Scientific Research Publishing

Journal of Computer and Communications, 2025, 13(12), 186-200
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

A Proposal of Genuine Computing Methods in Materials Informatics

Raymond Wu*, Susumu Otsuki

MI-6. Ltd., Tokyo, Japan
Email: *rwu.academic@gmail.com

How to cite this paper: Wu, R. and Otsuki, S. (2025) A Proposal of Genuine Computing Methods in Materials Informatics. *Journal of Computer and Communications*, **13**, 186-200.
https://doi.org/10.4236/jcc.2025.1312011

Received: October 19, 2025
Accepted: December 23, 2025
Published: December 26, 2025

Abstract

In the world of Material Informatics (MI), conventional methods involve tremendous laboratory work or extensive simulations that may not yield the expected results. Our objectives are to contribute to the originality of the standardization of the whole process of MI, starting with data quality measurement, the reusability of data and models, the development of a genuine data structure for machine learning, the alteration of design space, and finally, a novel coefficient-based analysis that can optimize the target yield and the likelihood. To achieve the objectives, our research has focused on the numerical analysis of informatics in materials science, supported by innovations in measurement and optimization processes. It also provides an overview of some of the recent successful data-driven MI strategies undertaken in this decade. The research also identifies some challenges the community is facing and those that should be overcome shortly, and streamlines a genuine process of MI.

Keywords

Materials Informatics, Machine Learning, Bayesian, Data Foundation

1. Introduction

Materials Informatics (MI) has become critical; however, several areas can be bottlenecks to the applications of Machine Learning (ML) in materials science. Common issues of MI are listed below, where the significant problems are classified into data quality issues, data reliability, and data and model reusability issues. Many institutes have started to seek advanced methods of computational algorithms to support their materials design. However, the lack of a genuine process in MI affects standard industry practices in materials science, leading to a loss of interoperability across the entire industry.

From the data reliability perspective, a lack of reliable data makes it challenging

to achieve the objectives of data reuse or replication of the design patterns. Consequently, material discovery and applications have been slowed down, which has had a profound impact on industry innovations.

As shown in the guidelines of the industry leaders; early detection and effective communication are crucial to ensuring data reliability. Addressing data quality issues requires both technical solutions and human intervention.

From the viewpoints of ML, ML methods are in high demand for MI; however, their reusability is low, and such a limitation significantly impacts the extrapolation capability of ML and makes the discovery of new materials difficult. The issues of reusability are linked to the validity and credibility of models to be reused, as well as the cost and time required for familiarization [1].

While attempting to reproduce a section of the results presented in a general-purpose ML framework for predicting properties of materials, many organizations have encountered several challenges, including:

1) reporting software dependencies
2) recording and sharing version logs
3) sequential code organization, and
4) clarifying code issues [2].

On the other hand, system parameters have a close linkage to model reusability, and system parameters of ML in materials discovery can be crucial to the macroscopic properties and characteristics of composition, microstructure, and processing conditions.

With the guidance of these key parameters to constrain the input and output, the prediction accuracy of the ML model can be significantly improved during the development of high-performance materials in any given system (Yuan, J., Li, Z., Wang, Q., 2024). However, in the current situation, the reusability of key parameters is low, and most situations require starting from scratch.

Under such circumstances, applications of ML in materials science often involve the use of a specific method tailored to a particular type of material. Alternatively, selecting a particular process through a comparison of the results of multiple ML methods can also be a common approach. This limits the application range of any specific model to a very narrow scope [3].

2. Literature Review

ML has been adopted in MI to reuse existing experimental and computational data in materials science, allowing well-trained models to be distributed across similar research in MI. However, in the current situation, some tactical issues need to be resolved to address the bottlenecks.

Data reusability can be one of the significant factors supporting MI in new materials discovery and materials design. However, the standard issue is that most experimental materials scientists lack experience in data-driven research.

Furthermore, materials scientists are unfamiliar with ML and fail to deploy genuine methods for MI that can be shared across the MI platform [4]. Most of the

literature focuses on specific conditions and struggles to establish generalization across a more exhaustive range of MI.

In the current situation, because most organizations are still focused on laboratory work rather than computation through a consolidated platform, the barrier remains high, and many organizations have spent considerable effort preparing raw data instead of reusing or refining existing patterns. A comparison of lab-based data and ML-driven data is illustrated in Figure 1 (Patterns A and B).

Most of the investigations do not cover the foundation work. The lack of literature in the investigation of the foundation of the reusability issue leaves a gap that can be difficult to fill. Consequently, achieving global best practices and genuine methods is not easily accomplished.

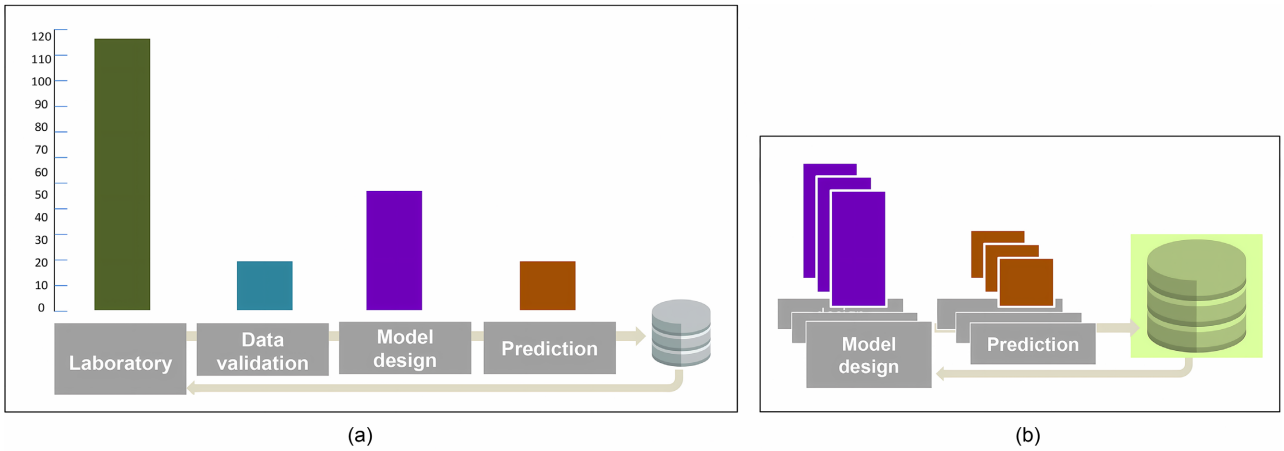

Figure 1. (a) Pattern A. conventional laboratory-driven ML; (b) Pattern B. data-driven ML.

In this investigation, our observations reveal that the tactical issues were caused by the gap between materials scientists and data scientists, which is linked to data validation in the context of data science. In other words, data produced by the laboratories are far from ready for data analysis work.

This also implies that most laboratory scientists do not understand how ML works on MI and are misled about the definition of data quality because they consider data quality to be related to the accuracy of experimental data.

Therefore, most MI scientists may spend over 50% of their effort fixing defects or leave some defects behind that impact the results of MI, even though such issues can be easily discovered in the laboratory work or the data architecture issues can be fixed through a strategic plan of the organization to settle the problems.

By utilizing suitable AI algorithms to optimize the design and discovery of materials, new materials can be systematically developed, the gap can be minimized, and the waste of resources and environmental pollution caused by numerous unnecessary experiments can be significantly reduced.

A framework is critical to minimizing the gap. In the MI world, the lack of an industry framework to guide standard practices is a significant concern. Examples include the fact that ANN models are based on feedforward neural networks.

In contrast, various types of neural networks, such as time series networks, recurrent neural networks, and adversarial networks, have emerged in ML [5]. This implies that a complex of ML methods is required in analyzing MI, and the use of various new neural networks mixed together with materials science will become a future trend. However, a framework is now available to guide the selection of suitable MI and its dependencies under these conditions.

Several studies in the literature have indicated the quality issues of ML algorithms and standard methods, such as Perdew-Burke-Ernzerhof (PBE), density functional theory (DFT), strongly constrained and appropriately normed (SCAN), r2SCAN, and generalized gradient approximation (GGA).

Among these, r2SCAN predicts more accurately than other methods, and it has delivered on its promises of efficiency and accuracy, requiring modestly fewer computational resources than others while offering much more reliable convergence [6].

However, when we compare the utilization of the MI software, the deployment of those tools, which were recognized as having higher accuracy, does not show a better adoption rate. Most of the reasons are linked to reliability in a particular situation. These are indicated in Figure 2 and Figure 3.

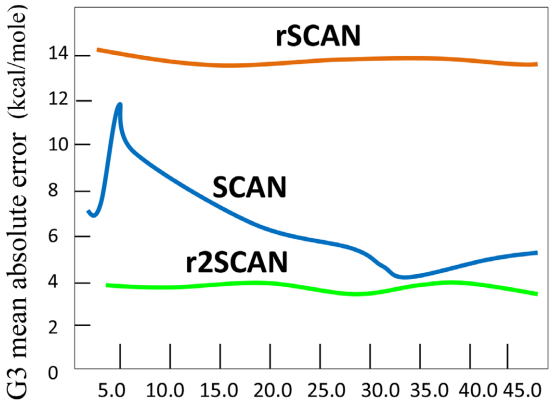

Figure 2. Relative grid-point counts.

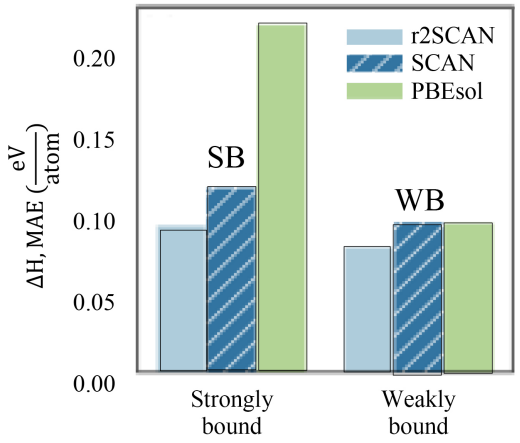

Figure 3. Strongly bound and weakly bound.

From the above review, reliability, accuracy, system performance, and computational resource requirements are the critical factors for algorithm selection. However, the tactical issues are caused by the data foundation, which is linked to data quality, data analysis, and data representation.

Several fundamental aspects of data-driven materials science are built upon the key elements of open science. As an example, the European Commission considers open science a new approach to the scientific process. The foundation is based on cooperative work and new ways of disseminating knowledge using digital technologies and collaborative tools [7]. These aspects reflect those that are particularly relevant to the birth and future of data-driven materials science [8].

3. Methods of MI

In order to achieve our objectives in the standardization of data quality measurement, a genuine data structure for ML, and the building of a robust MI platform, this paper is organized as follows:

- The data quality measurement methods and standards are discussed.
- The cross-validation process, DBSCAN for data clustering, Automatic relevance determination (ARD) feature weight analysis, and Bayesian optimization methods are introduced to streamline the MI processes.
- The concept of the MI platform is proposed, which supports an end-to-end process.

Data quality measurement is crucial because it reflects the improvement through each change. In this paper, R^2 (R-squared) and MSE are used to measure data quality. The primary purpose of deploying R-squared is to predict future outcomes.

Simultaneously, R-squared is used in the testing of hypotheses based on other related information, which is commonly referred to as the coefficient of determination. A value of 0 indicates that the response variable cannot be explained by the predictor variable at all. A value of 1 indicates that the response variable can be perfectly explained without error by the predictor variable. R-squared can be represented as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2}{\sum_{i=1}^{n}\left(y_i - \bar{y}_i\right)^2}.$$

where R^2 can be derived by using the r2_score function within the Python scikit-learn package. MSE measures the average of the squares of the errors—that is, the average squared difference between the estimated values and the actual values.

MSE of the predictor is computed as follows: $\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{y}_i\right)^2$.

In Python, the MSE can be zero if the estimation is 100% accurate, matching the observation perfectly.

3.1. R:M Ratio

This paper proposes the R:M ratio as the key measurement method for data quality.

R:M ratio = R^2/MSE, which divides the prediction of future outcomes by the average of the squares of the errors. Theoretically, the R:M ratio can be within the range of 0 and indefinite; however, a value greater than 5.0 can be considered a good value. The goal of this investigation is to maximize the target value (yield) through the optimization process. The R:M ratio supports the measurements and optimization of the objective function.

The objective function optimizes the variant features during training; hence, it is critical. For example, the probability of generating a training set in the maximum likelihood approach is one of the objective functions in this investigation.

To optimize the objective function (the likelihood), we need to find the maximum likelihood estimation (MLE). The MLE solution models the distribution through the Gaussian distribution, where $\theta = (\mu, \sigma^2)$.

Through the optimization process, we derive θ from the MLE, allowing us to achieve our objective function and maximize the likelihood. The other objective is to investigate the relationship between the level of regression and data quality performance; for this purpose, we apply a threshold to filter the data based on variant target values.

On the other hand, feature evaluation is critical to the optimization process. In the feature evaluation model, the "feature importance" report and F-score are generated through XGB and other models, which are helpful in the identification of key features. Both regression and classification methods are applied in the evaluation.

3.2. Bayesian Optimization (BO)

BO is a strategy for optimizing objective functions that are expensive to evaluate. It operates by building a probabilistic model of the objective function and using this model to select the most promising points to evaluate next. In addition, Bayesian likelihood can be an important indicator of the optimization that balances the target yield. Through this pattern, BO gives us a principled way of optimizing hyperparameters as follows:

Hyperparameters assignment: $x \in X$ Objective function: $\theta : X \to R$

Acquisition function (Ă): $X \to$ Optimized (μ, σ, $y_{\max}$)

To address the hyperparameter scaling issue, the ARD squared exponential kernel is applied. The formula is expressed as follows:

$$K_{\text{ARD}}(x, y) = \theta_0 \cdot \exp\left(-\frac{1}{2} r^2 (x - x')\right) \quad \text{where} \quad r^2 (x - y) = \sum_{i=1}^{d} \frac{(x_i - x_i')^2}{\theta^2}$$

The ARD regression provides a sparser solution. Each ARD coefficient ω_i can be drawn from a Gaussian distribution, centered on zero and with a given precision. ARD coefficients are generated by using coeff_ARD = ARDRegressor.coef. Some of the non-informative coefficients are set precisely to zero, while others are shifted closer to zero. Some non-informative coefficients still retain large values [9].

3.3. Regression Deviation Degree Threshold

Regression Deviation Degree Threshold (RDDT) was proposed in this research to filter in a certain group of regression levels of data. In the experiments, RDDT is set to 1 ··· 30 (%) to investigate the impacts of regression level on the optimization process.

The general equation can be expressed as:

$$D\left\{\forall y_i \in Y, y_i(\varepsilon) \le \Delta\right\} \Rightarrow y_{\text{pred}}[i] - y_{\text{obs}}[i] \le \varepsilon$$

such that, $\text{abs}\left(\left(y_{\text{pred}}[i] - y_{\text{obs}}[i]\right) / y_{\text{pred}}[i]\right) \le \text{RDDT}_{\text{m}}$

In the experiments, we apply different RDDT to filter in those data of different regression levels. For example, if we set RDDT = 10%, then among all 132 records of data, only 24 records meet the requirement of:
$\text{abs}\left(\left(y_{\text{pred}}[m] - y_{\text{obs}}[m]\right) / y_{\text{pred}}[m]\right) < 0.10$. These 24 records are the results of the filtering process under RDDT = 10%. As demonstrated in Figure 4, different RDDT (*i.e.*, 5% or 10%) will cover records with different levels of data quality.

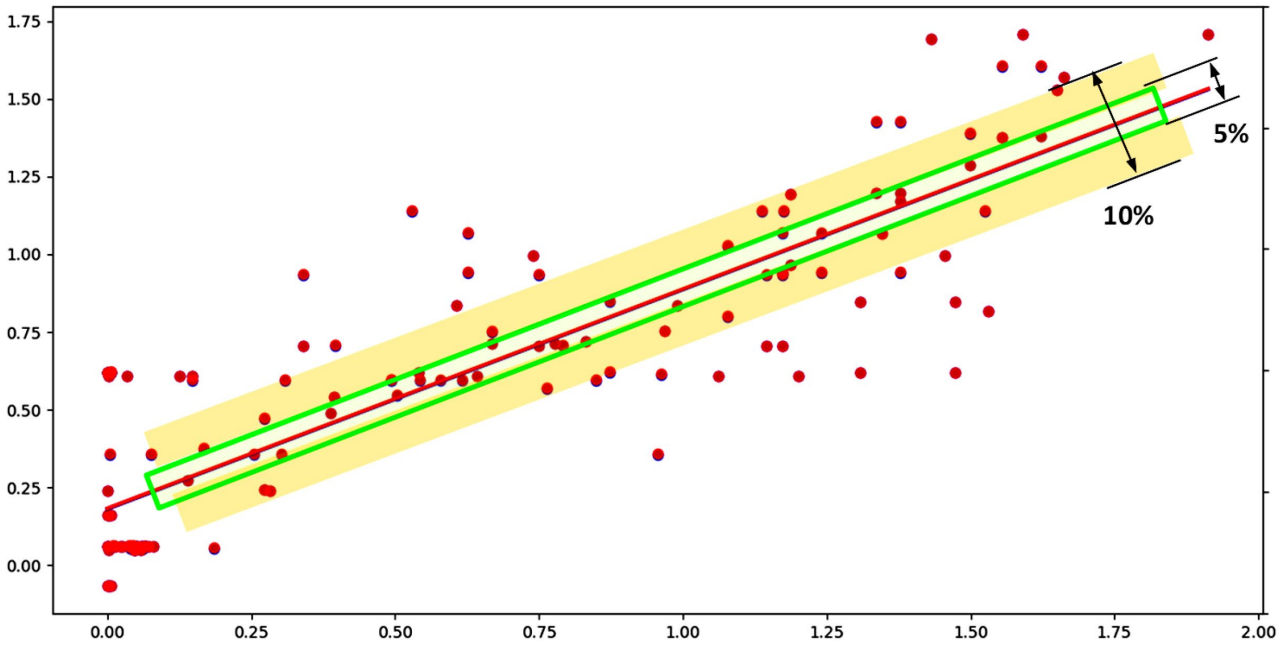

Figure 4. RDDT control in the filtering process.

For the same reason, when we set RDDT = 5%, which is stricter than 10%, only 11 records among 132 records are filtered in, and these 11 records are the results of the filtering process under RDDT = 5%. The statistics of record sizes for different RDDT values are shown in Table 1.

Table 1. Statistics of record counts of different RDDT.

RDDT	Original	30%	10%	5%
Data-frame	132 rows × 235	52 rows × 235	24 rows × 235	11 rows × 235

RDDT helps to identify those data with high regression and supports the optimization process. The machine learning and the laboratories need close collaboration to source the experimental conditions that produce the data with low RDDT.

When RDDT = 1%, the regression degree is very high, which means the prediction is very close to the observed value. The experiments conclude the following: a narrow control of RDDT can increase the mean value of the target yield and minimize the deviation, and this can be represented by Gaussian distribution:

$$g(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

In Python, this can be represented by:

```
1/np.sqrt(2np.piv) * np.exp(-(x - mu)**2/(2*v))
```

As shown in Figure 5, the blue (the target) and the red (the problem) are compared.

$g(x)$ = (1/100.5)(1/(2pi)0.5) * math.exp((-(0)**2)/(20)) = 0.12616 (the blue)

$g(x)$ = (1/500.5)(1/(2pi)0.5) * math.exp((-(0)**2)/(100)) = 0.05642 (the red)

Our optimization of $\theta=(\mu,\sigma)$ mainly works on maximizing the mean (μ) and secondarily on minimizing the deviation (σ). According to: $g(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$, both factors (maximizing μ and minimizing σ) will achieve a higher likelihood ($g(x)$).

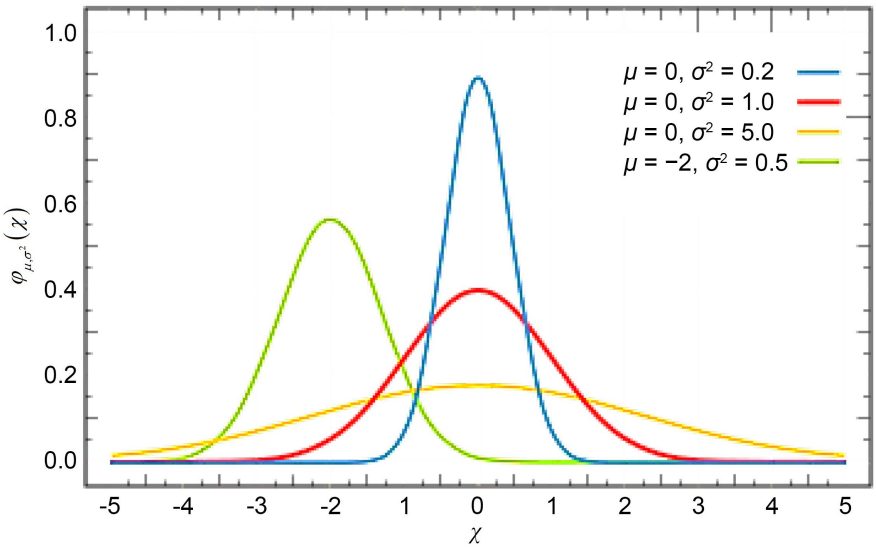

Figure 5. Regression analysis of different RDDTs.

Through the RDDT definition: $\text{abs}\left(\left(y_{\text{pred}}[i]-y_{\text{obs}}[i]\right)/y_{\text{pred}}[i]\right)$, we understand that a lower RDDT can achieve a lower σ value and a higher μ value. The experiments are elaborated in the next section (experiment results of RDDT).

Parallel to $\theta = (\mu, \sigma)$ calculation, our experiments were extended to likelihood estimation.

Likelihood is used as an objective function through the Gaussian process to check each point (of 10,000 samples) for their likelihood of $X\_sum = y_{pred} = y_{obs}$.

The area > max(X_sum) is integrated as the objective. Instead of changing the data, the parameters are adjusted. The likelihood is generated by an underlying Gaussian process, and the likelihood function (L) is the Gaussian itself. Meanwhile, as the log() function is monotonically increasing, log likelihood is used to find the θ that maximizes $f(\theta)$.

4. Experiment Results

The experiments on the method of model evaluation, R:M ratio, RDDT setting, and Bayesian optimization as proposed in Section 3 (Methodology) produced the following results;

4.1. RDDT Setting and Regression Analysis

In RDDT experiments, y_{obs} is used to denote the observed values, and y_{exp} is used to represent the expected values. The testing data comprise 565 records, with each record consisting of 264 features (the multivariate x values) and a single target value (y). The RDDT is set to 1 ··· 30 (%) in the experiment, where RDDT = {1 ··· 30%}, $n = 30$, and $RDDT_1 = 1\%$ ··· $RDDT_{30} = 30\%$.

For example, when RDDT1 = 1% is applied, only those records whose (|yi exp - yi obs|)/yi obs ≤ 1% will be filtered in. The lower the value of RDDT is, the fewer records will be filtered out, and the higher the degree of regression can be expected. In Figure 6, variant RDDT (1:30) is evaluated against the R:M ratio of 565 records. In the range of 3-8% of RDDT settings, a higher R:M ratio (> 10) can be achieved. In Figure 7, our experiment further achieves a higher μ (μ) value and higher likelihood.

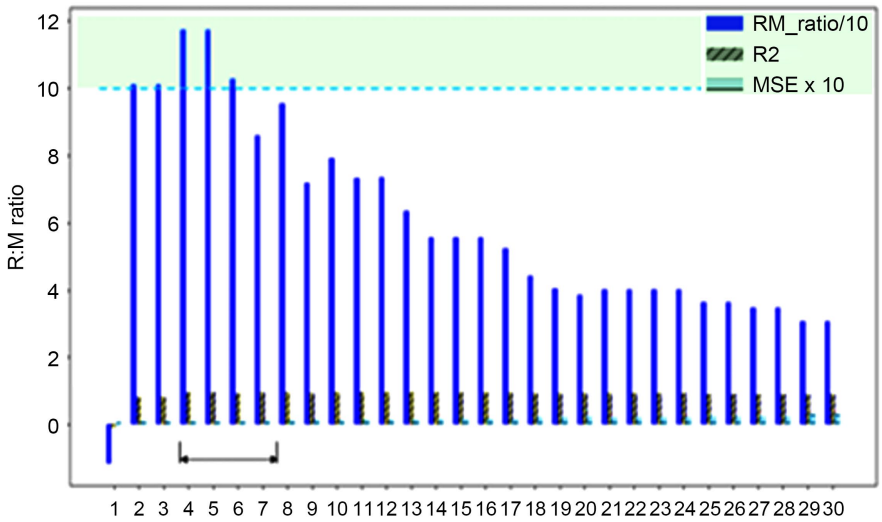

Figure 6. R:M ratio distribution across variant RDDT (1:30).

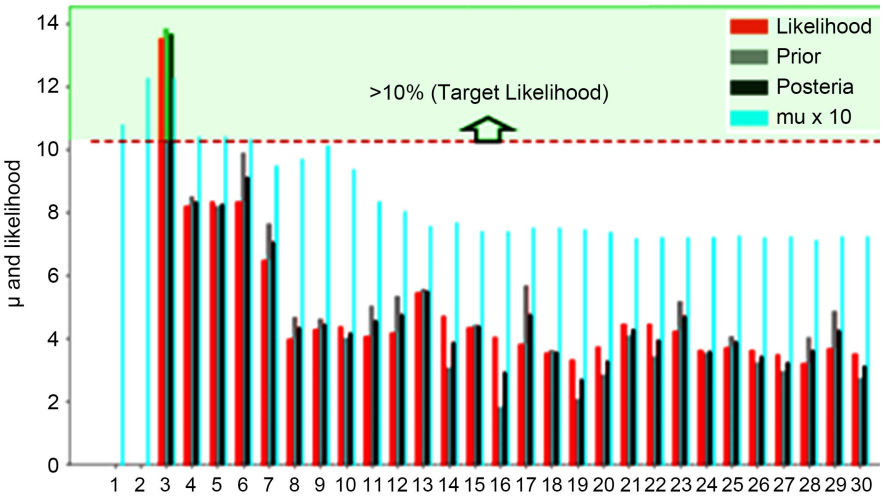

Figure 7. μ and likelihood distribution across variant RDDT (1:30).

4.2. Relationship between RDDT and R:M Ratio

The combination of the R:M ratio and RDDT can be a good solution for data quality measurement. First, when we use samples from different RDDT groups of 1, 5, 10, 15, 20, 25, 30, and all (100%), each group has a different value for the R:M ratio and the regression of the target value (Y_{obs} vs. Y_{pred}).

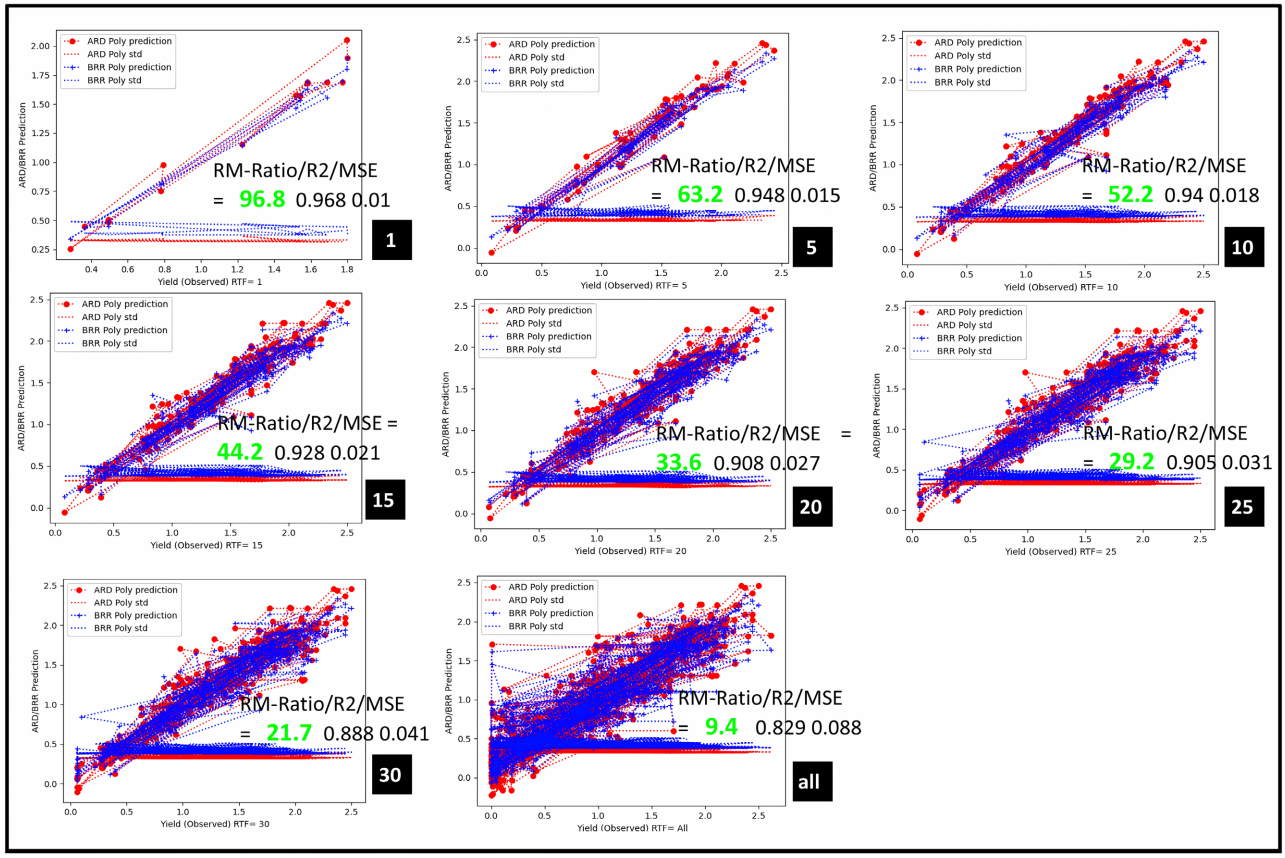

Figure 8. R:M ratio versus RDDT.

As shown in Figure 8, the R:M ratio can be an indicator of regression analysis. In order to achieve our objective function by optimizing $\theta(\mu, \sigma^2)$, the regression analysis is critical in the optimization process.

This implies that we attempt to maximize the y_{pred} target value and, at the same time, as shown in the following equations. We hope to filter in the high regression data through a minimum RTF;

$$y_{\text{pred}} = \sum_{j=0}^{m} \sum_{i=0}^{n} \left(\omega_i * x_{ji} \right) = X\_sum_j + \varepsilon$$

$$\text{abs}\left(X\_sum + \varepsilon - y_{\text{obs}} \right) \quad \text{or} \quad \text{abs}\left(y_{\text{pred}} - y_{\text{obs}} \right) < \text{RTF}(\%)$$

$\mathbb{R}_{\text{RDDTn}}$ denotes the domain of association of the multivariant $\sum_{k=1}^{n} x_{ik} \cdots x_{jk}$ and $y_i \cdots y_j$ (where $i, \cdots, j$ denotes the row records, and k denotes the columns' variant features).

Therefore, $\mathbb{R}_{\text{RDDTn}} \in \left\{ \sum_{k=1}^{n} x_{ik} \cdots x_{jk} \mid y_i \cdots y_j \right\}$, which satisfies $\left| y_{i\text{exp}} - y_{i\text{obs}} \right| / y_{i\text{obs}} \le \text{RDDT}_{\text{n}}$.

4.3. Exploitation and Exploration Process

The exploitation and exploration process is applied to narrow down the search criteria for the optimization strategy [10]. To expedite the search process in exploitation and exploration, we demonstrate how the original domain space, 565 × 264 (row-column), was filtered using DBSCAN feature selection and ARD row selection.

On the one hand, the exploration process is applied to expand horizons to discover unique possibilities. Because it involves unfamiliar territory, exploration often leads to the acquisition of new knowledge, insights, and skills that can be valuable in the long run. It fosters creativity, adaptability, and resilience, as individuals and organizations learn to navigate ambiguity and embrace change (Sinha, S., *et al.*, 2015).

On the other hand, the exploitation process is applied to maximize the benefits derived from existing opportunities. Exploitative strategies often involve incremental improvements, allowing for a gradual and controlled evolution. This approach is particularly valuable in that it emphasizes maintaining and optimizing existing systems rather than venturing into the unknown.

First, when acquisition in the exploration process is initiated, BO provides the optimization of hyperparameters by using a Gaussian process. The acquisition function can be expressed as Ψ: $X \rightarrow$ Optimized (μ, σ, y_{best}), and $\forall x \in X$, x is assigned to the objective function $\theta : X \rightarrow R$.

Let ψ denote the acquisition function and determine what we search. Typically, ψ is a function of the mean, variance, and the best observed value—that is, $\psi(\mu, \sigma, y_{\text{optimized}})$.

As shown in the Figure 9, 200 sample data points ranging between 2.16 and 5.48 are provided, and $\mu(x)$ - $3\sigma(x)$ is calculated for each data point. Among the 200 points, the min($\mu(\dot{d})$ - $3\sigma(\dot{d})$) is derived, and the associated $\dot{d}$ value (4.10 in this case) will be used to divide and conquer the sample space into two areas. Such

a divide-and-conquer algorithm will be continued for the remaining regions until no further candidate can be found.

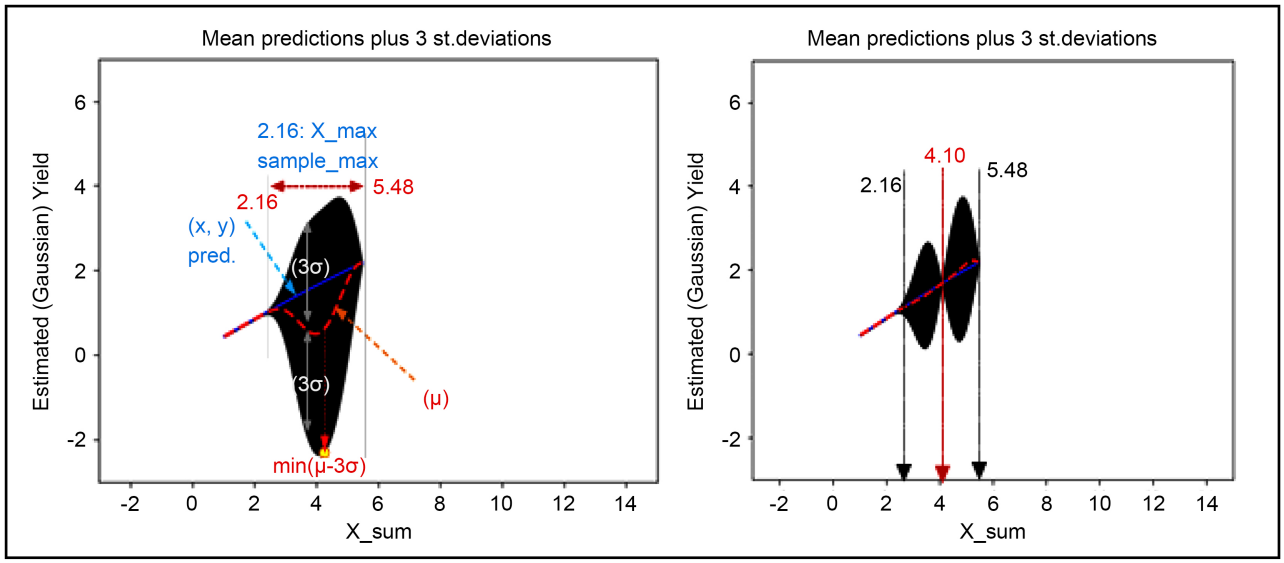

Figure 9. Exploitation and Exploration process (3σ).

In a similar algorithm in BO, the priority is to choose the next sampling point by maximizing the acquisition function $\tilde{a}(x)$: $x^* = \arg\max_{x \in X} \tilde{a}(x)$.

In the exploitation and exploration process, the acquisition function is key to striking a balance between exploration and exploitation. In the acquisition function, upper confidence bound (UCB) is applied. In UCB, let $\tilde{a}(x) = \mu(x) + \kappa\sigma(x)$, where κ is a constant that implies the following:

Exploration: By comparing two points x_1 and x_2, if their means are the same, then the one that has the larger $\chi^2(x)$ will be picked. Exploitation: By comparing two points x_1 and x_2, if their variances are the same, then the one that has the larger $\mu(x)$ will be picked.

Based on the above algorithm, the original 500-sample data are decomposed into multiple divisions in several steps to discover the optimized target value.

5. Discussions

The optimization process in machine learning can maximize the target yield and increase the likelihood or possibility that requires a standard process. Our research aims to standardize the machine learning process of materials informatics, as in the current situation, data quality measurement, data reusability, and model reusability are randomly distributed across the industry, and very often, data scientists of materials informatics miscommunicate due to inconsistent methods or ad hoc approaches.

In order to solve the problem, we first start with the data quality measurement, followed by the reusability of data and models, and finally, we develop a novel coefficient-based analysis that can optimize the target yield and the likelihood.

In terms of methodology, we propose the concept of RDDT as a parameter and the R:M ratio to control data quality, and trace back to the original processing conditions that can support continuous improvement. In reviewing the above investigation, we propose approaches to shape the design space in the column direction, row direction, and regression level direction.

This implies that only the essential features, high-performance clusters, and records showing high regression are filtered in. The R:M ratio is used to measure data quality when the data scientist receives data sets from the laboratory, while RDDT is applied in model evaluation and process control in a recursive cycle of continuous improvement.

After shaping the column and row directions, the retained design space achieves a better R:M ratio (14.3) compared to the filter-out space, which has R:M ratios of 2.9 and 1.4. Meanwhile, the target yield averages 0.514, compared to the rest of the space (0.501 and 0.463).

Furthermore, the experimental results in the regression-level direction reveal that when the records' R:M ratio is higher than 10.0, the mu (μ) value and likelihood can be maximized when the RDDT is controlled under 3% - 8% (averaged 5%).

One of the common issues in the MI platform is the computing capability of cache memory and the reusability of the MI algorithms. To facilitate a highly efficient MI platform, we utilize a semantic cache (supported by SCM) that enables persistent data retention in the cache. This approach supports a high hit ratio and high-performance computing, allowing all similar patterns and data to be reused by all users.

6. Conclusions

Our research aims to envision the criticality of standardization of the machine learning process in materials informatics. Data quality highly impacts the results of machine learning and needs a standardized measurement of source data before machine learning. In addition, data reusability and model reusability are crucial to materials informatics.

In this research, we propose methods of data quality measurement, followed by the reusability of data and models. We propose RDDT data measurement and the R:M ratio to control data quality. Under this framework, high regression data quality and machine learning optimization are improved through a recursive cycle of continuous improvement.

To address the current situation of fragmentation and abnormal MI processes, this research examines the necessary methods for governance in multistep MI processes. The investigation begins with data quality, data structure, data integrity, and system performance, followed by a review of the design space at all column, row, and regression levels.

Based on the averaged feature importance derived from XGB, RF, and SVM, feature selection becomes possible. Additionally, based on cluster selection

through the DBSCAN tool, row selection is feasible.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Robinson, S., Nance, R.E., Paul, R.J., Pidd, M. and Taylor, S.J.E. (2004) Simulation Model Reuse: Definitions, Benefits and Obstacles. *Simulation Modelling Practice and Theory*, **12**, 479-494. https://doi.org/10.1016/j.simpat.2003.11.006

[2] Persaud, D., Ward, L. and Hattrick-Simpers, J. (2024) Reproducibility in Materials Informatics: Lessons from 'A General-Purpose Machine Learning Framework for Predicting Properties of Inorganic Materials'. *Digital Discovery*, **3**, 281-286. https://doi.org/10.1039/d3dd00199g

[3] Yuan, J., Li, Z. and Wang, Q. (2024) Applications of Machine Learning Method in High-Performance Materials Design: A Review. *Journal of Materials Informatics*, **4**, 14. https://doi.org/10.20517/jmi.2024.15

[4] Wang, A.Y., Murdock, R.J., Kauwe, S.K., Oliynyk, A.O., Gurlo, A., Brgoch, J., *et al.* (2020) Machine Learning for Materials Scientists: An Introductory Guide toward Best Practices. *Chemistry of Materials*, **32**, 4954-4965. https://doi.org/10.1021/acs.chemmater.0c01907

[5] Fu, Z., Liu, W., Huang, C. and Mei, T. (2022) A Review of Performance Prediction Based on Machine Learning in Materials Science. *Nanomaterials*, **12**, Article 2957. https://doi.org/10.3390/nano12172957

[6] Furness, J. (2020) Accurate and Numerically Efficient r2SCAN Meta-Generalized Gradient Approximation. *The Journal of Physical Chemistry Letters*, **11**, 8208-8215.

[7] Foster, E. and Deardorff, A. (2017) Open Science Framework (OSF). *Journal of the Medical Library Association*, **105**, 203-206. https://doi.org/10.5195/jmla.2017.88

[8] Himanen, L., Geurts, A., Foster, A.S. and Rinke, P. (2019) Data-Driven Materials Science: Status, Challenges, and Perspectives. *Advanced Science*, **6**, Article 1900808. https://doi.org/10.1002/advs.201900808

[9] Rodrigues, F., Ortelli, N., Bierlaire, M. and Pereira, F.C. (2022) Bayesian Automatic Relevance Determination for Utility Function Specification in Discrete Choice Models. *IEEE Transactions on Intelligent Transportation Systems*, **23**, 3126-3136. https://doi.org/10.1109/tits.2020.3031965

[10] Wilson, R.C., Bonawitz, E., Costa, V.D. and Ebitz, R.B. (2021) Balancing Exploration and Exploitation with Information and Randomization. *Current Opinion in Behavioral Sciences*, **38**, 49-56. https://doi.org/10.1016/j.cobeha.2020.10.001