Sequential learning to accelerate materials discovery – App Description

An app for accelerating the experimental search for suitable materials is presented here. It can be used for method development and for exploring the configuration of Sequential Learning (SL) methods. To determine the performance of SL methods, it is common to use simulated experiments where the ground truth labels for all data points are already known. Initially, only a small fraction is provided to the SL algorithm (although more training data would be available). This is extended with one new data point from the remainder of the available data at each iteration. It is investigated which approach requires the least amount of data to achieve the goal. Approaches that require less data simply lead to faster success in laboratory practice. Thus, the goal is not to actually discover new materials using all available data, but to validate material discovery methods for scenarios where fewer labels are known (e.g., for new materials).

The app provides flexible and low-threshold access to SL methods via user interfaces.

The app, based on "Jupyter Notebooks", integrates seamlessly with the "AIIDA" workflow environment. The underlying code can be easily customized and extended. The app has intuitive and interactive user interfaces for data import and cleaning/selection, (statistical) data analysis, visualization for exploration and plausibility, AI environment and data evaluation and result visualization. This uses structured material data from CSV files.

Hands on

Upload

In the upload window of the app, the material data in CSV format can be imported via a dialog. An example file is provided in this repository. An option to set the CSV separator, the decimal separator and to delete non-numeric data is available. Additionally, lines at the beginning of the file can be skipped (e.g., header information, etc.). At the end of this process, the data is displayed to the user to allow a plausibility check. Here it can be checked quickly and easily whether decimal places are specified correctly, and all data are numeric.

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

Data Info

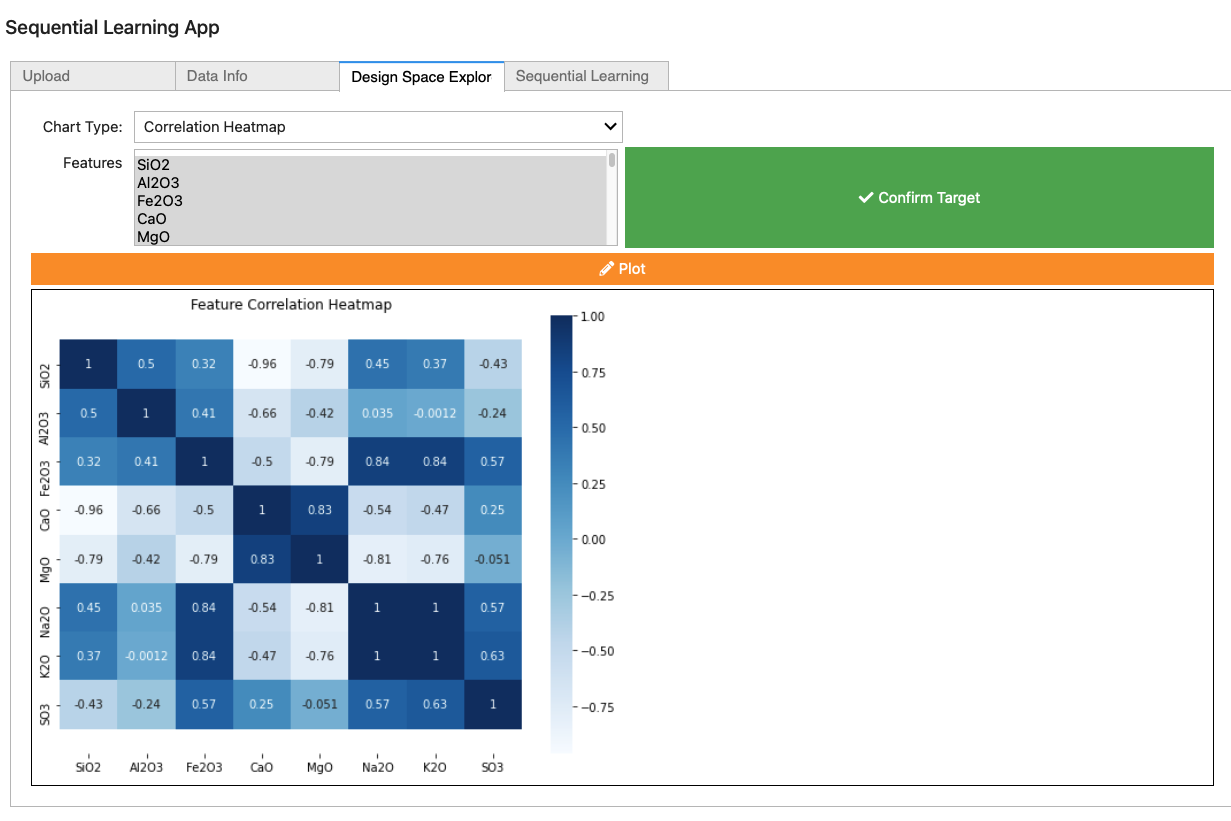
This window gives a detailed overview of the uploaded data. Besides the data preview, there is a detailed list of all variables (Info button) and some basic statistical characteristics of the variables (Stats button).

Ein Bild, das Tisch enthält.

Automatisch generierte Beschreibung

Design Space Explorer

The Design Space Explorer allows the visualization of complex relationships in the data. Here, specific dependencies between selected variables can be displayed as a scatter plot, the interrelationships and distributions of the variables can be mapped as a scatter matrix, and correlations can be visualized as a correlation heatmap. These tools allow a quick visual overview, e.g. of co-linearities of the characteristics for feature selection or trade-offs between different material properties, which are to be optimized.



Sequential Learning

This window provides a SL framework.

Settings

This window lets the user interactively set up the boundary conditions of the SL problem. The input feature and target properties can be selected simply by mouse click. It is possible to select multiple target properties (Multi-Objective Optimization). The optimization is then based on the sum (or difference - depending on whether maximization or minimization is desired) of the normalized properties.

The target can be specified as a quantile of the given properties (or their combinations in case of Multi-Objective Optimization). A lower target threshold (e.g. 90%) accelerates the SL optimization. However, this makes it increasingly difficult for SL to outperform a random process.

The sample threshold determines the restriction of the initial training data. If it was not restricted, it would be possible that the searched material is already contained in the initial data - which would make SL superfluous. The sample threshold is therefore always lower than the target threshold.

The initial sample size can be chosen below. Some SL algorithms require at least 3 samples. It is recommended to not choose less than 4 initial samples.

Sequential Learning Parameters

This tab lets the user select from several Machine Learning algorithms and utility functions. Some utility functions, such as MEID and MLID, allow to adjust further hyper parameters. The number of randomized SL runs can be set with the “# of SL runs” slider (standard value=30). The “Run” button executes a simulated experiments where the selected SL algorithms solve the optimization problem that has been specified in the “Settings” tab for the set number of SL runs.

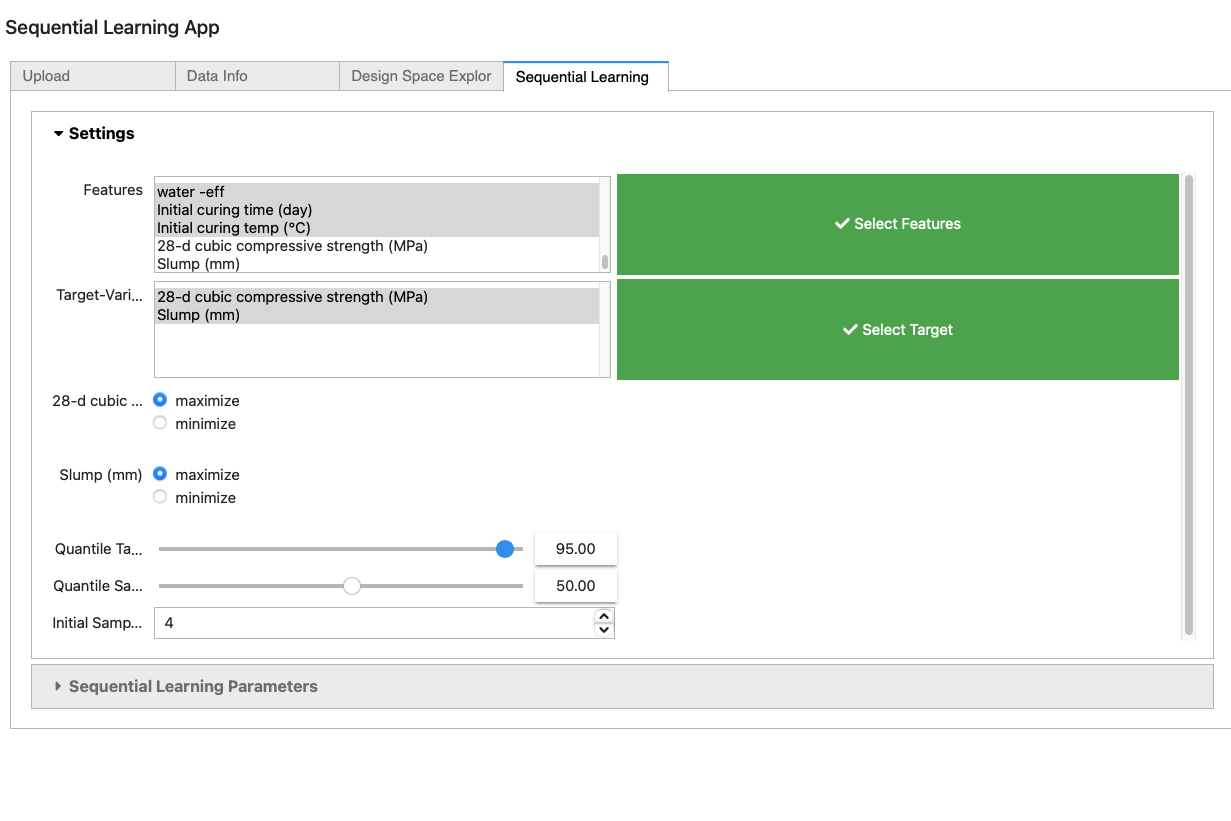
Result diagrams

The first diagram shows how fast a selected SL algorithm can find its way to the target. This is shown for each SL run as a linineplot in terms of the minimum distances in the design space from the already discovered materials to the targets per SL iterations. If the discovered materials remain far from the target solution for many iterations, a more explorative approach may help to improve performance. If it converges quickly, a more exploiting algorithm may improve performance even further.

The histogram below compares the performance in terms of experiments required of the SL algorithm VS a random process. SL is typically compared to a random process (RP) (i.e., without strategy or model) as a baseline benchmark. RPs consider each candidate as equally likely to succeed (uniform distribution). However, the success rate of RP has a nonlinear relationship with the required draws for the case of multiple targets, (The set of targets is controlled by the target threshold). A low target threshold means that RP becomes a much more difficult benchmark.

Conclusion

Serial data collection of SL, even if more successful than RP, can be detrimental in a real-world application, as waiting for experimental results could delay experimental progress. This is especially the case for materials whose synthesis is complex and whose material properties take time to develop or characterize (e.g., 28-day compressive strength of concrete). Collecting all samples at once or in batches may be more successful.

SL therefore targets material innovations for which data are not available or large-scale data collection would be too expensive. However, what actually provides an advantage in the lab depends heavily on how the SL problem is designed. The purpose of this app is to provide a tool to explore exactly under what conditions SL can help accelerate research.

SL and the closely related Bayesian optimization have repeatedly been reported to have great potential in accelerating drug and material discovery [1, 2]. The basic idea is to reduce the number of unsuccessful experiments (i.e., that lead to materials with unwanted properties) so that an ideal sequence of successive experiments is achieved. This is accomplished by coupling a prediction model (e.g., a machine learning model) with a decision-making rule based on a so-called utility function that guides the experimental program.



Figure 1: Schematic of sampling for discovering high performance materials (pink areas) with data-based methods; Left: (nearly) identically distributed sampling (black dots) as aspired in classic Machine Learning; right: sampling with Sequential Learning with initial data set (black dots) and sequential sampling (red arrows) of new samples (white dots).

Figure 1 illustrates the conceptual differences between classic Machine Learning (ML) (Figure 1, left) and the novel SL approach (Figure 1, right). The figure shows a mathematical space spanned by two base materials X1 and by X2 (in the real scenario, additional features maybe considered). Depending on the respective mixing ratio, the synthesis of X1 and X2 results in materials whose resulting material properties are shown in color (turquoise areas for low performance and pink areas for high performance materials).

ML and SL model this relationship using sampled data (black and white dots in Figure 1, respectively). These models can be employed to predict ideal combinations of X1 and X2. Classical ML approaches accomplish this (mostly) through interpolation, whereas SL is inherently extrapolative and thus potentially requires much fewer sample data. However, serial data collection of SL, even if successful, may be disadvantageous, because waiting for experimental results could delay experimental progress. This is especially the case for materials whose synthesis is complex and whose material properties require time to develop or characterize (e.g., 28-day compressive strength of binders). Collecting all samples at once or in batches can be more successful. SL could therefore fill a gap in material innovation where data are not available or large-scale data collection would be too expensive.

The scope of this app is to provide a tool to investigate exactly under which conditions SL can contribute to accelerating research regarding the properties of Materials. For this, data from the Lab can be imported to the app.

Based on this data, this app allows to investigate how the SL-approach, the quantitative research objective, and the desired success rate influence the performance of SL and deduce some of the critical circumstances under which SL can potentially enhance AAB research practice. We introduce a novel utility function that adapts common utility functions for applications with minimal training data i.e., lower number of experiments to reach the optimal sample design.

## **Structure of the Documentation**

We then briefly describe the SL algorithms which were used in this work. We use Gaussian process regression (GPR) – not yet implemented, decision tree regression (DT), tree ensemble regression (TE), random forest regression (RF) and four different utility functions, namely maximum expected improvement (MEI), maximum likelihood of improvement (MLI), maximum uncertainty (MU) and introduce a novel combination of MEI & MLI with a maximized distance measure, respectively. We go on to describe how we benchmark the acceleration in the experimental program.

# **Theoretical background and methods**

Despite success in other experimental sciences such as drug discovery, relatively few publications of SL exist in materials science. Lookman et al. [1] give an overview of the SL landscape in materials science. Despite the fact that SL regularly starts with a machine learning model based on only a few very high-dimensional data points and requires relatively little additional data from the lab, it often outperforms baseline benchmarks. However, these benchmarks are mostly statistically reached and do not necessarily mean that SL will also accelerate research in practice. Here, experimental designs often determine the speed and SL would possibly entail an adjustment of the entire process. Yet even in simple statistically motivated scenarios, specific performance is usually highly dependent on the data and the problem, and the exact relationships are still largely unknown in [3]. Lookman et al. conclude that despite many studies that use machine learning to make predictions, feedback between ML and experiments via SL has only recently been investigated in materials science.

## **2.2 Setting up Sequential Learning for materials discovery**

The underlying idea of SL is that not all experiments are equally useful. Some experiments provide more information than others. In contrast to classical design of experiments, where (only) the experimental parameters are optimized, the potential outcomes of the experiments themselves are the decisive factor. The most promising experiments are preferred over dead-end experiments and experiments whose outcome is already known. Experimental results are used to iteratively improve the ML model with high quality data. Each new experiment is selected to maximize the amount of useful information, e.g., according to [4], using previous experiments as a guide for the next experiment.

Figure 2 provides a chart that depicts the workflow of SL. The prediction of material properties in SL is based on a list of candidate materials given by experts using their domain knowledge, see (1) in Figure 2. Materials may be of interest because they are available, cheap, known to have further desirable properties, or simply because they seem generally promising. Although the exact criteria are not specified, it is recognized that the performance of the SL for material discovery is related to the quality of the candidates [3]. The candidate materials are represented in the so-called design space (DS) - a vector space that is comparable to the feature space in classical ML approaches. In the DS, the coordinates of each material are parameterized information about raw material, (micro-) structure and processing. An initial training data set with known target properties serves as an input for the prediction model in the first round (see (2) in Figure 2).

Sequential learning

No

add new data to training data

Yes

Preparation

(6) finish

Figure 2: Workflow of sequential learning for materials discovery.

At the core of the iterative SL task is the prediction of experimental outcomes ((3) in Figure 2), weighting the expected utility and deciding which candidate to investigate next ((4) in Figure 2). The utility is commonly estimated based on the predicted material property (the closer a predicted experimental result is to the desired value, the more useful it is) and a measure of uncertainty. The latter is a key driver for discovering new relationships and the basis of experimenting in general. As Reyes et al. [2] aptly summarize it, "Actually, the outcome of an experiment is the deviation from what we expected." In other words, if the outcome of an experiment is already known, there is no reason to conduct it and an experiment can be more useful if the uncertainty of its outcome is large. Lookman et al. [1] even state that they are not aware of an SL study where a new material has been discovered without utilizing uncertainties. In this sense, uncertainty can be considered an essential factor in the decision-making process. The SL task is finished as soon as the desired property is obtained, (5) and (6) in Figure 2.

The following section lays out the prediction methods, uncertainty estimates, and selection strategies utilized in this work.

## **2.2.1 Prediction methods and uncertainty estimates**

Originally, **decision trees** and tree ensembles are classification algorithms that learn the segmentation of an input data space, e.g., the DS, from pairs of data and labels [5]. By introducing one class per discrete label value (and interpolated intermediate values), pseudo-regression is performed, meaning that interpolated predictions are possible, but extrapolations outside the range of values of the label set are not. The core of tree-based algorithms is the sequential decision-making alongside the values of the respective input variables. In that sense, the data points are not considered as a "whole", but each coordinate is independently partitioned into discrete label values. By nature, this makes it relatively ill-suited to capture inter-parameter correlations. However, this can be advantageous for high-dimensional data, where unwanted correlations (so-called co-linear behavior) often result from the limited amount of available data (as expected with material data). In addition, the set of successive decisions is limited. This can result in a prioritization of relevant DS-parameters, which helps to further reduce problem complexity.

**Ensemble trees** resample the training data - e.g., by a random draw with replacement - and train a new decision tree on each of the draws. The resulting ensemble tree is, depending on the respective algorithm, the average of the tree ensemble (so-called bagging) but can also have a more complex algorithmic nature that includes, for example, an error-weighted average of the trees (as in boosting) [6]. Ensemble learners generally reduce the influence of noisy training data on the prediction and create more refined decision rules. However, resampling requires slightly more data which could have a negative effect for very small data sets (as is to be expected in an early experimental stage). In this work of the shelf MATLAB functions “fitrtree” [7] and “fitrensemble” [8] with ten surrogate splits and 10 ensemble trees were used with standard settings.

A crucial parameter of many SL methods is the **uncertainty** of a prediction (see section SL). More precisely, the epistemic uncertainty from the potentially erroneous assumptions of a model due to incomplete information is sought. Most ML methods do not provide an estimate of this by default because they are point estimates. However, it can be calculated as the dispersion of the prediction under slightly varying boundary conditions. To this end, varying training datasets can be created by resampling (such as jackknife bootstrapping) from the original training set as in [9]. The uncertainty then corresponds to the prediction scattering of the models trained on different samples of the training.

**Gaussian Process Regression** has been introduced by Krige in the year 1951 [10] and is a probabilistic model. The core concept of GPR is to assume an underlying distribution, i.e., to treat the data as random variables. Unlike DT and TE, which learn exact values for each parameter in a function, GPR derives a probability distribution using Bayes' rule. It updates prior knowledge (in the case of GPR, a specified distribution function) with observations (the training data) to compute a joint posterior probability distribution over all possible values. It contains information from both the prior distribution and the data set. Predictions are made through the joint distribution by weighting all possible predictions with their calculated posterior distribution. The output is the point estimate at the considered point of interest, which yields an expected value and a variance. As the latter can be naturally used as a measure of uncertainty, GPR has been a popular method for sequential learning or the closely related Bayesian optimization applications [1]. The probability functions in GPR are commonly specified by a multivariate Gaussian distribution (in theory, other distribution can be used, too) - a so-called Gaussian process (GP) - which is defined by a mean and covariance function. The selection of GPs can incorporate a priori knowledge about boundary conditions, e.g., when periodicity, dependencies, various length scales or general trends are known. However, this is rather relevant for time series and location-dependent data and has no proximate applicability for the presented case. Furthermore, GPs control the smoothness of the (interpolated) predictions. We compared all GPs that are implemented in MATLABs statistics toolbox [11] and found that the exponential GP performed best in the SL task.

## **2.2.2 Strategies and utility functions**

SL executes a strategy to select the next input by prioritizing the predictions, which are weighted by a utility function. The prioritization is conducted by – depending on whether the objective is to minimize or maximize a criterion – choosing the minimum or maximum weighted value. For simplicity, only the maximization case will be considered in the remainder of the manuscript, which can be described by the equation (1),

(1)

where is the selected next candidate and corresponds to finding the maximum utility . Three general strategies can be distinguished. 1. Explorative strategies attempt to reduce model uncertainty by using utility functions that favor candidates with large prediction uncertainties. 2. Exploitative strategies tend to reinforce the current model perception by considering only the predicted values by the utility function (without considering uncertainties). 3. The third group is balancing between exploring and exploiting. Only 2. and 3. are greedy strategies and thus suitable for most material finding problems.

**Maximum Expected Improvement** **(MEI)**

The (MEI) strategy [9] purely exploits by simply selecting the next candidate according to the maximum prediction value. The utility of the prediction is simply:

(2)

where is the mean prediction of the candidate.

**Maximum Likelihood of Improvement** **(MLI)**

The MLI strategy [9] is an explore and exploit strategy. It selects the candidate with the highest likelihood to exhibit the desired target property. In the case of normally distributed prediction, the candidate with the highest 95 percent likelihood can be determined according to equation (3),

(3)

where is the 95 % quantile, is the mean prediction of the i-th candidate and is the standard deviation of the i-th candidate.

### **MEI and MLI with maximum Euclidean distance (MEI+D and MLI+D, respectively)**

At the beginning of an SL run, the predictive power of ML algorithms is relatively poor due to the small amount of training data. The data are further reduced by sampling for uncertainty estimation by DT and TE, with only a portion of the data available for each sample. This causes a situation where many candidates yield the same prediction and uncertainty value, despite the fact that their composition and processing’s are very different. Candidates that have a large average Euclidean distance to the known DS candidates differ naturally more in their design. Their choice would increase the data variability and, in turn, the predictive model's performance will be most improved. This a-priori knowledge is naturally part of GPR, such that it outputs higher uncertainties for more distant data points. The utility function can be adjusted in a similar way by choosing the value that has the largest mean distance to the known DS candidates from a given range of prediction values. The MEI+D or MLI+D utilities were estimated according to equation (4) and (5), respectively.

(4)

(5)

where is the mean Euclidian distance, are j-th coordinates of the known training data with samples and and are the DS coordinates of the candidate with a greater than 90 % quantile of the MEI or MLI utility. The MEI+D and MLI+D strategies aim at boosting the initial rounds of a SL run and hence were restricted to the 15 first iterations in the presented work. The utility was then calculated according to the MEI and MLI.

## **2.2.3 Benchmarking SL against a Random Process** **(RP)**

Although, SL is based on ML methods, classical error-based ML benchmarks typically do not apply in this context. This is because the target of SL in materials discovery is to find a candidate with - depending on the property - maximum or minimum value of a said property. In a reasonably set scenario, this goal is always achieved with zero error and is merely a matter of iterations. Although this comparison is somewhat odd from a mathematical point of view, it underscores the fact that the focus here is on the effort required to reach this threshold as a measure of performance. A common metric is the required number of experiments between (5) and (3) in Figure 2 until a set target is reached ((6) in Figure 2).

To determine the performance of SL methods, it is common to use simulated experiments where the ground truth labels for all data points are already known [1]. Initially, only a small fraction is provided to the SL algorithm (although more training data would be available). This is extended with one new data point from the remainder of the available data at each iteration. It is investigated which approach requires the least amount of data to achieve the goal ((6) in Figure 2). Approaches that require less data simply lead to faster success in laboratory practice.

Thus, the goal is not to actually discover new materials using all available data, but to validate material discovery methods for scenarios where fewer labels are known (e.g., for new materials). In this approach, the generalizability is statistically demonstrated by quantifying the performance of SL methods under randomized initial conditions and then expressing it, for example, as a mean value and standard deviation. This allows meaningful comparisons between different SL approaches. To generate randomized initial conditions in an in-progress experimental study would require significant additional effort and is unrealistic in most cases. Therefore, comparisons of performance and repeatability between different SL methods in actual material discovery are usually not possible.

This approach also differs from the classical ML approach, where generalizability is demonstrated on retained test data. However, this luxury is often not afforded in experimental science, where data are extremely limited due to costly acquisition [1].

SL is commonly compared against a Random Process (RP) (i.e., acting without a strategy and model) as a baseline benchmark. RPs consider each candidate equally likely to succeed (uniform distribution). The average number of draws necessary to find the maximum target property is 50 % of the given candidates, respectively. This is the benchmark against which SL competes.

Despite the fact that this benchmark is often surpassed by SL, a significant use of SL cannot be found in practice. One reason for this may be the significantly higher effort that is caused by the sequentialization of the experimental procedure in SL. This means that from a purely functional point of view, RP can produce the desired results faster if the parallelization of experiments is more effective. In view of this situation, it is worthwhile to include further parameters for the consideration of the usefulness of SL in practice.

The specific value of the target threshold (i.e., the property value to be exceeded) inherently affects the iteration required; the smaller , the fewer iterations are required for SL to succeed. From a practical perspective, relatively small deviations of the highest cement strengths contradict a special significance of a unique strength value as the target (especially considering the aleatory uncertainties of this value). To accelerate experimental progress, one can argue to reduce , to a value that lies in the upper quantile of strengths (e.g., ) without losing much significance of the results.

Furthermore, the aspired success rate determines the number of experiments required. The relationship is simply: the higher a desired success rate, the more experiments are needed. The performance of SL at a certain success rate can be empirically determined as the quantile of the required draws from multiple SL runs. In the laboratory practice, the required success rate is expected to be much higher than the 50 % rate, which is, as mentioned above, the typical benchmark for SL.

The relationship between success rate and target threshold can be described analytically for RP as the hypergeometric cumulative distribution according to the following equation:

(6)

where corresponds to the success rate, is the size of the population, is the number of items with the desired characteristic in the population, and is the number of samples drawn. The threshold of success can be defined in terms of the parameters and . According to equation (6), the success rate has a non-linear relationship with the required draws for the case of multiple targets (where ), i.e., the before mentioned rule that a 50 % success rate requires 50 % of the possible experiments holds not for those cases. Instead, much less data is required. The exact amount further depends on the size of the population where a greater leads to fewer required draws .

This means firstly that RP becomes a much tougher benchmark when can be reduced to include more successful candidates. Secondly, a fair comparison against RP must consider the maximum available size of the population . For example, if the DS is fragmented into several smaller DSs to parallelize SL, the RP performance must still be considered on the whole DS, as its parallelization are independent of the segmentation. In other words, benchmarks can only be compared amongst same size DS and the perception of the performance compared is skewed favorably to SL when smaller DSs are used.

* 1. **Functionality**

**First**, combinations of the three prediction algorithms with four utility functions and three initial training set sizes were investigated.

**Second,** the complexity of the SL task was altered, with SL performed in a common DS containing the 131 data points and a segmented DS. The segmentation breaks the complex optimization problem into separate, expectedly simpler problems - potentially creating better predictability and possibly further speed-up. The segmentation was achieved with k-means clustering (with ) of the data in two-dimensional T-SNE coordinates using the standard functions in MATLAB, respectively [12, 13]. This created clusters that correspond to the three data resources (comparison Figure 3, left). In addition, the segmentation of the problems allows parallelization, since the SL runs could be executed in batches consisting of the segments.

**Third**, the initial training set size was varied between four and twelve. The minimal initial training set size of four was restricted by two sub-samplings in the TE with uncertainties algorithm (each leaving out one candidate) and a minimum of two points in a sensible regression problem.

**Fourth**, the task of the SL has been varied in terms of the success threshold between finding the absolute maximum within one DS and finding e a definend quantile of the sought after target.

**Fifth**, the benchmarking was carried out statistically in terms of a 50 % (Figures 4 and 6) and 97% (Figures 5 and 7) success rate. The former is the usual benchmark indicating the average performance and the latter indicates the robustness, which has special importance in scientific studies. The success rates were estimated from 30 SL runs with randomly sampled initial training sets. These initial samples can be restricted, e.g. to the lower half (50% quantile) of the strengths to enforce a significant improvement in the discovery process. Since extrapolation is much more difficult than interpolation for many learning algorithms, it is hoped that this will provide a more realistic understanding of SL performance. It is acknowledged that this restriction has a considerable influence on the performance of SL. However, in practice, whether the increased performance would be achieved remains largely unknown and depends on the given candidate group (see further details in [3]).

Table 2: Influencing factor on the success of SL and its variations that are investigated in this paper.

|  |  |  |
| --- | --- | --- |
|  | **Influencing factor** | **Investigated variation** |
| 1 | Prediction algorithm and  utility function | GPR, DT, TE |
| MEI, MLI, MEI+D, MLI+D |
| 2 | DS complexity | segmented DS vs. joint DS |
| 3 | Initial training set size | 4, 8, 12 |
| 4 | Target threshold | = maximum strength vs. 90% quantile strength |
| 5 | Success rate | 50% vs. 97%, i.e., performance in 15/30 and 29/30 SL runs respectively |

Figure 8 illustrates how the relationship between model complexity and model efficiency (note the non-linear scaling of the Y-axis) affect feasibility for laboratory practice. Here the characterization of AABs, on the one hand, involves considering many parameters that potentially vary widely between candidate materials (as discussed in the literature review Section 2.1). On the other hand, collecting data is time consuming and usually must be repeated for each new material. Both aspects limit the applicability of current ML models in practice. Although the specific course of the feasibility boundary depends on the respective laboratory capacity and the complexity of the underlying material, it is clear that the presented SL approach (blue cross) is much more effective than the state-of-the-art ML models (black dots).



**x**

less feasible

more feasible

**x**

Figure 8: A comparison between state-of-the-art ML models and SL for predicting cementitious material properties in terms of feasibility in laboratory practice as a function of model complexity and model efficiency.

# **Conclusions**

# **References**

|  |  |
| --- | --- |
| [1] | T. Lookman, P. . V. Balachandran, D. Xue and R. Yuan, "Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design," *NPJ Computational Materials,* pp. 2-17, 21 5 2019. |
| [2] | K. Reyes and W. B. Powell, "Optimal Learning for Sequential Decisions in Laboratory Experimentation," *SIAM Review (preprint),* pp. 1-47; DOI: arXiv:2004.05417 , 14 April 2020. |
| [3] | Y. Kim, E. Kim, E. Antono, B. Meredig and J. Ling, "Machine-learned metrics for predicting the likelihood of success in materials discovery," *npj Computational Materials ,* vol. 6, no. 131, 2020. |
| [4] | D. V. Lindley, "On a measure of the information provided by an experiment," *Ann. Math. Stat. 27,* p. 986–1005; https://doi.org/10.1214/aoms/1177728069, 1956. |
| [5] | R. J. Quinlan, Induction of Decision Trees, Boston, USA: Kluwer Academic Publishers, 1986. |
| [6] | D. Opitz and D. Maclin, "Popular Ensemble Methods: An Empirical Study," *Journal of Artificial Inteligence Research,* p. DOI: https://doi.org/10.1613/jair.614, 1. August 1999. |
| [7] | Mathworks, "fitrtree - Fit binary decision tree for regression," 2021. [Online]. Available: https://www.mathworks.com/help/stats/fitrtree.html. [Accessed 29 03 2021]. |
| [8] | Mathworks, "fitrensemble - Fit ensemble of learners for regression," 2021. [Online]. Available: https://www.mathworks.com/help/stats/fitrensemble.html?searchHighlight=fitrensemble&s\_tid=srchtitle. [Accessed 29 March 2021]. |
| [9] | J. Ling, M. Hutchinson, E. Antono, S. Paradiso and B. Meredig, "High-Dimensional Materials and Process Optimization using Data-driven Experimental Design with Well-Calibrated Uncertainty Estimates," *Integrating Materials and Manufacturing Innovation 6,* pp. 207–217; DOI: https://doi.org/10.1007/s40192-017-0098-z, 05 July 2017. |
| [10] | D. G. Krige, A statistical approach to some mine valuations and allied problems at the Witwatersrand, University of Witwatersrand: Master thesis, 1951. |
| [11] | Mathworks, "Kernel (Covariance) Function Options," 26 03 2021. [Online]. Available: https://www.mathworks.com/help/stats/kernel-covariance-function-options.html. |
| [12] | Mathworks, "k-means clustering," 2021. [Online]. Available: https://www.mathworks.com/help/stats/kmeans.html. [Accessed 4 April 2021]. |
| [13] | Mathworks, "t-Distributed Stochastic Neighbor Embedding," 2021. [Online]. Available: https://www.mathworks.com/help/stats/tsne.html. [Accessed 4. April 2021]. |
| [14] | A. Ahmad, F. Farooq, P. Niewiadomski, K. Ostrowski, A. A, F. Aslam and R. Alyousef, "Prediction of Compressive Strength of Fly Ash Based Concrete Using Individual and Ensemble Algorithm," *Materials (Basel). 14(4):794,* p. doi: 10.3390/ma14040794, 8. Feb. 2021. |
| [15] | A. Bagheri, A. Nazari and J. Sanjayan, "The use of machine learning in boron-based geopolymers: Function approximation of compressive strength by ANN and GP," *Measurement,* p. Measurement 141 (2019) 241–249; DOI: https://doi.org/10.1016/j.measurement.2019.03.001, 18 April 2019. |
| [16] | E. Gomaa, T. Han, M. ElGawady, J. Huang and A. Kumar, "Machine learning to predict properties of fresh and hardened alkali-activated concrete," *Cement and Concrete Composites,* p. 115 (2021) 103863; DOI: https://doi.org/10.1016/j.cemconcomp.2020.103863, 4. November 2020. |
| [17] | T. Xie, P. Visintin, X. Zhao and R. Gravina, "Mix design and mechanical properties of geopolymer and alkali activated concrete: Review of the state-of-the-art and the development of a new unified approach," *Construction and Building Materials,* p. 256 (2020) 119380; DOI: https://doi.org/10.1016/j.conbuildmat.2020.119380, 17. April 2020. |
| [18] | S. A. Addy and N. P. Koshla, "Computer Aided Design of Portland Cement Concrete Mixes," *Transportation Research Record (1062),* pp. 63-69, 1986. |
| [19] | H. Naderpour, A. H. Rafiean and P. Fakharian, "Compressive strength prediction of environmentally friendly concrete using artificial neural networks," *Journal of Building Engineering,* pp. 213-219; https://doi.org/10.1016/j.jobe.2018.01.007, 17 January 2018. |
| [20] | A. Kandiri, F. Sartipi and M. Kioumarsi, "Predicting Compressive Strength of Concrete Containing Recycled Aggregate Using Modified ANN with Different Optimization Algorithms," *Applied Sciences,* pp. 1-19; 11, 485. https://doi.org/10.3390/ app11020485 , 6. January 2021. |
| [21] | Y. Yu, W. Li, J. Li and T. N. Nguyen, "A novel optimised self-learning method for compressive strength prediction of high performance concrete," *Construction and Building Materials,* pp. 229-247; https://doi.org/10.1016/j.conbuildmat.2018.06.219, 3. July 2018. |
| [22] | C.-Y. Kao, C.-H. Shen, J.-C. Jan and S.-L. Hung, "A Computer-Aided Approach to Pozzolanic Concrete Mix Design," *Advances in Civil Engineering,* pp. 1-15; DOI: https://doi.org/10.1155/2018/4398017, 10 October 2018. |
| [23] | P. Ziolkowski and M. Niedostatkiewicz, "Machine Learning Techniques in Concrete Mix Design," *Materials,* pp. 1-16; 12, 1256; DOI: 10.3390/ma12081256, 17 April 2019. |
| [24] | D.-C. Feng, Z.-T. Liu, X.-D. Wang, Y. Chen, J.-Q. Chang, D.-F. Wei and Z.-M. Jiang, "Machine learning-based compressive strength prediction for concrete: An adaptive boosting approach," *Construction and Building Materials,* pp. 1-11; DOI: https://doi.org/10.1016/j.conbuildmat.2019.117000, 15 September 2020. |
| [25] | J. Lapeyre, T. Han, B. Wiles, H. Ma, J. Huang, G. Sant and A. Kumar, "Machine learning enables prompt prediction of hydration kinetics of multicomponent cementitious systems," *scientific reports,* pp. 1-16; DOI: https://doi.org/10.1038/s41598-021-83582-6, 11:3922 2021. |
| [26] | C. Moro, H. E. Fil, V. Francioso and M. Velay-Lizancos, "Influence of water-to-binder ratio on the optimum percentage of nano- TiO2 addition in terms of compressive strength of mortars: A laboratory and virtual experimental study based on ANN model," *Construction and Building Materials,* pp. 1-11; DOI: https://doi.org/10.1016/j.conbuildmat.2020.120960, 13 September 2020. |
| [27] | E. M. Golafshani, A. Behnood and M. Arashpour, "Predicting the compressive strength of normal and High-Performance Concretes using ANN and ANFIS hybridized with Grey Wolf Optimizer," *Construction and Building Materials,* pp. 1-14, 19 October 2020. |
| [28] | J. Zhang, G. Ma, Y. Huang, J. sun, F. Aslani and B. Nener, "Modelling uniaxial compressive strength of lightweight self-compacting concrete using random forest regression," *Construction and Building Materials,* pp. 713-719; DOI: 10.1016/j.conbuildmat.2019.03.189 , 14 March 2019. |
| [29] | R. M. Andrew, "Global CO2 emissions from cement production," *Earth System Science Data,* p. 195–217, 26 January 2018. |
| [30] | R. Dell, "Making the Concrete and Steel We Need Doesn’t Have to Bake the Planet," 4 March 2021. [Online]. Available: https://www.nytimes.com/2021/03/04/opinion/climate-change-infrastructure.html . [Accessed 29. March 2021]. |
| [31] | World Business Council for Sustainable Development, Cement Sustainability Initiative. , "Technology Roadmap - Low-Carbon Transition in the Cement Industry," International Energy Agency, France, 2018. |
| [32] | J. L. Provis and J. van Deventer, Alkali Activated Materials, State-of-the-Art Report, RILEM TC 224-AAM, Springer, Dordrecht, Heidelberg, New York, London, 2014. |
| [33] | R. Firdous and D. Stephan, "Effect of silica modulus on the geopolymerization activity of natural pozzolans," *Construction and Building Materials,* pp. 31-43, 219; DOI: https://doi.org/10.1016/j.conbuildmat.2019.05.161. , 31 May 2019. |
| [34] | V. Shobeiri, B. Bennett, T. Xie and P. Visintin, "A comprehensive assessment of the global warming potential of geopolymer concrete," *Journal of Cleaner Production, 297,* p. https://doi.org/10.1016/j.jclepro.2021.126669. , 15 May 2021. |
| [35] | J. S. Damtoft, J. Lukasik, D. Herfort, D. Sorrentino and E. M. Gartner, "Sustainable development and climate change initiatives," *Cement and Concrete Research 38 (2),* p. 115–127. DOI: 10.1016/j.cemconres.2007.09.008., 2008. |
| [36] | M. Guess, "It’ll cost $45 billion, but Germany proposes to eliminate coal in 19 years. Ars Technica," 28 January 2019. [Online]. Available: https://arstechnica.com/tech-policy/2019/01/itll-cost-45-billion-but-germany-proposes-to-eliminate-coal-in-19-years/?comments=1&pos. [Accessed 27 03 2021]. |
| [37] | J. G. Mathieson, M. A. Somerville, A. Deev and Jahanshahi, "Utilization of biomass as an alternative fuel in ironmaking.," in *Iron Ore, vol. 42*, Elsevier, 2015, p. 581–613. |
| [38] | O. E. Manz, "Worldwide production of coal ash and utilization in concrete and other products," *In Fuel 76 (8),* p. 691–696, 1997. |
| [39] | J. L. Provis, "Alkali-activated materials.," *In Cement and Concrete Research 114,* p. 40–48. DOI: 10.1016/j.cemconres.2017.02.009., 2018. |
| [40] | R. Firdous and D. Stephan, "Impact of the mineralogical composition of natural pozzolan on properties of resultant geopolymers," *Journal of Sustainable Cement-Based Materials. 176 (1),* p. 1–16. DOI: 10.1080/21650373.2020.1809028., 2020. |
| [41] | D. Bondar, C. J. Lynsdale, N. B. Milestone, N. Hassani and A. A. Ramezanianpour, "Effect of type, form, and dosage of activators on strength of alkali-activated natural pozzolans," *Cem. Concr. Compos. 33 (2),* 2011. |
| [42] | C. Shi, P. V. Krivenko and D. Roy, Alkali-Activated Cements and Concretes, Abingdon: Taylor & Francis Group, 2006. |
| [43] | R. Firdous and D. Stephan, "Influence of heat treatment and mechanical activation on reactivity of natural pozzolan for geopolymer synthesis," 2 – 15 August, Eindhoven, The Netherlands, 2019. |
| [44] | E. N. Kani and A. Allahverdi, "Effect of Chemical Composition on Basic Engineering Properties of Inorganic Polymeric Binder Based on Natural Pozzolan," *Ceram. – Silik. 53 (3),* p. 195–204, 2009. |
| [45] | R. Tänzer, Alkalisch aktivierte Systeme auf Basis von Hüttensandmehlen (Dissertation), Berlin: Technische Universität Berlin, 2018. |
| [46] | CEN/TC 51, DIN EN 196-6:2010-05 Methods of testing cement - Part 6: Determination of fineness; German version EN 196-6:2010, 2010. |
| [47] | R. Firdous and D. Stephan, "Factors effecting the properties of alkali activated natural pozzolan based geopolymers," Aachen Germany, pp. 100–106. , 2019. |
| [48] | ASTM C311 , Standard Test Methods for Sampling and Testing Fly Ash or Natural Pozzolans for Use in Portland-Cement Concrete, Vol 05.05 ed., Annual Book of ASTM Standards : ASTM International, 2001. |
| [49] | EN 196-6, Methods of testing cement – Part 6: Determination of fineness, 2010. |
| [50] | DIN 66126, Determination of specific surface area of disperse solids by the gas permeability technique - Blaine method, 2015. |
| [51] | DIN EN ISO 11885, Water quality - Determination of selected elements by inductively coupled plasma optical emission spectrometry (ICP-OES) (ISO 11885:2007); German version EN ISO 11885:2009, 2009. |
| [52] | Hatfield, Ashley, "National Minerals Information Center," 2021. [Online]. Available: https://www.usgs.gov/centers/nmic/cement-statistics-and-information. [Accessed 20 04 2021]. |
| [53] | EN 196-2, Methods of testing cement - Part 2: Chemical analysis of cement, 2005. |
| [54] | R. Firdous, D. Stephan and J. N. Y. Djobo, "Natural pozzolan based geopolymers: A review on mechanical, microstructural and durability characteristics. ,," *Construction and Building Materials, 190,* p. 1251–1263; DOI:10.1016/j.conbuildmat.2018.09.191, 27 September 2018. |
| [55] | R. J. Thomas and S. Peethamparan, "Stepwise regression modeling for compressive strength of alkali-activated concrete," *Construction and Building Materials,* p. 141 (2017) 315–324; DOI: http://dx.doi.org/10.1016/j.conbuildmat.2017.03.006, 1. March 2017. |
| [56] | M. Hameed, M. AlOmar, W. Baniya and M. A. AlSaadi, "Prediction of high‑strength concrete: high‑order response surface methodology modeling approach," *Engineering with Computers,* pp. DOI: https://doi.org/10.1007/s00366-021-01284-z, 06 Februar 2021. |
| [57] | Z. Mundher, R. Deo, A. Hilal, A. Abd, L. Cornejo, S. Salcedo-sanz and M. Nehdi, "Software Predicting compressive strength of lightweight foamed concrete using extreme learning machine model," *Advances in Engineering Software ,* vol. 115, pp. 112–125, https://doi.org/10.1016/j. advengsoft.2017.09.004, 2018. |
| [58] | J. Xu, X. Zhao, Y. Yu, T. Xie, G. Yang and J. Xue, "arametric sensitivity analysis and modelling of mechanical properties of normal- and high-strength recycled aggregate concrete using grey theory, multiple nonlinear regression and artificial neural networks," *Construction and Building Materials,* vol. 211, pp. 479–491, https:// doi.org/10.1016/j.conbuildmat.2019.03.234., 2019. |
| [59] | O. B. Douma, B. Boukhatem, M. Ghrici and A. Tagnit-Hamou, "Prediction of properties of self-compacting concrete containing fly ash using artificial neural network," *Neural Comput. Appl.,* vol. 28, pp. 707–718, https://doi.org/ 10.1007/s00521-016-2368-7, 2017. |
| [60] | D. Van Dao, H. Ly, S. Trinh, T. Le and B. Pham, "Artificial intelligence approaches for prediction of compressive strength of geopolymer concrete," *Materials (Basel),* vol. 12, p. https://doi.org/10.3390/ma12060983, 2019. |
| [61] | F. Deng, Y. He, S. Zhou, Y. Yu, H. Cheng and X. Wu, "Compressive strength prediction of recycled concrete based on deep learning," *Construction and Building Materials,* vol. 175, pp. 562-569, https://doi.org/10.1016/j.conbuildmat.2018.04.169., 2018. |
| [62] | G. Asteris and K. Kolovos, "Self-compacting concrete strength prediction using surrogate models," *Neural Computing and Applications ,* vol. 31, pp. 409–424, https://doi.org/10.1007/s00521-017-3007-7, 2019. |
| [63] | E. Mohammadi and A. Behnood, "Application of soft computing methods for predicting the elastic modulus of recycled aggregate concrete Input layer Output layer," *J. Clean. Prod. ,* vol. 176 , pp. 1163–1176, https://doi.org/10.1016/j. jclepro.2017.11.186, 2018. |
| [64] | A. Behnood, V. Behnood, M. Modiri and K. Esat, "Prediction of the compressive strength of normal and high-performance concretes using M5P model tree algorithm," *Constr. Build. Mater. ,* vol. 142, pp. 199–207, https://doi.org/10.1016/ j.conbuildmat.2017.03.061, 2017. |
| [65] | Q. Han, J. X. C. Gui and G. Lacidogna, "A generalized method to predict the compressive strength of high-performance concrete by improved random forest algorithm," *Constr. Build. Mater.,* vol. 226, pp. 734-742, https://doi.org/10.1016/ j.conbuildmat.2019.07.315, 2019. |
| [66] | T. Gupta, K. Patel, S. Siddique, R. Sharma and S. Chaudhary, "Prediction of mechanical properties of rubberised concrete exposed to elevated temperature using ANN," *Measurement,* vol. 147, p. https://doi.org/10.1016/j. measurement.2019.106870, 2019. |
| [67] | R. Bachir, A. Mohammed and T. Habib, "Using artificial neural networks approach to estimate compressive strength for rubberized concrete," *Period. Polytech. Civ. Eng.,* vol. 62, pp. 858–865, https://doi.org/10.3311/PPci.11928, 2018. |
| [68] | A. Behnood and E. Golafshani, "Machine learning study of the mechanical properties of concretes containing waste foundry sand,," *Constr. Build. Mater.,* vol. 243, p. https://doi.org/10.1016/j.conbuildmat.2020.118152 118152., 2020. |
| [69] | A. Gholampour, I. Mansouri and O. Kisi, "Evaluation of mechanical properties of concretes containing coarse recycled concrete aggregates using multivariate adaptive regression splines (MARS), M5 model tree (M5Tree), and least squares support vector regression (LSSVR) models," *Neural Comput & Applic,* vol. 32, pp. 295–308, https://doi.org/10.1007/s00521-018-3630-y, 2020. |
| [70] | W. B. Chaabene, M. Flah and M. L. Nehdi, "Machine learning prediction of mechanical properties of concrete: Critical review," *Construction and Building Materials,* vol. 260, p. https://doi.org/10.1016/j.conbuildmat.2020.119889, 2020. |
| [71] | L. van der Maaten and G. Hinton, "Visualizing Data Using t-SNE," *Journal of Machine Learning Research,* vol. 9, p. 2579–2605, 2008. |
| [72] | P. A. Ortega, J. X. Wang, M. Rowland, T. Genewein, Z. Kurth-Nelson, R. Pascanu, N. Heess, J. Veness, A. Pritzel, P. Sprechmann, S. M. Jayakumar, T. McGrath, K. Miller, M. Azar, I. Osband and Rabinowitz, "Meta-learning of Sequential Strategies," *CoRR,* p. DOI: http://arxiv.org/abs/1905.03030, 2019. |