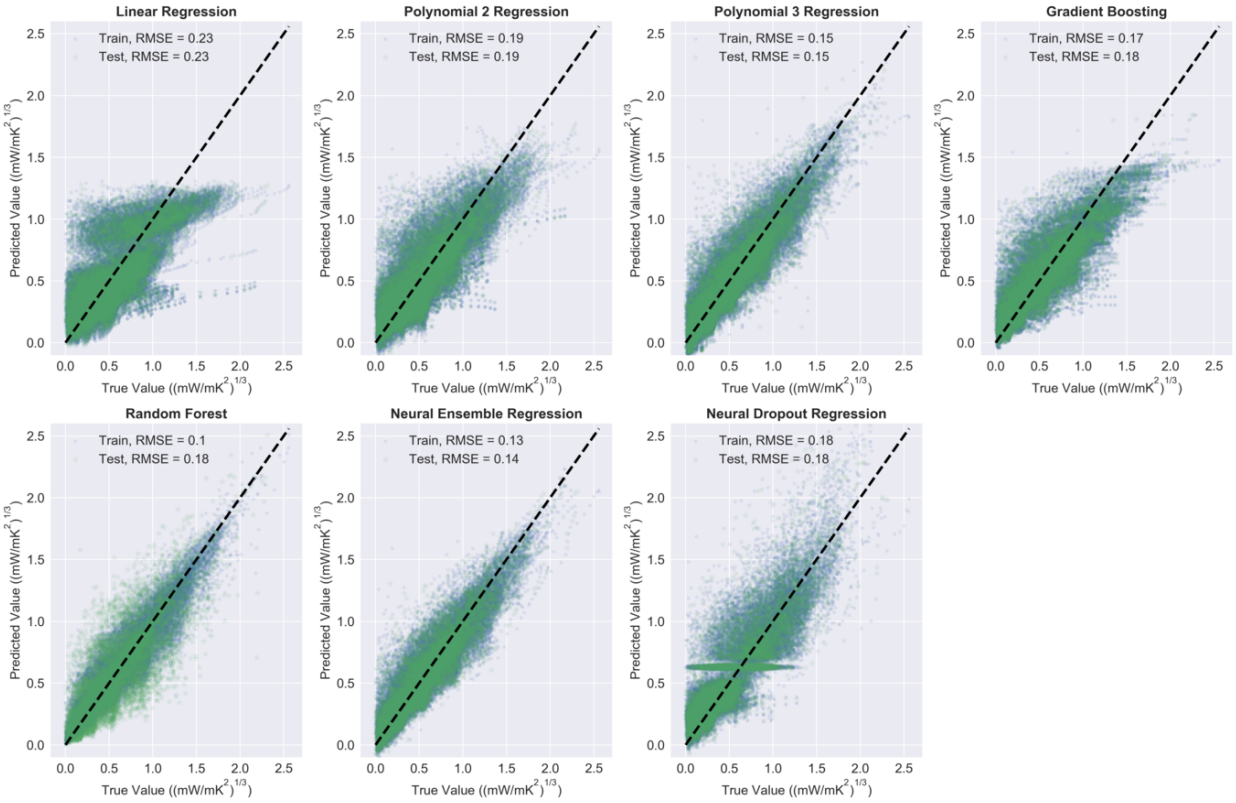

Figure. S11. Scatter plot visualizations of the actual thermoelectric power factor (cube root) vs values predicted by each of the trained regressors, with the training and test sets labeled separately. Also indicated are the RMSE scores for each regressor.

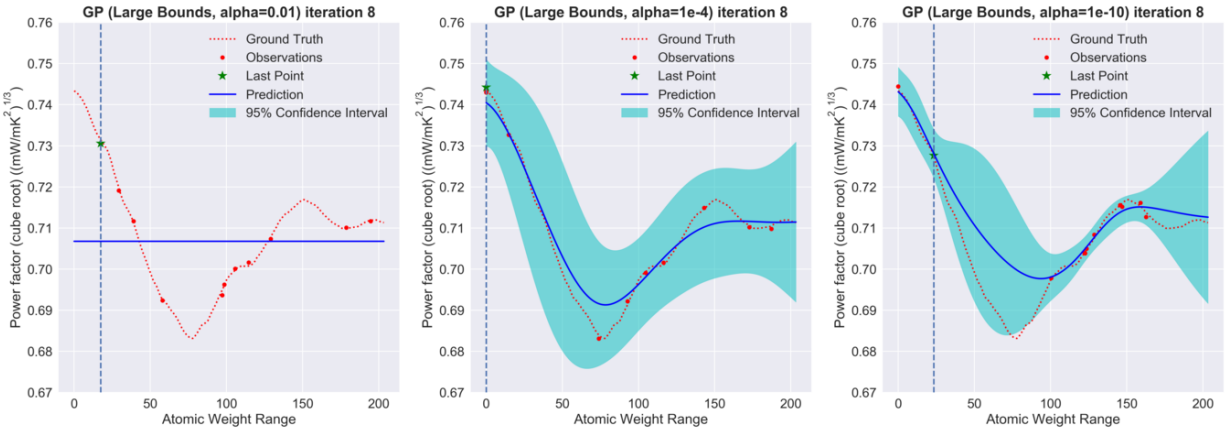

Figure. S12. Visualization of the prediction and uncertainty manifold for the Gaussian Process surrogate model applied to the thermoelectric dataset, for different values of *alpha*. The *alpha* parameter is an indication of the noise level inherent in the dataset. It should be noted that if *alpha* is set inappropriately, the

GP is unable to accurately model the given data.

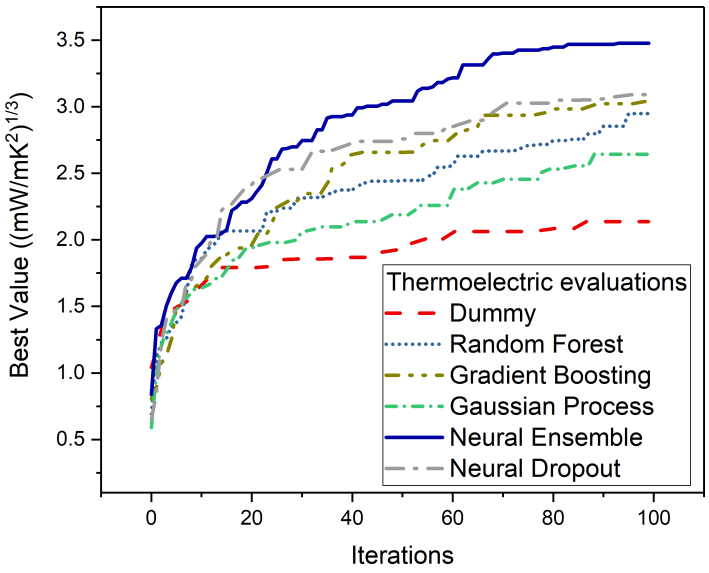

Figure. S13. Comparison of the speed of the optimization for the thermoelectric dataset with different surrogate models. Here, the Gaussian Process model parameters have not been optimized. With an alpha value of 2 (similar to that used for the concrete dataset), the GP model can barely outperform random search, whereas a tuned GP model can be the top performer. These results illustrate the importance of proper tuning of the GP parameters.

Index	**Input Name**	***Lambda* parameter**
1	Atomic Weight Range	0.73
2	Mean Atomic Weight	0.37
3	Atomic Weight Deviation	0.50
4	Covalent Radius Range	1.00
5	Mean Covalent Radius	0.72
6	Covalent Radius Deviation	0.88
7	Electronegativity Range	1.14
8	Mean Electronegativity	1.30
9	Electronegativity Deviation	0.98
10	Molecular Weight	0.01
11	Temperature	1.00
12	Doping	-1.74
13	S Fraction	0.33
14	D Fraction	1.16
15	P Fraction	0.72
16	Formation Energy Per Atom	1.12

17	Energy Above Hull	-2.33
18	Final Energy Per Atom	1.08
19	Volume	-0.70
20	Density	0.47
21	Bandgap	0.52
22	Fermi Level	1.02

Table S1. Fitted Yeo-Johnson *Lambda* parameters for the thermoelectric dataset inputs

Figures with Data and Code

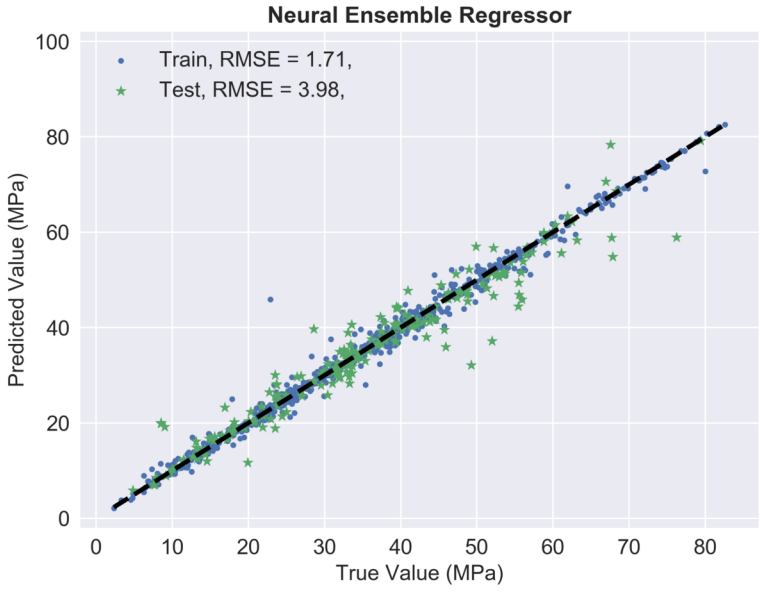

Figure 1: Concrete dataset machine learning compressive strength predictions using Neural Ensemble Regressor

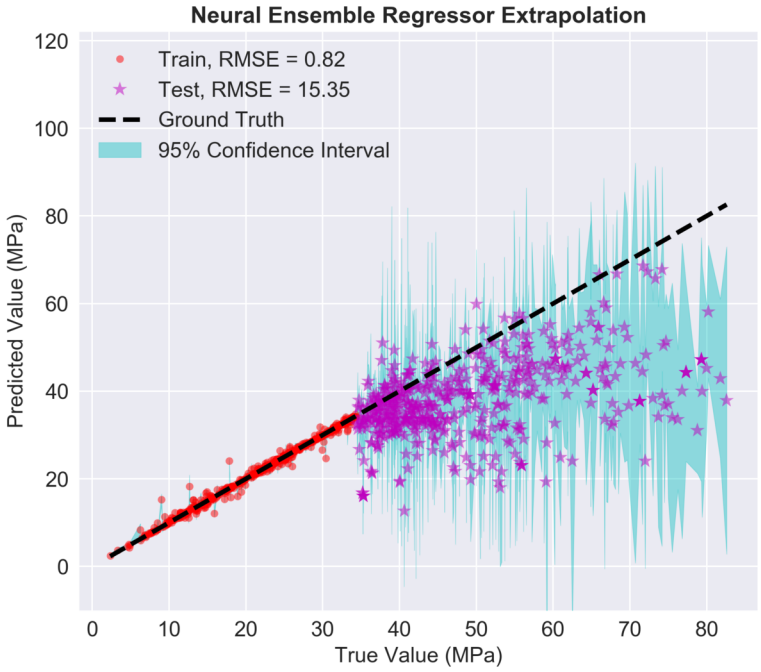

Figure 2: Visualizing the extrapolative performance of the Neural Ensemble Regressor

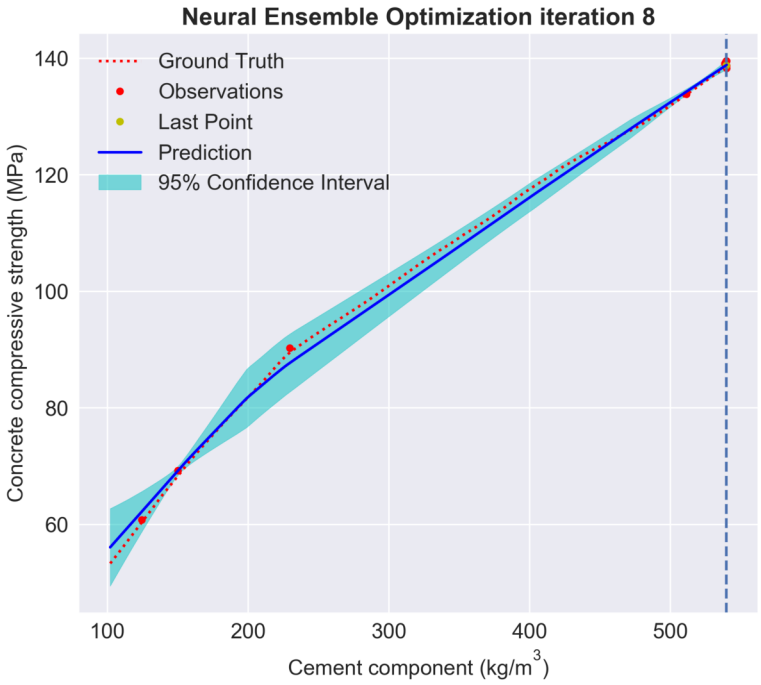

Figure 3: Visualizing the concrete Bayesian Optimization in 1D using the Neural Ensemble as the surrogate model

Data

Concrete compressive strength dataset

Hosted file

`Concrete_Data.xls` available at https://authorea.com/users/418278/articles/525086-supporting-information-for-extrapolative-bayesian-optimization-with-gaussian-process-and-neural-network-ensemble-surrogate-models

Thermoelectrics power factor dataset

Hosted file

`cubic.xlsx` available at https://authorea.com/users/418278/articles/525086-supporting-information-for-extrapolative-bayesian-optimization-with-gaussian-process-and-neural-network-ensemble-surrogate-models

Power plant output dataset

Hosted file

`Power Plant.xlsx` available at https://authorea.com/users/418278/articles/525086-supporting-information-for-extrapolative-bayesian-optimization-with-gaussian-process-and-neural-

network-ensemble-surrogate-models

Code

Concrete optimization (Jupyter ipython notebook)

Hosted file

`Concrete Optimization.ipynb` available at https://authorea.com/users/418278/articles/525086-supporting-information-for-extrapolative-bayesian-optimization-with-gaussian-process-and-neural-network-ensemble-surrogate-models

Thermoelectrics optimization (Jupyter ipython notebook)

Hosted file

`Thermoelectric Optimization.ipynb` available at https://authorea.com/users/418278/articles/525086-supporting-information-for-extrapolative-bayesian-optimization-with-gaussian-process-and-neural-network-ensemble-surrogate-models

Power plant optimizaton (Jupyter ipython notebook)

Hosted file

`Power Optimization.ipynb` available at https://authorea.com/users/418278/articles/525086-supporting-information-for-extrapolative-bayesian-optimization-with-gaussian-process-and-neural-network-ensemble-surrogate-models

Synthetic functions optimization (Jupyter ipython notebook)

Hosted file

`SynFn Optimization.ipynb` available at https://authorea.com/users/418278/articles/525086-supporting-information-for-extrapolative-bayesian-optimization-with-gaussian-process-and-neural-network-ensemble-surrogate-models

Neural Ensemble regressor (python file)

Hosted file

`NeuralEnsemble.py` available at https://authorea.com/users/418278/articles/525086-supporting-information-for-extrapolative-bayesian-optimization-with-gaussian-process-and-neural-network-ensemble-surrogate-models

Neural Dropout regressor (python file)

Hosted file

`NeuralEnsembleDropout.py` available at https://authorea.com/users/418278/articles/525086-supporting-information-for-extrapolative-bayesian-optimization-with-gaussian-process-and-neural-network-ensemble-surrogate-models