Electronic Supplementary Material (ESI) for Nanoscale.

Designing the Ultrasonic Treatment of Nanoparticle-Dispersions via Machine Learning

*Christina Glaubitz,[a] Barbara Rothen-Rutishauser,[a] Marco Lattuada,[b] Sandor Balog,[a]*and Alke Petri-Fink[a,b]**

[a]Adolphe Merkle Institute, University of Fribourg, Chemin des Verdiers 4, 1700 Fribourg, Switzerland

[b]Chemistry Department, University of Fribourg, Chemin du Musée 9, 1700 Fribourg, Switzerland

Supporting Information

Contents

1. Particle synthesis

Table SI 1. List and weight of chemicals used for particle synthesis.

Nominal particle diameter / nm	Water / g	Ethanol / g (absolute, VWR)	Ammonia / g (25 %, Merck)	Tetraethyl orthosilicate / g
40	46.5	510	10	64.5
70	60	477	10	78.5
100	73.5	453	10	99.9
80	67	465	10	89.2

2. Particle characterization

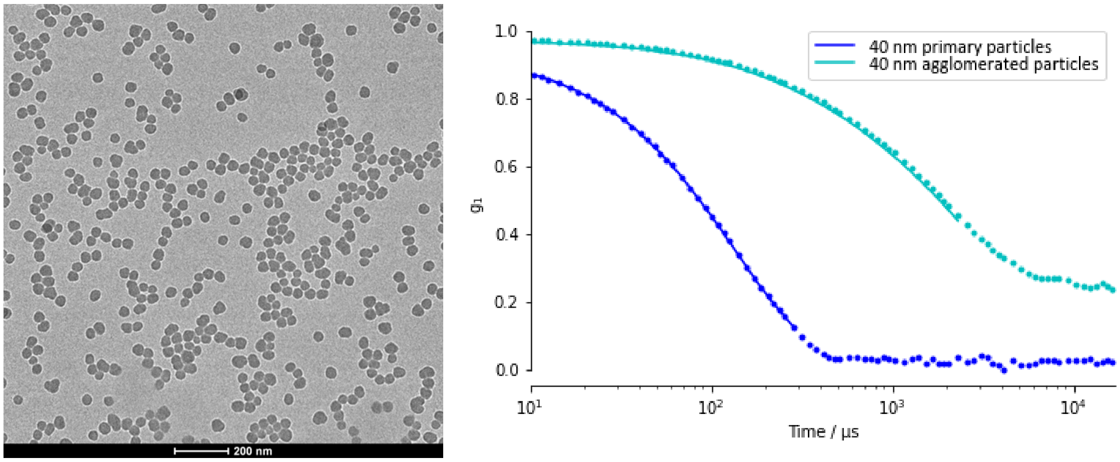

Figure SI 1. Representative transmission electron microscopy (TEM) image and dynamic light scattering (DLS) auto-correlation functions of 40 nm SiO_2 ENPs (Z-average diameter of 46 nm and a PDI of 0.05). The experimental auto-correlation functions (symbols) are analyzed with a cumulant-type nonlinear regression (solid lines).

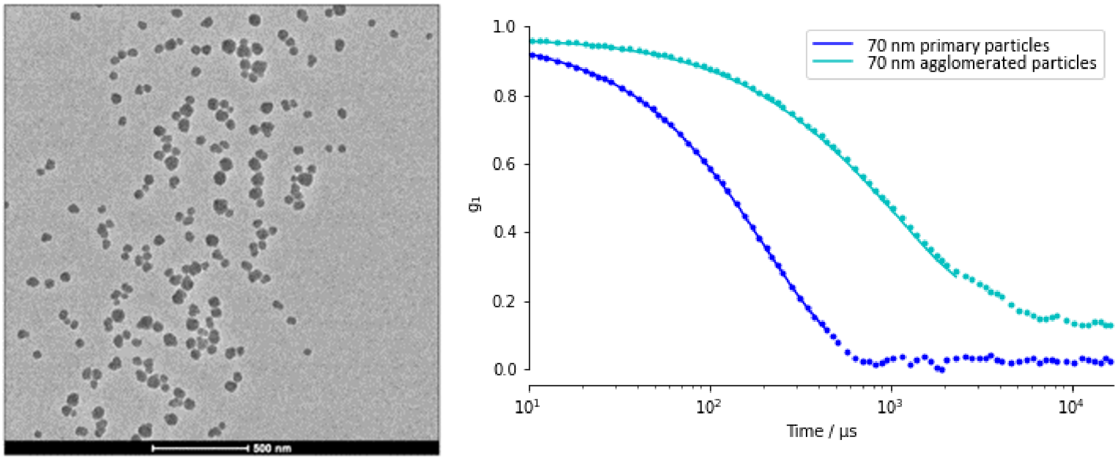

Figure SI 2. Representative TEM image and DLS auto-correlation functions of 70 nm SiO_2 ENPs (Z-average diameter of 69 nm and a PDI of 0.04). The experimental auto-correlation functions (symbols) are analyzed with a cumulant-type nonlinear regression (solid lines).

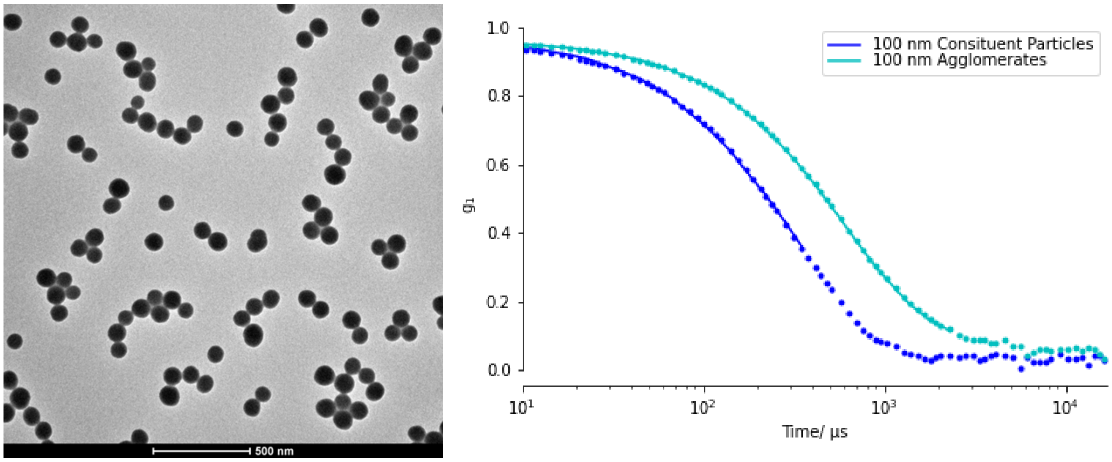

Figure SI 3. Representative TEM image and DLS auto-correlation functions of 100 nm SiO_2 ENPs (Z-average diameter of 116 nm and a PDI of 0.05). The experimental auto-correlation functions (symbols) are analyzed with a cumulant-type nonlinear regression (solid lines).

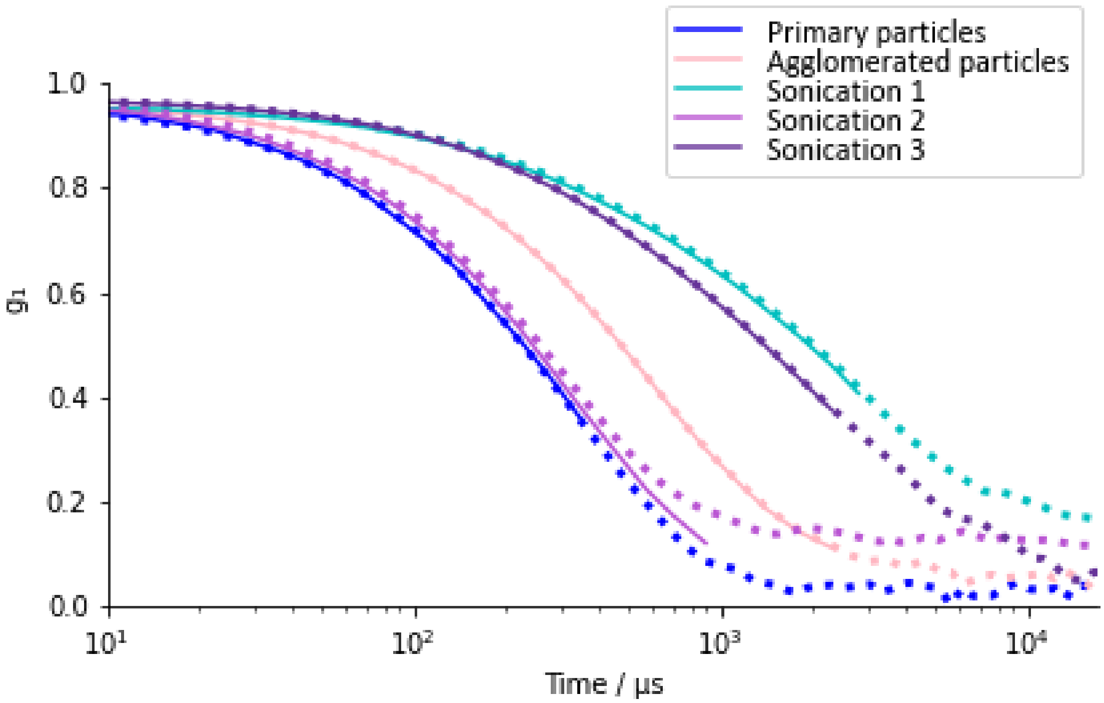

Figure SI 3. The impact of different sonication parameters on the DLS autocorrelation function of agglomerated 100 nm SiO_2 ENPs. The symbols are experimental data, and the solid lines are cumulant-type nonlinear regressions. The impact of sonication on DLS is evident. Sonication 1 (Run 1 in Table 1): Volume = 1 mL, Concentration = 1 mg mL^{-1}, Bath sonication, Amplitude = 30 %, Duration = 45 min, Energy density = 2025 J mL$^{-1.}$ Sonication 2 (Run 7 in Table 1): Volume = 5 mL, Concentration = 1 mg mL^{-1}, Probe sonication, Amplitude = 10 %, Duration = 20 min, Energy density = 715 J mL^{-1}. Sonication 3 (Run 13 in Table 1): Volume = 10 mL, Concentration = 5 mg mL^{-1}, Bath sonication, Amplitude = 30 %, Duration = 1 min, Energy density = 5 J mL^{-1}.

3. Machine learning terms

- *Categorical Data*: Categorical labels have no intrinsic order or is based on two or more categories rather than numerical values (e.g. sonicator type: “bath” or “probe”; NP Chemical formula: “SiO_2”, “TiO2”, “ZnO” or “CeO_2”).[1]
- *One-Hot Encoding*: As some ML methods cannot work with categorical data, One-Hot Encoding is needed to convert categories into numerical labels by creating a feature column for each category and using value “1” to encode the presence and “0” to encode the absence (Table SI 1).[1]

Table SI 1. Encoding categorical data (particle material) into numerical data using One-Hot-Encoding.

Material		SiO_2	TiO_2	ZnO	CeO_2
SiO_2		1	0	0	0
TiO_2	→	0	1	0	0
ZnO		0	0	1	0
CeO_2		0	0	0	1

- *Training set:* Part of the data, on which the learning algorithm is trained on.[2]
- *Validation set:* Data, which is taken from the training data set used for hyperparameter tuning. It is necessary for the training procedure to see if the model is able to predict unseen data, but to test the final predictability of a model the test data set should not be used for training processes. Therefore, the validation data is part of the training data, but

withhold in the training procedure (Cross-validation). Based on the predictability of the validation set, the models training procedure is adjusted.[2]

- *Cross-validation:* Sampling procedure that splits the training set into subsets. One of the splits is withhold during the training of the model and used as a validation set (Figure SI 5).[2]

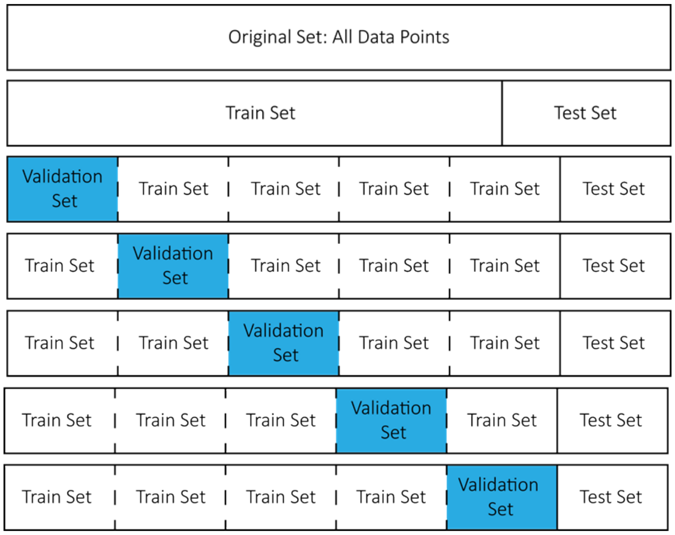

Figure SI 5. Schematic presentation of splitting data to obtain training set, validation set, and test sets, and the concept of five-fold cross validation.

- *Test set:* Part of the data, on which the performance of the algorithm is tested by comparing the measured results with the predicted results. To correctly asses the performance of the model, the testing needs to be performed on data the model has not seen in the training process.[2]

- *Model hyperparameter:* Parameters that are dedicated to configurate the algorithm and are adjusted by the operator. In tree-based models, hyper-parameters include the maximum depth of the tree, the number of trees to grow, the number of variables to consider when building each tree, the minimum number of samples on a leaf or the fraction of observations used to build a tree. There are multiple ways to determine the hyper-parameter set.[3, 4]

- *Tree-structured Parzen estimator (TPE) algorithm:* Algorithm to automatically determine the optimal hyper-parameter set. This is done by mapping a response surface on the objective function $p(y \mid x)$ of the probability of a score y (here root mean squared error) regards to a hyperparameter x using Equation SI 1.

(SI 1) $$p(y \mid x) = \frac{p(y \mid x) \cdot p(y)}{p(x)}$$

The objective function is expressed with Equation SI 2.

(SI 2) $$p(y \mid x) = \frac{p(y \mid x) \cdot p(y)}{p(x)}$$

with y^* being a certain threshold and $l(x)$ and $g(x)$ being two different distribution of hyperparameter sets. As the goal of the algorithm is to find the parameter set with the minimal score, the scores of these two functions are compared with each other. While the algorithm keeps the more successful function (so the one that outputs the lower score), the other one is replaced with a new function, and again the two scores are compared. This is repeated until the optimal parameter set is found.[3, 4]

- *Hyperparameter optimization grid:* The optimization was done with a Tree-structured Parzen estimator[5] as implemented in the Optuna-library.[6] To find the optimal parameter set for the best predictability of the model, a five-fold stratified cross validation and the mean absolute error as metric is used over this grid:

```
params = {
    # "scaler": trial.suggest_categorical("scaler", ["standard", "robust"]),

    "n_estimators": trial.suggest_int("n_estimators", 10, 500),

    "colsample_bytree": trial.suggest_uniform("colsample_bytree", 0.4, 0.9),

    "colsample_bylevel": trial.suggest_uniform("colsample_bylevel", 0.4, 0.9),

    # "min_child_weight": trial.suggest_int("min_child_weight", 0, 350),

    "subsample": trial.suggest_uniform("subsample", 0.4, 0.9),

    # "gamma": trial.suggest_uniform("gamma", 0, 1000),

    "max_depth": trial.suggest_int("max_depth", 2, 60),

  }
```

- *Model "Learnable" parameters / Node weights:* The learnable parameters are the choice of decision variables at each node and the numeric thresholds used to decide whether to take the left or right branch when generating predictive rules of a model. These parameters are determined in the training and validating process of the model (in our case Gradient Boosting Decision Tree).[2]

- *Gradient Boosting Decision Tree:* Gradient boosting means, that a cycle starts with fitting an initial model (this can be a tree or linear regression) to the data. A second model is built which focuses on predicting the cases where the first model performs poorly more accurately. The error of prediction is reduced by determining the targets outcome for this second model and changing the model's node weights based on their impacts to the prediction error. This is repeated multiple times until the error is satisfying.[2]

- *Feature Importance Analysis:* Features act next to the data as input to our model. Those features can stand alone or can be built from other features by applying feature engineering. The better the combination of features is describing the underlying physical and chemical processes, the better the insights, and therefore the predictability, of the model will be. To understand the feature hierarchy the model is following for predicting the target, one can determine the Feature Importance which can be done in multiple ways.[2]

- *SHAP (SHapley Additive exPlanations):* The SHAP method is a way to determine the order of importance of the used features by determining the Shapley-value of every possible feature permutation. The Shapley-values $\varphi_{i,j}$ are calculated using the Equation SI-3.**[7]**

(SI 3)
$$\varphi_{i,j} = \sum_{S \subseteq M \setminus i} \frac{|S|!(|M| - |S| - 1)!}{|M|!}[f(S \cup i) - f(S)]$$

Here, the difference of models' prediction with and without feature i with respect to feature j are determined with S as the subset of features that are not including feature i, $(S \cup i)$ the subsets of features in S plus feature i, S the set of all features, $f(S \cup i)$ as the model trained with all features and $f(S)$ as the model trained without feature i. To determine the global importance of features, especially for correlated features, the Shapley-values are determined for every possible feature combination. The main effect for the prediction can then be obtained as the difference between SHAP value $\emptyset_i$ and sum of SHAP interaction values for a feature using Formula SI-4.**[8, 9]**

(SI 4)
$$\varphi_{i,i} = \emptyset_i - \sum_{j \neq i} \varphi_{i,j}$$

4. Calibration of sonication energy

The delivered sonication energy was determined via calorimetry, as described elsewhere,[10] using 50 mL and 100 mL of MilliQ water. The increase of water temperature (T) was measured during sonication at timepoints of 60 s, 5 min, 10 min, 15 min, 20 min, 25 min, 30 min, 35 min, 40 min, and 45 min. All measurements were repeated three times. Given that the sonicators were thermally not isolated, part of the sonication power was lost to the environment via heat transfer. At the same time, we may safely assume that the environment behaves as a heat sink with a quasi-constant temperature (T_{env}). Therefore, the heat transfer could be described by Fourier's law, and the power balance may be expressed by a simple first order and linear ordinary differential equation:

(SI 5)
$$C_p \cdot m \cdot \frac{dT}{dt} = P_S - L \cdot (T(t) - T_{env})$$

where C_p is the specific heat of water (4.18 J/g K), m is the mass of the MilliQ water (g), and L is a loss coefficient (J s^{-1} K^{-1}).

Then the effective power is simply

(SI 6)
$$P = \frac{dT}{dt} m \, C_p,$$

and the energy released by sonication to the water is then obtained by integrating the time-dependent effective power over time. By solving Equation SI 5 we obtain the model to describe the temperature as a function of sonication time

(SI 7)
$$T(t) = T_{env} + \frac{P_S}{L}\left(1 - e^{-\frac{L}{C_p m}t}\right)$$

where from the temperature vs time curve is at hand and so is the function $\frac{dT}{dt}$, and the effective power and energy may be determined via Equation SI 6.

5. Model validation

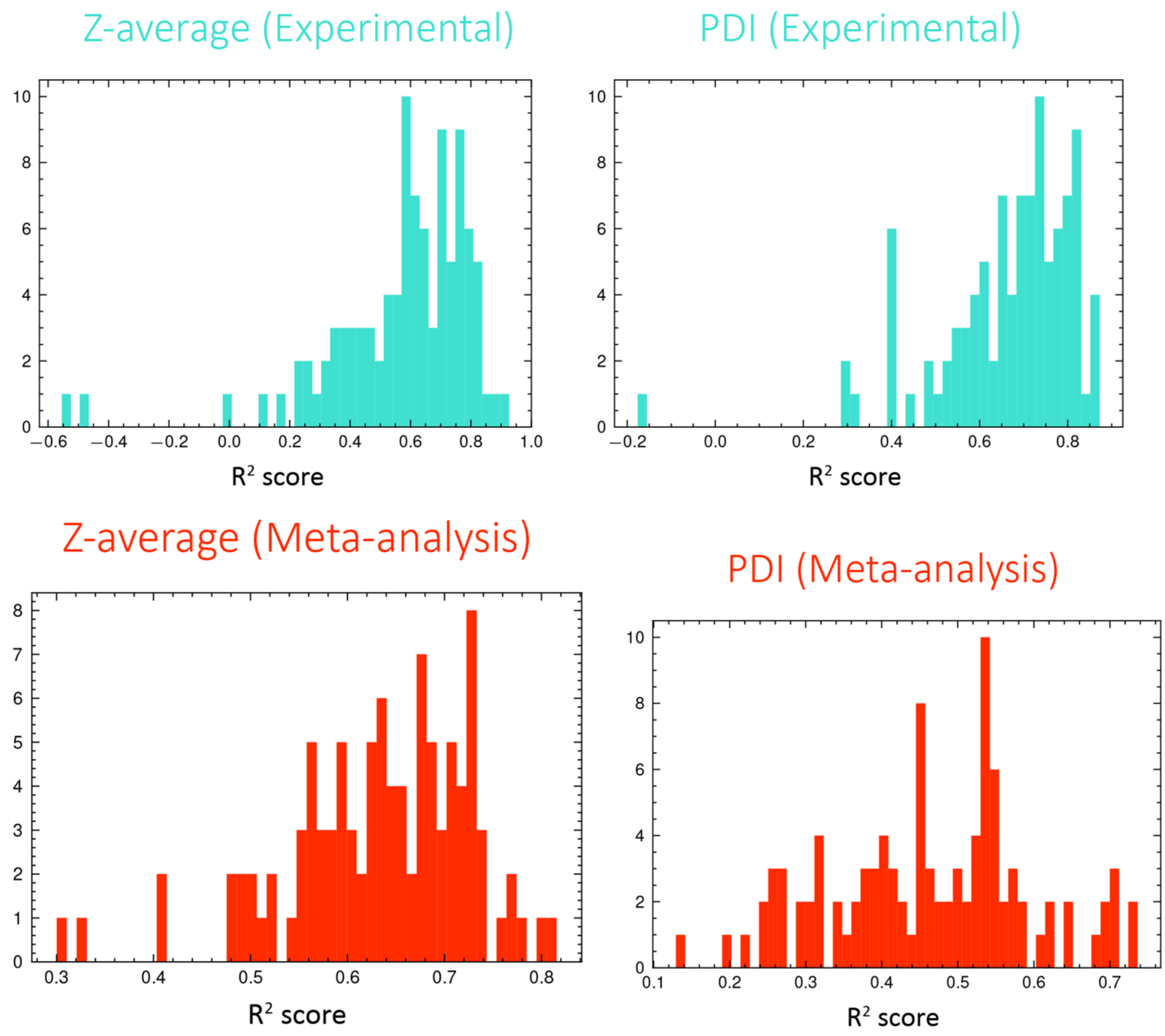

Figure SI 7. Distribution of the R^2 score for the test set of 100 newly randomly seeded and trained models. Upper left panel: Model based on lab generated data, predicting the Z-average. Upper right panel: Model based on lab generated data, predicting the PDI. Lower left panel: Model based on meta data, predicting the Z-average. Downer right panel: Model based on meta generated data, predicting the PDI.

6. Feature importance analysis

The global importance of each feature is based on its relative, mean absolute SHAP value

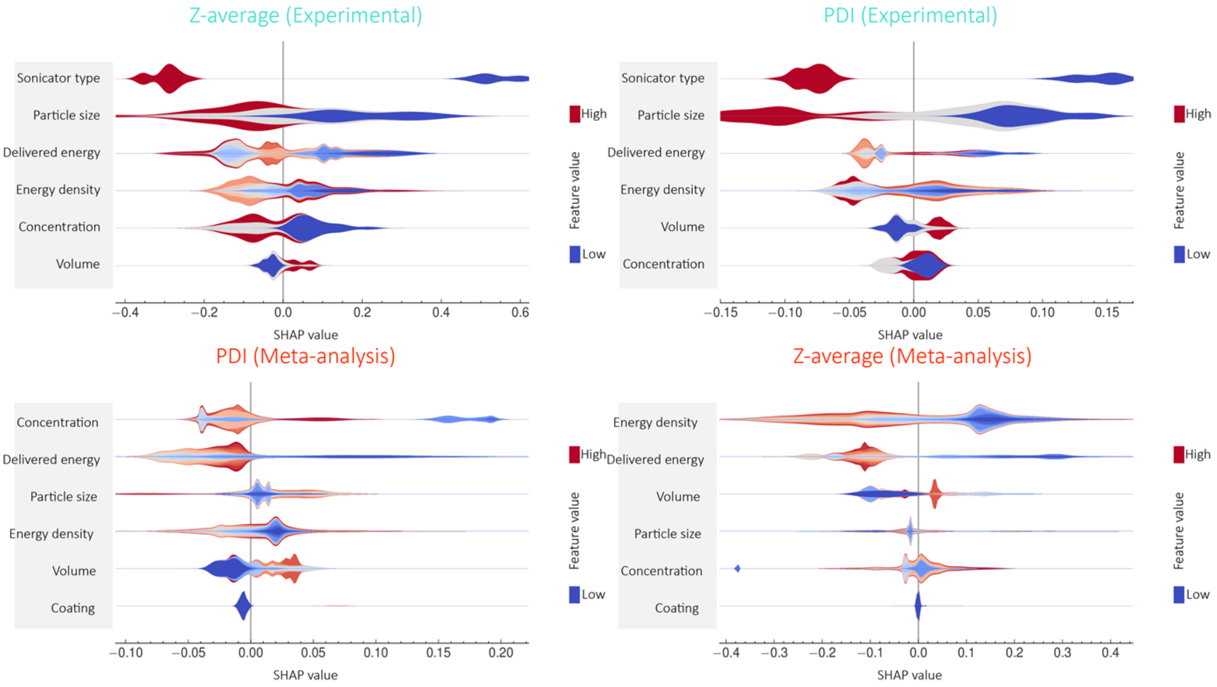

Figure SI 8. Summary of the SHAP analysis to determine the feature importance for the model predicting the Z-average based on experimental data. The violin plots show how many data points exist for a certain feature and SHAP value. The grey, vertical line indicates the baseline, so the average predicted value for all data points. A negative SHAP value (shown on the abscissa) indicates a low predicted Z-average with regards to the baseline, and a positive SHAP value a higher predicted Z-average. Upper left panel: Distribution of SHAP values for the model predicting the Z-average based on experimental data. Upper right panel: Distribution of SHAP values for the model predicting the PDI based on experimental data. Lower left panel: Distribution of SHAP values for the model predicting the Z-average based on meta-analysis. Lower right panel: Distribution of SHAP values for the model predicting the PDI based on meta-analysis.

References

1. Harris, D.; Harris, S., *Digital design and computer architecture*. 2 ed.; Morgan Kaufmann: San Francisco, 2012; p 129.

2. Jablonka, K. M.; Ongari, D.; Moosavi, S. M.; Smit, B., Big-Data Science in Porous Materials: Materials Genomics and Machine Learning. *Chemical Reviews* **2020,** *120* (16), 8066–8129.

3. Bergstra, J. S.; Bardenet, R.; Bengio, Y.; Kégl, B., Algorithms for Hyper-Parameter Optimization. *Advances in Neural Information Processing Systems* **2011,** *24*, 2546–2554.

4. Bergstra, J. S.; Yamins, D.; Cox, D. D., Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. In *Proceedings of the 30th International Conference on Machine Learning*, 2013; Vol. 28, pp 115-123.

5. Bergstra, J.; Bardenet, R.; Bengio, Y.; Kégl, B., Algorithms for Hyper-Parameter Optimization. *Advances in Neural Information Processing Systems* **2011**.

6. Takuya, A.; Shotaro, S.; Toshihiko, Y.; Takeru, O.; Masanori, K., Optuna: A Next-generation Hyperparameter Optimization Framework. *KDD (arXiv)* **2019**.

7. Lipovetsky, S.; Conklin, M., Analysis of regression in game theory approach. *Applied Stochastic Models in Business and Industry* **2001,** *17* (4), 319–330.

8. Lundberg, S.; Lee, S.-I., A Unified Approach to Interpreting Model Predictions. In *31st Conference on Neural Information Processing System*, Long Beach, United States, 2017.

9. Lundberg, S. M.; Erion, G. G.; Lee, S.-I., Consistent Individualized Feature Attribution for Tree Ensembles. **2019**.

10. Gilliland, D. *Standardised dispersion protocols for high priority materials groups*; NanoDefine Technical Report D2.3: Wageningen, 2016.