Combining Machine Learning and Formal Techniques for Small Data Applications - A Framework to Explore New Structural Materials

(Invited Paper)

Rolf Drechsler*† Sebastian Huhn*† Christina Plump*

*Institute of Computer Science
University of Bremen
28359 Bremen, Germany
{drechsle,huhn,cplump}@informatik.uni-bremen.de

†Cyber-Physical Systems
DFKI GmbH
28359 Bremen, Germany

***Abstract*—The massive increase in computation power leads to a renaissance of supervised learning techniques, which were published decades ago but have so far been confined to theory. These techniques form the increasingly important field of *Machine Learning* (ML), which contributes to a large variety of research concerning industrial, automotive but also consumer applications strongly influencing our daily life. Commonly, the learning techniques require a set of labeled data, which involves a resource-intensive generation, to conduct the training. Depending on the dimensionality of the data and the required precision as needed by the application, the amount of training data varies. In case of insufficient training data, the prediction is of low-quality or not even possible at all, restricting the applicability of ML.**

This work proposes a combination of formal techniques and ML to implement a framework that allows coping with high-dimensional, training data while retaining a high prediction quality. The efficacy of this method is exemplarily demonstrated on the basis of an interdisciplinary material science research problem concerning the development of new structural materials, though it can be adapted to further applications.

I. Introduction

Machine Learning (ML) subsumes a wide range of techniques – like deep neural networks or support vector machines –, whose algorithmic foundations have mostly been laid decades ago. However, due to the limited computation power at that time, these approaches had been confined to theory. With the increase in computation power of central processing units and the availability of high-performance hardware accelerators, ML takes center stage of recent research and industry again.

ML provides powerful mechanisms to classify objects - as frequently performed in computer vision - or to approximate the output value (*prediction*) of highly complex or even unknown relations for given input data (*patterns*). In particular, ML has shaped systems as used for advanced driver assistance systems and, even more importantly, enabled the pioneering work in the field autonomous driving [1], which simply cannot be addressed by classical approaches.

The leading class of ML techniques follows a supervised learning scheme, i.e., the algorithm has to be trained initially prior to its invocation. This training requires (training) data, which consists of patterns in conjunction with the corresponding expected output value - acting as the label of the individual pattern. The generation of these labels forms a resource-intensive task since these labels have to be determined by hand or by computing-intensive simulations and, hence, this number is strictly limited.

Generally, large training data sets are required to achieve a certain prediction quality, i.e., the predicted value deviates from the exact value only within a certain range. This yields a confidence interval, which is then considered during the system's design, for instance, by introducing a certain robustness [2]. In contrast to this, several applications exist, which could take a strong advantage of ML, though these applications rarely invoke ML since they require a high prediction quality that cannot be achieved yet . Such a demanding application concerns the *Electronic Design Automation* (EDA), which orchestrates more and more ML techniques during the design flow. For instance, ML is utilized to consider non-functional aspects about the power-consumption [3], to pave the way for new debugging methodology [4] or to reveal malefic components like Hardware Trojans [5], which are all essential aspects for the next generation of circuit design.

The restricted usage of ML for these new, demanding applications is even more emphasized by steadily increasing dimensionality of the patterns while the amount of training data remains relatively the same. Consequently, it is not possible to sustain the required training rate, which unavoidably leads to a decrease in the prediction quality or makes it even impossible to predict at all. As said, supervised ML is meant to be applied whenever a large set of training data are available rather than being faced with small data.

This work proposes a new approach, which combines ML with formal techniques to improve the prediction quality and, more importantly, enable the applicability of ML within small data applications dealing with a large dimensional input space. The proposed methodology yields a data processing flow, which seamlessly integrates so-called formalized descriptions and state-of-the-art ML techniques, which represent application-

specific knowledge [6] about the targeted domain. More precisely, formal descriptions are formulated, which hold *static* and *volatile* domain-specific information about the application. *Static* information refers to, for instance, physical key facts. In contrast to this static information, *volatile* information refers to the kind of information that may be altered over time, for instance, due to new insights of the application-domain. Thus, considering the dynamic character of the latter type of information is of high relevance when modeling these formal descriptions and a mechanism is required to validate the volatile information. The proposed methodology allows for the first time to utilize ML techniques regardless of high dimensional patterns and small training data.

This paper demonstrates the valuable contribution of the proposed methodology on the basis of an interdisciplinary research question of the highest relevance about the development of new structural materials. However, the proposed techniques can be easily adapted to other domains. In particular, this work succeeds in implementing a Predictive Function considering data from evolutionary (material) testing procedures to, finally, predict resulting material properties without having the time-consuming and cost-intensive experiments be conducted.

The structure of this paper is as follows: Section II describes the background of the addressed application. Section III describes the proposed methodology and defines the formal descriptions. The implementation of the framework is briefly given in Section IV. Section V demonstrates the framework's efficacy for exemplary test procedures and, finally, Section VI concludes the paper.

II. Background

¸ Recent advantages in the field of *Computer-Aided Design* (CAD) allow designing integrated circuits of steadily increasing complexity concerning the number of transistors or processing cores. Analogously to the field of computer engineering, the major progress of CAD techniques enables to pursue completely new designs, which lead to novel constructions and breakthroughs, for instance, in automotive or aerospace applications. However, these new applications involve complex (constructional) designs, which require, among others, high-performance structural materials to tackle the arising challenges concerning strength, weight or durability – forming a requirement profile of desired material properties.

The properties of structural materials depend on the alloy composition in conjunction with the thermal and mechanical treatment [7]. Typically, a set of specific treatments is applied to adjust and, more importantly, to improve the resulting properties depending on the intended application of the material.

These treatments can be performed with a large number of different parameter sets, like temperature gradients or the cold forming forces. Due to the fact that it is neither possible to model the underlying physics adequately nor to describe the interdependence analytically, the development of new structural material still follows a trial-and-error principle [8]–[10].

The generation and evaluation of new structural material is a cost-intensive and time-consuming process. Consequently, a strictly combinational search for new structural material will not succeed with respect to the given constraints in the sense of time and costs [11]. This is even more unlikely when considering the large search-space of potential candidates as spanned by the high numbers of parameters during the initial generation and the subsequent treatments.

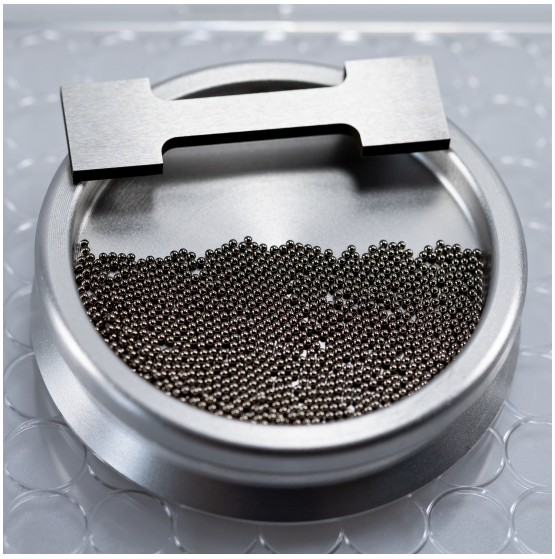

Fig. 1. Comparison between tensile test specimen and spherical micro samples (of same mass) ©Jan Rathke

A high throughput approach for exploring new structural materials has been proposed in [12] to address the shortcomings of the trial-and-error principle. This approach invokes, among others, supervised machine learning techniques and formal methods. This combination should allow approximating the material property of a given sample effectively without conducting the cost-intensive and time-consuming experiment itself. However, the invocation of supervised learning techniques requires a considerable amount of labeled data to be conducted prior to the actual *training*. The generation of this training data forms a cost-intensive task since the candidates (samples) have to be generated and treated by using the most promising parameter sets. Furthermore, the resulting sample has then to be evaluated by applying specific (standardized) test procedures like the tensile test to determine the label in the sense of supervised learning.

In work [13], a shift from the macro-level to the micro-level has been proposed to reduce the required resources for determining the required training (and validation) data significantly. At this micro-level, (micro) samples are provided as spheres with diameter sizes of $0.6 - 1.2mm$. In comparison, a standardized testing specimen of a regular tensile test holds a geometry of 12 times 2.5 cm [11]. From a mass perspective, this conventional one – the macro sample – equals roughly 2,000 spherical micro-samples when considering a frequently used commercial steel such as 100Cr6 (cf. Figure 1).

As the standardized testing procedures can not be performed on these micro samples, completely new test procedures on

the micro-level have been developed leading to so-called characteristic values [14].

Definition 1 (Characteristic Value)**.** *A characteristic value is either a measured value from a testing procedure or a determined value from a series of measured ones. Ideally, this value is independent of the parameters of the experiment. If this independence is not given, the parameters are kept fixed to ensure comparability later on. Depending on whether the value is obtained from testing a micro or macro sample, it is called a micro characteristic value or a macro characteristic value. D_μ describes the set of all micro characteristic values and, analogously, the set of the macro characteristic values is described by D_M.*

Definition 2 (Material Property)**.** *A material property is a normed value computed from the results of a standardized testing procedure. The set containing all material properties is given by WE.*

The determined characteristic values describe the structural material's behavior from a chemical, thermal or a mechanical perspective. By following this idea of down-scaling the samples' size, a Predictive Function has been introduced in [12], [15] to significantly save resources. More precisely, this function allows projecting characteristic values onto material properties by orchestrating supervised machine learning techniques. By this, it is possible to produce a large number of samples on the micro-level, evaluate these samples by newly developed methods, and compute the corresponding material properties, which makes testing much more resource-efficient.

A. Predictive Function

The Predictive Function Ψ is the central predicting component of the high throughput development approach. In particular, the precision of the prediction of material properties and the capability to cope with a small amount of training data are both critical aspects since testing material properties is a resource intensive task. Consequently, the training has to be conducted on just a few well-chosen grid points.

When comparing characteristic values and material properties, one may notice that these differ not only in size but also in the testing procedure. This circumstance further impedes the prediction since scaling effects are implicitly given but not modeled. For this reason, a two-stage measurement principle has been proposed in [13], which separates between the micro- and macro-level measurement to exclude such scaling effects during the final prediction of the material properties. Figure 3 reflects the obtained relationships of this two stage measurement.

This middle layer allows for the definition of two functions whose composition yields the Predictive Function:

Definition 3. *The Scaling Function is defined as $\Theta : D_\mu \to D_M$ and maps D_μ to D_M. Analogously, the Transfer Function is defined as $\Lambda : D_M \to WE$ and maps D_M to WE. Consequently, the Predictive Function Ψ: $D_\mu \to WE$ then fulfills $\Psi = \Lambda \circ \Theta$.*

The mentioned grid points are necessary to conduct the training of both the Scaling Function Θ and the Transfer Function Λ. Consequently, the Predictive Function Ψ requires a full set of training data (of all three layers) for each and every input dimension given by the characteristic values. Hereby, it can be ensured that all inter- as well as intradependencies are properly reflected.

Furthermore, a comprehensive search algorithm is required to close the loop of the development process of new structural materials, which is discussed next.

B. Search Algorithm

The *Search Algorithm* allows identifying the alloy compositions and subsequent treatments, which yield *promising* characteristic values being measured on the micro-level. Here, *promising* refers to the circumstance that the predictive function approximates material properties for measured characteristic values which should be as close as possible to a given requirement profile (of material properties) [12]. The efficacy of this search algorithm highly depends on the precision of the predictive function. Typically, the requirement profile acts as a starting point for this search.

Definition 4. *A requirement profile is defined as a vector $\boldsymbol{dp} \in WE$ containing the desired value for each material property. Besides this, it may contain a tolerance vector $tv \in \mathbb{R}^{dim(WE)}$ stating the length of a tolerable error margin*[1].

The search algorithm identifies one or multiple of these vectors of characteristic values, i.e., $\boldsymbol{x} \in D_\mu$, which, in turn, map to the given profile $\boldsymbol{dp}$ (given the error margin) when applying Ψ. This search is conducted by a multi-objective optimization approach, which follows state-of-the-art techniques like [16].

A high throughput approach is presented in [13], which seamlessly aggregates the recently presented techniques. The resulting data and information flow is presented in Figure 2. However, in the light of material expense, grid points can not be produced in a quantity that enables regular supervised learning techniques typically expected *Big Data* instead of *Small Data*. For this reason, new measures have to be developed and seamlessly integrated into the high throughput approach, which allow a prediction of high quality even if the amount of data – with respect to the number of input dimensions – is strictly limited.

This paper proposes a framework, which allows to utilize ML techniques by taking advantage of application-specific knowledge in terms of formal descriptions effectively even if high dimensional patterns and relatively small amount of training data are faced.

III. Proposed Methodology

As introduced in Section II, the Predictive Function acts as a central component of the high throughput approach [12]. In this context, a first data processing framework has been proposed in [15], which allows the prediction of two material

[1]Here, it is assumed that the error margin is always centralized.

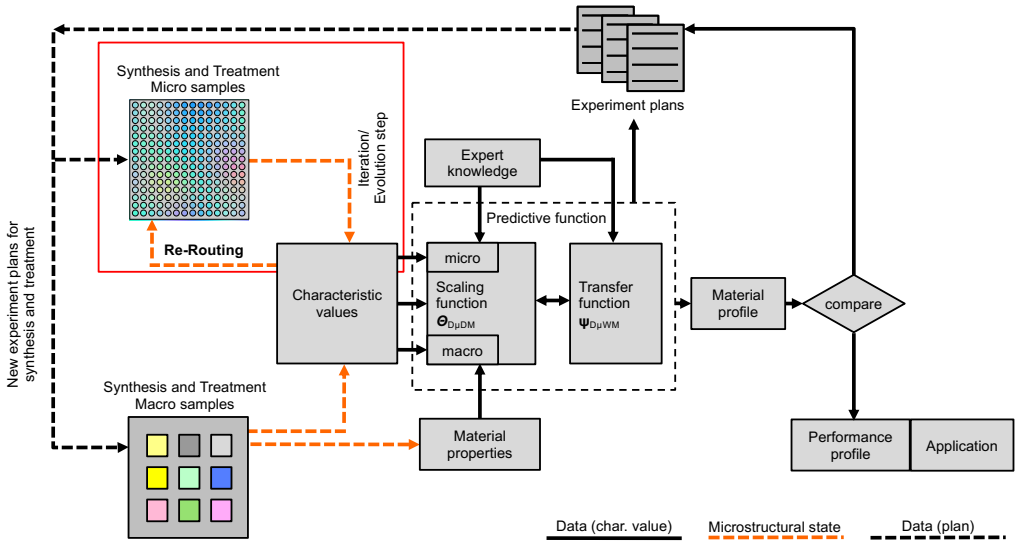

Fig. 2. Information flow of the high throughput approach [13]

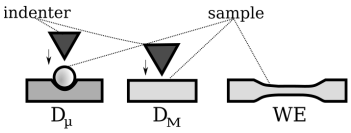

Fig. 3. Relationship between different levels: Shown for the exemplary testing procedure *nano indentation*. From D_μ to D_M only the size changes but not the testing procedure. From D_M to WE the testing procedure changes but the size level remains comparable.

properties by considering different characteristic values of both the micro- and macro-size of a single test procedure at once.

Similar machine-learning techniques are utilized as done in [15] to build the basis of the proposed framework. However, one major extension concerns the seamless integration of formal descriptions. More precisely, the intended framework faces multiple challenges, which all have to be thoroughly addressed to succeed. These challenges are as follows:

1) A high number of dimensions of the input and the output space exists.
2) An online learning capability is required since new data are continuously generated by conducting experiments.
3) It is not ensured that all input dimensions are *valued* for each and every of the grid points.
4) A non-linear relation between input and output space exists.
5) Only a small number of data is available (in relation to the high number of dimensions).

The three aspects 1)-3) are addressed by using a kernel-based regression, as stated in [15]. However, the remaining two aspects 4)-5) are not satisfactorily solved yet. As a consequence, the efficacy of the high throughput approach [12], [13] is still limited and, by this, the full potential of exploring new structural material is not yet exploited. The proposed methodology of this paper closes the gap between high dimensional input space and the prediction. By this, a fundamental contribution to the overall flow is added.

A. Kernel Regression-based Learning

A kernel function provides a powerful learning mechanism, which has been frequently used for classification as well as for regression tasks. The main idea behind such a kernel is the fact that a higher dimension might lead to a linear separability (when used for classification) or a linear dependency (when used for regression). A kernel is function $\kappa : X \times X \to \mathbb{R}$, if there is a Hilbert space $\mathcal{H}$ such that a transformation $f : X \to \mathcal{H}$ exists and $\kappa(x, y) = \langle f(x), f(y) \rangle$ holds [17]. $\mathcal{H}$ is often called feature space. It is generally unfeasible to compute either this space or the inner product explicitly since $\mathcal{H}$ often has a very high dimension. To tackle this, it is taken advantage of the so-called *kernel-trick*, i.e., making use of the equation $\kappa(x, y) = \langle f(x), f(y) \rangle$. By this, it is not required to compute the Hilbert-Space $\mathcal{H}$. The computation $\kappa(x, y)$ yields the same result as long as all computations in $\mathcal{H}$ are expressed in terms of the inner product $\langle \cdot, \cdot \rangle$ and, hence, no transformation to $\mathcal{H}$ has to be conducted.

Different kernel types exist such as *Gaussian kernels*, *sigmoid kernels*, *polynomial kernels* and *radial basis kernels*. According to *Mercer's theorem* [18], as long as $\mathcal{H}$ does have an inner product, i.e., is a Hilbert space, the existence of $f : X \to \mathcal{H}$ is given. For the actual regression a *Nadaraya-Watson-Estimator* is generally invoked, where a kernel-density estimator is employed to estimate the regression function. The support of online (recursive) learning is essential since the underlying learning procedure is still computationally expensive and, hence, it is not feasible to re-train from scratch. For this reason, the proposed framework orchestrates the *Kernel Recursive Least Squares* (KRLS) algorithm as proposed in [19]. This KRLS algorithm [19] uses a *least squares* technique to estimate the regression. The required online learning capability is given by construction since the algorithm is defined recursively. Besides this, the KRLS technique invokes a kernel-estimator within the RLS-core to support even non-linear regressions. According to the situation at hand, any kernel function can be used, which

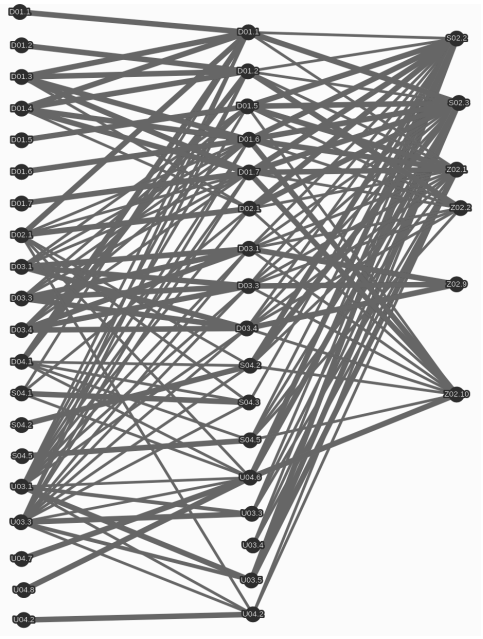

Fig. 4. Adjacency matrix represented as *gene graph*

satisfies *Mercer's theorem*.

Three out of five challenges are addressed by embedding a KLRS technique in the proposed framework. This is since a non-linear online learning algorithm is provided as required by challenges 2 and 4. Furthermore, the use of kernel-functions allows to explore a higher dimensional space and, hence, does not restrict to the ones as chosen initially (cf. challenge 3). Besides this, kernel functions are an effective way to cope with a high number of dimensions (challenge 1).

The remainder of this section focuses on the difficulties, which arise from the combination of challenges 1) and 5), i.e., the high dimensionality and the fact that only a small amount of training data is available.

B. Formal descriptions

The combination of the high dimensionality and the small amount of training data is a challenging one for every prediction framework. This work addresses both challenges by introducing formal descriptions of information – that is known or derived from the data – to guide the prediction process.

Two types of information are considered as follows:

1) *Physically-based information*: This information stores restrictions on characteristic values (on the micro- and macro-level) and material properties that stem from certain physical facts. As a simple example, a length can not be negative, which halves the search space and, more importantly, the prediction space for one dimension.
2) *Experience-based information*: This information is gathered from the material scientists, which has been gained through experiments or has been published in the literature. This information is rather used for identifying main correlations between features than used for restriction of the prediction space. However, it can be used for preferred training of kernel functions.

The proposed methodology focuses on the second type of information. The desired experience-based information (*experts' knowledge*) is represented by four different types of *adjacency matrices*. These matrices encode whether an expert considers that a relation exists between two dimensions.

Definition 5. *An adjacency matrix is a symmetric matrix* $A_{I,O} = (a_{ij})_{i\in\{1,..,dimI\},j\in\{1,...,dimO\}}$*, where* $a_{ij} \in \{0, 0.5, 1\} \cup \{-1\} = V \cup F$*. V can be expanded to allow for more detailed input of the relation's degree if necessary.*

a_{ij} reflects the experts' knowledge about the assumed relations between the i-th component of I and the j-th component of O. If the expert assumes

- no relation at all: $a_{ij} = 0$
- some relation, but unsure about degree: $a_{ij} = 0.5$
- definite relation: $a_{ij} = 1$
- has no insight: $a_{ij} = -1$

$F = \{-1\}$ is included as a flag. Furthermore, as the expert's knowledge about possible relations advances due to a higher number of conducted experiments, the step size for V might be adapted to enable a more precise specification. For example, V might be expanded to $V = \{0, 0.25, 0.5, 0.75, 1.0\}$.

Such matrices have been created for the Scaling Function as well as for the Transfer Function and, furthermore, for two different levels of abstraction. At first, the level of testing procedures is considered, describing whether one testing procedure's results are related to the result's of another one. Secondly, the level of characteristic values is considered. This yields the four different matrices as follows:

- $A^{TP}(\Theta)$: Matrix for scaling function Θ, high abstraction level.
- $A^{CV}(\Theta)$: Matrix for scaling function Θ, fine granularity.
- $A^{TP}(\Lambda)$: Matrix for transfer function Λ, high abstraction level.
- $A^{CV}(\Lambda)$: Matrix for transfer function Λ, fine granularity.

A visualization has been done with a gene graph and is shown in Figure 4 on page 5 for the high abstraction level.

This work utilizes the encoded relations for determining suitable input features to train the specific kernel function such that the likelihood, that the output features are appropriately reflected, is maximized.

Figure 4 states how to predict the characteristic value *D01.1* on the macro-level (first icon from the top in the middle layer). In particular, it is proposed to train the Scaling Function with the three highly related testing procedures on the micro-level (first, third and fourth icons from the top in the left layer). This drastically reduces the dimensionality such that the kernel-based technique, as explained above, is able to cope with the small amount of training data.

It is important to update the adjacency matrix regularly. On the one hand, experts might have obtained new results that lead to a different perception of relations and, on the other hand, the learning process could have yielded results that neglect a once assumed relation at all or reveals a weaker one than initially assumed.

Furthermore, the proposed framework implements a hypothesis system, which allows experts from material science to validate their *potential relations* against the current data basis. A domain-specific language has been developed to provide an easy but powerful mechanism to formulate such a hypothesis. If the hypothesis does not hold during the data-driven validation, the hypotheses system returns a list of counterexamples, i.e., data points including metadata like the corresponding test procedure, which yields this specific value. If no counterexample has been determined by the system, the hypothesis holds with respect to the obtained data. This mechanism significantly contributes to the increase in the expert's confidence when adjusting assumed relations within the postulated adjacency matrices.

IV. Implementation

This section briefly describes the overall implementation of the proposed framework combining formal descriptions with state-of-the-art ML techniques.

The developed framework has solely been written in C++ and the dlib [20] library is used as the back-end for the ML. A document-oriented *MongoDB Enterprise* server instance realizes the central database for the high-throughput approach. This class of databases holds several advantages when dealing with heterogeneous data of high volume [21]. A data exchange component implements a connector to this central database, which allows seamless access while applying required post-processing operations, e.g., filtering, trimming, or normalizing, to ensure that the data is compatible with the ML back-end.

The connected database stores the information of every conducted experiment on the micro- and macro-level in a strictly sample-oriented fashion. The information includes meta-information about the experiments like environmental conditions, the experimental data (characteristic values), the measured raw data, and the material properties in the case of infrequently performed standardized material tests. Further information about the generation and the mechanical/thermal treatment is stored in the database as well, which is required to identify the corresponding data sets - considering the relationships as stated in Figure 3 - for the grid points.

The database scheme is extended to store the formal descriptions in terms of the proposed adjacency matrices (cf. Definition 5). Four different matrices are utilized to store these descriptions, as stated in the previous section. To reflect the evolution of these knowledge base, i.e., the adjustments of the matrices due to new insight to the domain, the matrices are extended by a version control system on the document-oriented database level.

The framework invokes a KLRS algorithm [19] as core[2] of the prediction engine. A pool of different kernel functions is supported by the framework, for instance, linear, radial-basis, and sigmoid kernel functions.

For the later evaluation, the prediction core orchestrates a KRLS technique invoking a sigmoid kernel function. This

[2]Note that this core can be replaced by any arbitrary algorithm, however, depending on the current data pool, this technique seems to be well suitable.

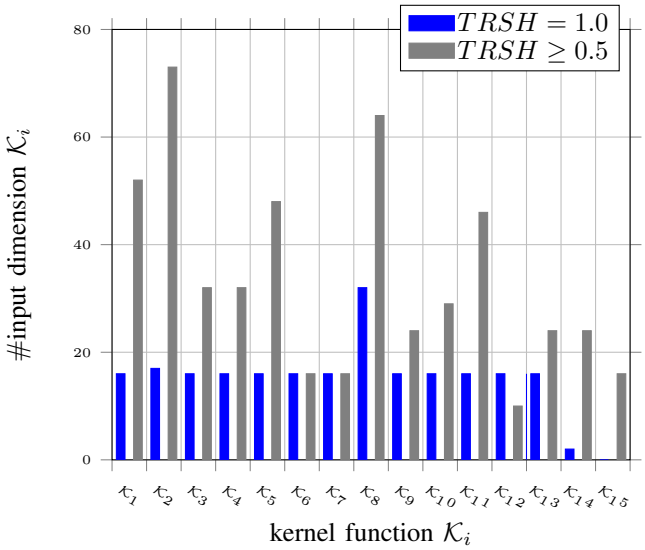

Fig. 5. Built and trained kernel functions for Transfer Function (using alloy 100Cr6 with 5 grid points) with two different level of certainty (TRSH)

function is defined as the hyperbolic tangent of the dot product in the original representation and has shown to approximate even complex, non-linear relationships. For instance, this function is frequently used as the activation function of neural networks [22]. This sigmoid kernel function κ for a pair of values (x_1, x_2) of the matrix M is defined as $\kappa(x_1, x_2) = \tanh(\gamma * M_{1:n}(x_1) \bullet M_{1:n}(x_2) + b)$. Consequently, the two parameters γ and b have to be adjusted to the data. These parameters are determined iteratively and set as follows: $\gamma = \frac{1}{1000}$ and $b = -\frac{1}{4}$. Besides this, the precision is set to $\frac{1}{100}$, which reflects the available data basis.

The general flow of the framework is as follows:

1) The framework receives a prediction request for a set of characteristic values, which should yield the (predicted) macro-material properties.
2) If no suitable kernel-function has been trained yet, a new kernel is built and trained as follows:
   a) A control unit processes user-defined settings like the alloy and the formal descriptions, i.e., the specific adjacency matrices, to determine the required type of data.
   b) The derived data types are used to configure the post-processor of the database connector to gather the corresponding (and available) grid points.
   c) A new kernel function is generated and trained with a default parameter set by considering the grid points. However, at least one grid point is excluded from the training data since one grid point is used to perform the validation.
   d) The last step is repeated iteratively while adjusting the parameter set of the kernel function or even altering the type of function completely.
   e) The most beneficial kernel function is stored in a serialized fashion to be loaded in the case of a prediction request.

3) The prediction core loads the specific kernel functions and invokes them on the given data.
4) Finally, the predicted macro-material properties are returned.

Figure 5 presents the individual kernel functions as yielded by the proposed framework to realize the Transfer Function. Depending on the used threshold (TRSH) value, up to 15 kernel functions are generated by using the currently available database for the 100Cr6 alloy. This threshold determines the required level of certainty in the sense of the introduced adjacency matrices. Consequently, more data sets are considered if a lower threshold is assumed. Note that a single kernel function considers multiple input dimensions (characteristic values) at once, as shown by the histogram in Figure 5, and, more importantly, predicts multiple material properties.

V. Evaluation

This section describes the experimental evaluation of the proposed methodology. At first, the standard approach is evaluated invoking solely kernel regressions without experts' knowledge. Second, the formalized experts' knowledge is combined with the prediction core following the proposed scheme of Section III. The mean absolute error of the prediction accuracy after a five-fold cross-validation [23], [24] is determined since the data is not uniformly distributed and, hence, a bias in the error estimation is prevented. Finally, the rate of impossible predictions, i.e., kernel functions returning `NaN`, is given.

The training data consists of grid points, i.e., fully classified structural materials including roughly 6,500 data points. The resulting dimensions are presented in Table I. Due to this high testing effort for the full classification and, particularly, intensive resource demand for the macro-level experiments, not more than five grid points were obtained.

The evaluation focuses on the Transfer Function, which projects the new test procedures (conducted on macro material) onto the standardized material testing procedures, since this function is even more interesting from a material scientific point of view.

All experiments were executed on an *AMD Ryzen 7 3700X 8-core* CPU running at 4.2 GHz with 32 GB DDR4 system memory. The framework has been compiled by *GCC v9.3.1* (including *dlib v19.2*) on a *Fedora 31* operation system. A *MongoDB Enterprise v3.6.17* runs as the central database server.

A. Evaluation Metric and Results

Let $T = \{1, 2, 3, 4, 5\}$ be the set representing the available data. A number $i \in T$ is chosen and the remaining data is used to train the Transfer Function denoted by Λ_{T_i}. The absolute error is determined by comparing the predicted values $(\Lambda_{T_i}(\boldsymbol{cv}_i^{macro}))$ - where $\boldsymbol{cv}_i^{macro}$ denotes the characteristic values on macro level of grid point i - to the actually measured values $\boldsymbol{mp}_i$.

$$absErr_{T_i} = |\Lambda_{T_i}(\boldsymbol{cv}_i^{macro}) - \boldsymbol{mp_i}|$$

TABLE I
Dimensions (not necessarily independent of each other) being considered during training of Predictive Function ψ.

level	# dimensions	# data points (avg.)
char. values (micro)	205	25
char. values (macro)	134	10
material properties	30	10

$absErr$ is a vector containing the absolute errors for every prediction dimension of the output space, i.e., all 15 kernel functions are considered, as presented in Figure 5. The calculation of an *overall relative error* is required since the range of measured values highly varies. To that end, every component of $absErr$ is divided by the corresponding value of $\boldsymbol{mp}_i$. This procedure can now be repeated for every $i \in T$ and the results are averaged.

$$predError = \frac{1}{5}\sum_{i=1}^{5} overErr_{T_i}$$

The overall run-time of every run is $46.53s$ in average $\pm 0.9s$. This run-time includes the data post-processing, the filtering as well as the training of the KRLS technique. When invoking the prediction framework without the formal descriptions, the *NaN* rate equals 100%. The *NaN* rate can be reduced by 86.7% to 13.3% when utilizing the formal descriptions in the considered scenario, as proposed by this work. Following the evaluation metric above, the $predError$ equals 22.13% and can be decreased to $\approx 17\%$ when excluding the outliers (N=1) from the prediction.

VI. Conclusion

This paper proposed a combination of formal and supervised ML techniques, which yielded the Predictive Function. This function succeeded with the prediction even when only small data had been available for the initial training. More precisely, this work considered two types of information - static as well as volatile information - for deriving formal descriptions, which were modeled in terms of adjacency matrices. Furthermore, a suitable hypothesis system was implemented, which allowed to validate the inserted volatile information. The feasibility of this approach has been proven on the basis of an interdisciplinary material science research problem about the development of new structural materials. Hereby, it was possible for the first time to predict the majority of resulting material properties with sufficient accuracy by processing experimental data, which were conducted by new evolutionary testing procedures without invoking the time- and cost-intensive standardized material tests.

VII. Acknowledgment

This work was supported by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – project number 276397488 – SFB 1232 in subprojects P01 'Predictive function' and P02 'Heuristic, Statistical and Analytical Experimental Design'.

REFERENCES

[1] P. Koopman and M. Wagner, "Autonomous vehicle safety: An interdisciplinary challenge," *IEEE Intelligent Transportation Systems Magazine (ITSM)*, vol. 9, no. 1, pp. 90–96, 2017.

[2] M. Weber and J. Weisbrod, "Requirements engineering in automotive development-experiences and challenges," in *IEEE Joint International Conference on Requirements Engineering (RE)*, 2002, pp. 331–340.

[3] H. Dhotre, S. Eggersglüß, K. Chakrabarty, and R. Drechsler, "Machine learning-based prediction of test power," in *IEEE European Test Symposium (ETS)*, 2019, pp. 1–6.

[4] F. Ye, Z. Zhang, K. Chakrabarty, and X. Gu, "Board-level functional fault diagnosis using multikernel support vector machines and incremental learning," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems (TCAD)*, vol. 33, no. 2, pp. 279–290, 2014.

[5] H. S. Choo, C. Y. Ooi, M. Inoue, N. Ismail, M. Moghbel, S. Baskara Dass, C. H. Kok, and F. A. Hussin, "Machine-learning-based multiple abstraction-level detection of hardware trojan inserted at register-transfer level," in *IEEE Asian Test Symposium (ATS)*, 2019, pp. 98–980.

[6] S. Huhn, S. Frehse, R. Wille, and R. Drechsler, "Determining application-specific knowledge for improving robustness of sequential circuits," *IEEE Transactions on Very Large Scale Integration (VLSI) Systems (TVLSI)*, vol. 27, no. 4, pp. 875–887, 2019.

[7] K. Theining, *Steel and its heat treatment*, 2nd ed. Butterworth-Heinemann, 2013.

[8] H. Cobb, *The history of stainless steel*. Materials Park, Ohio: ASM International, 2010.

[9] E. Hornbogen, "Hundred years of precipitation hardening," *Journal of Light Metals (JLM)*, vol. 1, pp. 127–132, 2001.

[10] J. García, G. Jones, K. Virwani, B. McCloskey, D. Boday, G. ter Huurne, H. Horn, D. Coady, A. Bintaleb, A. Alabdulrahman, F. Alsewailem, H. Almegren, and J. Hedrick, "Recyclable, strong thermosets and organogels via paraformaldehyde condensation with diamines," *Science*, vol. 344, no. 6185, pp. 732–735, 2014.

[11] H. Czichos, T. Saito, and L. Smith, *Springer Handbook of Materials Measurement Methods*. Springer Berlin Heidelberg, 2007.

[12] R. Drechsler, S. Eggersglüß, N. Ellendt, S. Huhn, and L. Mädler, "Exploring superior structural materials using multi-objective optimization and formal techniques," in *International Symposium on Embedded Computing and System Design (ISED)*, 2016, pp. 13–17.

[13] N. Ellendt and L. Mädler, "High-throughput exploration of evolutionary structural materials," *HTM Journal of Heat Treatment and Materials*, vol. 73, pp. 3–12, 2018.

[14] M. Steinbacher, G. Alexe, M. Baune, I. Bobrov, I. Bösing, B. Clausen, T. Czotscher, J. Epp, A. Fischer, L. Langstädtler, D. Meyer, S. Raj Menon, O. Riemer, H. Sonnenberg, A. Thomann, A. Toenjes, F. Vollertsen, N. Wielki, and N. Ellendt, "Descriptors for high throughput in structural materials development," *High-Throughput*, vol. 8, no. 4, 2019.

[15] S. Huhn, H. Sonnenberg, S. Eggersglüß, B. Clausen, and R. Drechsler, "Revealing properties of structural materials by combining regression-based algorithms and nano indentation measurements," in *IEEE Symposium Series on Computational Intelligence (SSCI)*, 2017, pp. 1–6.

[16] A. Sülflow, N. Göckel, and R. Drechsler, "Robust multi-objective optimization in high dimensional spaces," *International Conference on Evolutionary Multi-Criterion Optimization (EMO)*, vol. 4403, pp. 715–726, 2006.

[17] V. Vapnik, *The Nature of Statistical Learning Theory*. Springer, 2000.

[18] De-Gang Chen, Heng-You Wang, and E. C. C. Tsang, "Generalized mercer theorem and its application to feature space related to indefinite kernels," in *International Conference on Machine Learning and Cybernetics (ICMLC)*, vol. 2, 2008, pp. 774–777.

[19] Y. Engel, S. Mannor, and R. Meir, "The kernel recursive least-squares algorithm," *IEEE Transactions on Signal Processing (TSP)*, vol. 52, no. 8, pp. 2275–2285, 2004.

[20] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research (JMLR)*, vol. 10, pp. 1755–1758, 2009.

[21] C. Győrödi, R. Győrödi, G. Pecherle, and A. Olah, "A comparative study: MongoDB vs. MySQL," in *International Conference on Engineering of Modern Electric Systems (EMES)*, 2015, pp. 1–6.

[22] H.-T. Lin and C.-J. Lin, "A study on sigmoid kernels for SVM and the training of non-PSD kernels by SMO-type methods," *Neural Computation (NC)*, vol. 3, pp. 1–32, 2003.

[23] T. Fushiki, "Estimation of prediction error by using k-fold cross-validation," *Statistics and Computing*, vol. 21, pp. 137–146, 2011.

[24] J. D. Rodriguez, A. Perez, and J. A. Lozano, "Sensitivity analysis of k-fold cross validation in prediction error estimation," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32, no. 3, pp. 569–575, 2010.