Elemental Diffusion Coefficient Prediction in Conventional Alloys Using Machine Learning

Arjun S. Kulathuvayal,[†] Yi Rao,[‡] and Yanqing Su[*,†]

†*Department of Mechanical and Aerospace Engineering, Utah State University, Logan, UT 84322-4130, USA*

‡*Department of Chemistry and Biochemistry, Utah State University, Logan, UT 84322-4130, USA*

E-mail: yanqing.su@usu.edu

Phone: +1 435-797-0957

Abstract

This paper presents ML-DiCE (Machine Learned Diffusion Coefficient Estimator), a comprehensive machine learning framework designed to predict diffusion coefficients in impure metallic (IM) and multi-component alloy (MCA) media. The framework incorporates five ML models, each tailored to specific diffusion modes: (1) impurity and (2) self-diffusion in IM media, and (3) self, (4) impurity, and (5) chemical diffusion in MCA media. These models use statistical aggregations of atomic descriptors for both the diffusing elements and the diffusion media, along with the temperature of the diffusion process, as features. Models are trained using the random forest and deep neural network algorithms, with performance evaluated through the coefficient of determination (R^2), mean squared error (MSE), and uncertainty estimates. The models within this framework achieve an impressive R^2 score above 0.90 with MSE less than 10^{-16} m^2/s, demonstrating high predictive accuracy and reliability for diffusion coefficient.

Keywords

elemental diffusion, diffusivity prediction, machine learning, conventional alloys, atomic features

1 Introduction

Conventional metallic alloys are composed of one or two principal elements along with additional minor alloying elements in negligible quantities to modify their microstructures and properties. These traditional alloys have a long history of successful applications and have been extensively studied and optimized over the years. This study employs the application of Machine Learning (ML) in predicting an important elemental diffusion parameter – diffusion coefficient in conventional metallic alloys. The diffusion coefficient represents the proportionality constant in Fick's laws of diffusion, which describe the relationship between the flux of particles and their concentration gradient.[1] Typically, the elemental diffusion process in alloys is a thermally activated process, exerting a profound influence on alloy microstructures and hence mechanical properties.[2,3] The diffusion process can be classified based on modes of diffusion such as self-diffusion, interstitial diffusion, impurity diffusion, chemical diffusion, and grain boundary diffusion. The diffusion coefficient is a key factor in all the aforementioned diffusion modes, intricately linked to various factors of the diffusion medium and the diffusing elements, wherein especially their chemical and physical characteristics. Given the multifaceted association of the diffusion coefficient with intrinsic and extrinsic material properties, the prospect of employing ML algorithms to comprehend and decipher these relationships appears promising. Thus the goal of this study is to predict the diffusion coefficient for a given temperature by incorporating a wide range of physio-chemical features of diffusion medium and diffusing element for a specific diffusion mode.

Experimentally determining diffusion coefficients typically involves two main approaches: the tracer method and the interdiffusion method. The tracer method involves tracking diffusing species tagged with radioactive isotopes, allowing for the determination of tracer diffusion coefficients by applying known diffusion solutions to measured concentration profiles. However, this method necessitates numerous independent measurements across various homogeneous alloys, with the number of required compositions scaling exponentially with the number of components, making the tracer method a time-consuming process. In addition

to this, the availability and stability of radioactive isotopes present further challenges especially when elements like Al and Ca are employed as diffusing species.[4] In such scenarios, the secondary ion mass spectrometry-based thin-film technique mitigates the stability problem to an extent by employing enriched stable isotopes.[5] Additionally, the diffusion coefficient calculation is hampered by the dependence on heavy isotopes as tracers especially when comparing different isotopic species of the same molecule. This is because potential functions are invariant with isotopic substitution within the bounds of the Born-Oppenheimer approximation, potentially overlooking many important but poorly understood factors.[6] On the other hand, chemical diffusion, or interdiffusion, is measured by bringing alloys of different compositions into contact and inducing diffusion transport through chemical potential gradients. In this approach, techniques like the Boltzmann-Matano method in binary systems and the Matano-Kirkaldy method in ternary systems are commonly employed to determine the diffusion coefficients, offering efficient alternatives to the tracer method.[7,8] However, for multicomponent alloys (number of components ≥ 4) it is generally impossible to apply this Onsager-formalism-based scheme to estimate the entire matrix of independent interdiffusion coefficients, as the given number of independent diffusion paths, which are one-dimensional by definition, cannot intersect in a multi-component space.[9–11]

These days, first-principles calculations and molecular dynamics simulations have become efficient methods for determining diffusion coefficients with the increase in computing resources. Particularly, the nudged elastic band (NEB) and its modified version (climbing image algorithm) implemented in the first principle codes are well-proclaimed for accuracy in determining minimum energy paths and hence the diffusion coefficient.[12,13,13] NEB relies on harmonic transition state theory which uses quadratic approximations of the energy surface around saddle points where reaction intermediates situates.[14] Generally, reactions proceed through a minimum energy path (MEP), connected by saddle points in the potential energy terrain commonly known as images in NEB.[15] First-order saddle points, where energy is at a maximum along the MEP but at a minimum in all normal directions, determine

the activation energy which can be further solved for diffusion coefficient using Arrhenius equation.[16,17] One major drawback of NEB-based methods is the immense computational resources required when the system becomes sufficiently large and multi-component, as is typical in alloys.

To estimate of diffusion process using molecular dynamics within the assumption of a simple Lennard-Jones fluid model, the mean squared displacement (MSD) and velocity auto-correlation function (VACF) of ions can be used. The slope of the MSD versus time graph is proportional to the diffusion coefficient, while the time-integral of the VACF, in accordance with the Green–Kubo relation, is also proportional to the diffusion coefficient.[18] To ensure accurate results, it is crucial that this integral converges well. Furthermore, errors in the diffusion coefficient calculation can arise from deviations of the MSD from linear time dependence. To validate the accuracy in such instances, the diffusion coefficient can be cross-checked using the VACF where diffusion coefficient is estimated using Einstein's relation.[19] Simulations using the Lennard-Jones potential often modify to overcome its finite range nature.[20] These changes might not fully capture the diverse interactions as in alloy systems, potentially overstating the force between atoms.

Through various learning approaches, ML has already been established in material science to understand complex interdependencies among material parameters, thereby offering invaluable insights into tuning properties of materials for target-specific applications. Such approaches are particularly advantageous given the time-consuming process of experimentally understanding material properties. This opens a new dimension in alloy designing, enabling the identification and optimization of the most promising diffusive element for a given alloy. It also allows the fine-tuning of heat treatment and surface treatment processes for selective diffusive elements to enhance and tailor the alloy's characteristics for specific applications. By leveraging information on the host material's composition, as well as its physical and chemical attributes, the presented ML approach facilitates the estimation of diffusion parameters specific to a particular diffusive species. Subsequently, the acquired

knowledge of diffusion parameters can be applied to the design of alloys for various engineering applications. For instance, a study by Lü et al.[21] investigated the impact of a solution heat treatment at 693 K on Mg_2Si particles in Mg−Al−Si alloys. The study revealed that the treatment led to spheroidization of Mg_2Si particles due to Si atom diffusion along the Mg_2Si/Mg interface, resulting in superior mechanical properties for the alloys. Similarly in Titanium alloys with $(\alpha+\beta)$ dual phases, the diffusion of α-stabilizing elements (Al, O, etc.) and β-stabilizing elements (Mo, V, etc.) into corresponding phases during heat treatments is identified as a critical aspect of microstructural evolution, which in turn makes the alloy apt for aerospace and marine application.[22]

In general, predictive modeling in material science encompasses two main domains. The first domain revolves around the prediction of a material's mechanical properties, including fundamental factors like lattice parameters, lattice volume, density, and a range of elastic features. These properties play a crucial role in determining how a material responds to mechanical forces and deformations. For instance, the work of Li et al.[23] predicts lattice constants from the fundamental features of material composition. In a similar work, Peng et al.[24] examines the link between geometry parameters, relative densities, and range of lattice constants to predict mechanical and fatigue properties of lattice structures with different relative densities and crystalline systems. Another paper by Lee et al.[25] explores high-order Bézier curves to optimize lattice structures by using learning techniques, enhancing mechanical properties like modulus and strength while maintaining efficiency in load bearing and energy absorption. On the other hand, the second domain is mostly on predicting electrical and thermodynamic properties, encompassing key aspects like melting point, electrical conductivity, thermal conductivity, specific heat capacity, and diffusion parameters. Recent review articles have highlighted the optimum performance of ML-based models over analytical models in predicting thermal transport properties and advancing thermo-electric materials research.[26,27] Parallelly, similar studies are also accelerated in optimizing battery performance by targeting high-cycling efficiency, and durability with superior safety precau-

tions.[28–30] In the literature on ML-based diffusion studies, most research focuses on liquid or gas diffusion media. For instance, a study by Zhao et al.[31] utilized training data from molecular dynamics simulations to estimate the diffusion coefficients of binary and ternary supercritical water mixtures, employing neural networks with the aid of transfer learning. Similarly, other ML-based studies have predicted self-diffusion coefficients in pure liquids, Lennard-Jones fluids, and binary diffusion coefficients in gases.[32,33] A detailed literature survey reveals that while diffusion studies in fluid media are prevalent, there is a notable absence of ML-based diffusivity studies in solids, particularly in alloy media.[34–37]

Here, we introduce a novel ML-based computational framework – Machine Learned Diffusion Coefficient Estimator (ML - DiCE) that can predict the diffusion coefficient for two modes of diffusion in impure metallic (IM) media and three modes of diffusion in multi-component alloy (MCA) media with an accuracy above 90%. In particular, our model considers self and impurity diffusion modes in IM media, and self, impurity, and chemical modes of diffusion in MCA media. The choice of these models are based on three key criteria: 1) the quality of the data which ensures the diversity in DM and DE; 2) quantity of data that ensures the availability of sufficient number data points for a given ML algorithm to perform optimally; 3) variability in diffusion coefficient that aims for a lower standard deviation which results in reliably across different DM present in data. The model has been trained with a sufficiently large experimental dataset using the Random Forest(RF), Deep Neural Network (DNN) and Support Vector Regression (SVR) algorithms from the $Scikit-learn$ library. A separate featurization scheme was employed for diffusing elements and diffusion media that include fundamental atomic properties as well as composition level features of the diffusion media for multi-component diffusion media. Remarkably, our model encompasses 95% of elements from the periodic table, either as diffusing elements or part of the diffusion medium's composition, allowing it to predict diffusion coefficients across a wide verity of diffusion processes. The model performs optimally for a temperature range of 100–3500 K. Rigorous cross-testing against experimental data ensures the reliability and

accuracy of our predictions, establishing our model as a valuable tool in understanding and predicting diffusion processes.

The paper is structured as follows: Section Sec. 2 presents an overview of the data types used, data preprocessing, accessible experimental features, statistical considerations regarding the data, and the criteria employed for model selection. Section Sec. 3 delves into a comprehensive analysis of feature engineering alongside the training scheme. This section also includes an individual assessment of model performance, key feature extraction techniques, and an uncertainty analysis. In Sec. 4, we discuss a comparative analysis of model's performance with RF, DNN and SVR algorithms and potential bias and limitations of the model. Finally, the conclusions drawn and the model's accessibility are discussed in Section Sec. 5.

2 Materials and methods

2.1 Diffusion data

Input feature data : This study utilized diffusion data extracted from the material database popularly known as $MatNavi$, developed by the National Institute of Material Science, Japan.[38] $Kakusan$ is the diffusion database subset under $MatNavi$ material database that aims to encompass fundamental diffusion data of metallic and inorganic materials, primarily sourced from relevant literature references.[39,40] The dataset used for this study includes attributes such as the diffusion coefficient (m^2/s), temperature (K), diffusing element, diffusion mode, reference literature, and composition of alloy with weight percentages of constituent elements. The diffusing media presented in the dataset encompasses both multi-component alloys and impure metals, with purity quantified as a percentage. A comprehensive overview of the data collected for IM and MCA media is depicted in Figure 1. The choice selection of diffusion modes for modeling is based on the dataset size. For example, Figure 1a illustrates the IM media dataset, where the impurity and self diffusion

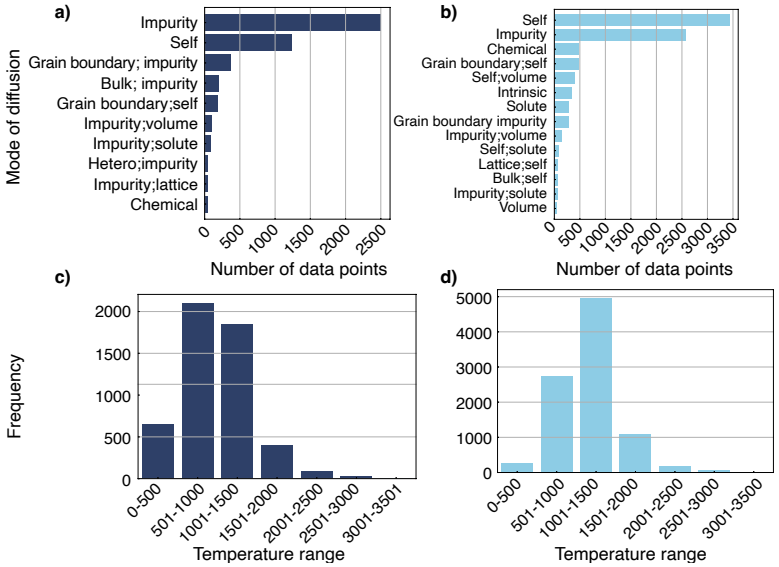

Figure 1: Dataset overview: a) diffusion mode present in dataset for IM media and b) MCA media. Frequency of temperature range of diffusion for c) IM media and d) MCA media.

modes comprise a larger number of diffusion data points (3431 and 2567, respectively) compared to other modes, thus making them suitable for training. Similarly, in MCA media (Figure 1b), self diffusion (3431 data points), impurity diffusion (2567 data points), and chemical diffusion (483 data points) modes have been selected due to the sufficient amount of data available. Some diffusion modes are ambiguously labeled in the dataset; for instance, 471 data points are categorized under a diffusion mode termed 'grain boundary, self' in MCA dataset. Despite the sufficient number of data points in such modes, the ambiguity in diffusion mode classification led to the exclusion of such modes of diffusion from further study. Figure 1c and d display the experimental temperature range, mostly between 500 and 1500 K, for calculating the diffusion coefficients in IM and MCA media respectively. The periodic tables given in Figure S1a and b highlight the diffusing elements considered for IM and MCA media respectively. The percentage contribution from each block is shown in the pie diagram given in the inset of the periodic table. The pie diagram illustrates the

distribution pattern of diffusing elements across both media types. It shows that a minority of diffusing elements originate from the s and f blocks, while the majority stem from the d and p blocks in both media. This visualization also facilitates the assessment of the model's applicability to specific diffusing elements and diffusion media, offering valuable insights for users.

Data preprocessing: The *MatNavi* database offers diffusion data along with essential descriptors such as temperature, diffusion mode, diffusion medium composition, and diffusion element for a diffusion coefficient. Specifically, in alloy compositions, it provides the constituent ratios of elements as percentages, while in elemental media, purity of the media is expressed as a percentage. Since most of the experimental data are based on isotope tracer method, corresponding diffusing elements were tagged with the information (mass number) of isotope used.

The data was initially categorized according to diffusion modes, with subsequent removal of rows containing missing data points. Additionally, we identified and eliminated extraneous alphanumeric characters within the diffusion media representation, along with isotope labels associated with diffusing elements. Similarly, the dataset was further treated to address outliers, missing values, and duplicate data. Before being input into the featurization algorithm, the diffusion system was subjected to representation alterations that guarantee the process of diffusing element X across AaBbCcDd (or Aa in the case of IM media) media at a specific temperature, where uppercase letters denote the element and lowercase letters indicate the corresponding percentage of the element. The overall preprocessing treatments are illustrated in a diagram given in Figure S3

Target data – the diffusion coefficient: The diffusion coefficient (D) presented in the

dataset is calculated based on Arrhenius representation of diffusivity given by Equation 1.

$$D = D_0 \exp\left(\frac{-Q}{RT}\right) \tag{1}$$

where D_0 is a temperature-independent constant, Q the activation energy for diffusion, R the gas constant and T the temperature. The statistical parameters of diffusion coefficient for respective models are given in Table 1. Prima facie, it can be seen that the diffusion

Table 1: Statistical parameters of target data – diffusion coefficient*.

Medium	Diffusion mode	Types of diffusion medium#	Number of data	Minimum ($\times 10^{-25}$)	Maximum ($\times 10^{-7}$)	Mean (μ) ($\times 10^{-10}$)	Standard deviation (σ) ($\times 10^{-9}$)	Coefficient of variation ($\frac{\sigma}{\mu}$) (%)
Impure metallic media	Impurity	33	2506	0.099	1800	1520	4220	3336.35
	self	32	1240	0.028	0.011	0.046	0.058	1260.25
Multi-component alloy media	self	191	3431	22.70	0.009	0.020	0.022	1126.92
	impurity	271	2567	0.230	6.840	9.080	15.601	1724.90
	chemical	14	483	5000	0.011	0.448	10.500	23.54

* Unit of Diffusion coefficient is given in m^2/s in the table.
# Number of different types of diffusion media present in data.

coefficient has extremely low order of magnitude ranging from $10^{-25} - 10^{-7}$ m^2/s in the overall data. Secondly, the aggregation of data points at certain frequencies is evident (Figure S4) because the diffusion coefficients corresponding to most of the alloys are collected for a widely varying range of temperatures from the experiment. This results in an elevated value for the coefficient of variation as given in Table 1. In particular, self-diffusion in IM media, and self and chemical diffusion in MCA media show higher variance compared to other modes of diffusion. Considering this higher variability in target data, log-transformed values of the diffusion coefficients were used for training. This study prefers the log-transformed diffusion coefficient as the target variable since it exhibits close proximity to the normal distribution behavior as seen in Figure S4, aiding in better performance of models.

3 Feature Engineering

Performance of predictive models in the ML domain is highly influenced by both the quantity and quality of data, as well as the choice of training algorithms. This study observes that the choice of descriptors plays a significant role in predicting diffusion coefficients. Descriptors encompass factors that have direct or indirect influence on diffusion coefficients and are therefore considered meticulously in this study. In light of this, we included a detailed set of descriptors separately for diffusing elements and diffusion media for predicting diffusion coefficients, acknowledging that certain descriptors may pertain specifically to the chemical and structural characteristics of the chemical species involved. In addition to this, the temperature corresponding to the diffusion is placed as an important feature considering its direct influence on the target variable.

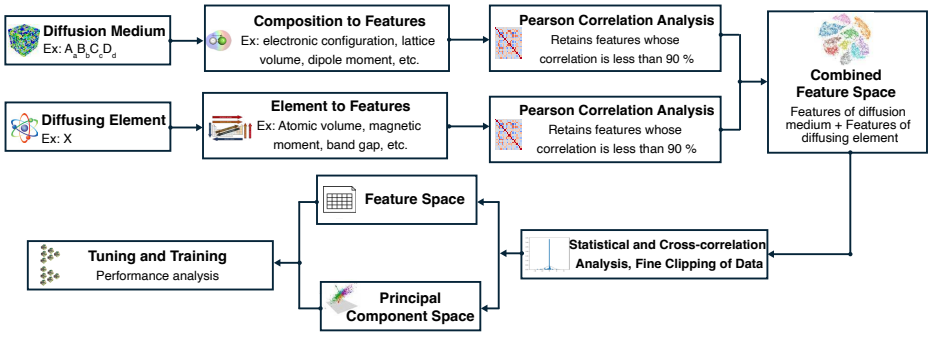

Figure 2: Workflow of modeling.

Featurization : Figure 2 presents an overview of our modeling approach through featurization, where we utilize the *Magpie* featurization preset available in the *Matminer* library to featurize the composition of the diffusion medium and the diffusing element. *Matminer*, an open-source Python library designed for material data mining, serves as a valuable resource in our featurization scheme and model development.[41] The *Magpie* featurization preset includes statistical computations for elemental attributes as mean, average deviation, range,

mode, minimum, and maximum.[42] A detailed overview of all such attributes along with the statistical operation considered is given in Table S1. When employing the *Magpie* algorithm to extract composition features, we considered descriptors weighted by the composition fraction of each element. However, this weighted scheme was omitted when featurizing the diffusion element. Consequently, two sets of descriptors were generated, one corresponding to the diffusion media and the other to the diffusing element. Then, these two sets of features underwent separate Pearson's correlation analyses implemented through *Pandas* – a Python library, wherein features exhibiting a correlation exceeding 90% were excluded.[43,44] For instance, in cases where two features displayed a correlation above 90%, we retained the feature with the greater Mean Absolute Deviation (MAD). Pearson's correlation ($r_{f1,f2}$) between feature-1 (f1) and feature-2 (f2), was computed using the Equation 2.

$$r_{f1,f2} = \frac{\sum \left(f1_i - \bar{f}1\right)\left(f2_i - \bar{f}2\right)}{\sqrt{\sum \left(f1_i - \bar{f}1\right)^2 \left(f2_i - \bar{f}2\right)^2}} \tag{2}$$

where $f1_i$ and $f2_i$ are the feature values, and $\bar{f}1$ and $\bar{f}2$ respective sample mean. Pearson's coefficient, a statistical metric, quantifies the magnitude and direction of a linear correlation between two continuous variables. This coefficient varies between -1 and 1, with 1 indicating a flawless positive linear relationship, -1 denoting a perfect negative linear correlation, and 0 signifying no linear association. The correlation-based filtering was applied separately to each set of features, the features of the diffusion medium and element were merged, and then another correlation analysis was performed in the combined feature space. However this time features with a correlation greater than 75% were omitted, by retaining the feature with the highest MAD ($= \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}|$ where n is the number of data and, x_i and $\bar{x}$ are data point and mean of the data respectively). This feature space is then considered for further analysis and training.

Feature space analysis: Scaled principal component analysis (PCA) based on the covari-

ance matrix was utilized to minimize the dimension of the data due to the large number of features. There are various ways to calculate the number of PCs.[45,46] In this study, the choice of the number of PCs to be retained was based on the explained variance. Finally, the training was performed using both decomposed feature space (PC space that explains 99% of variance in data) and pristine feature space for a comparative analysis of the model's performance.

Training scheme: Considering the size of data, number of features, and complexity among features, the RF and DNN have been chosen as the main algorithms for training purposes with hyperparameters optimized through *GridSearchCV* method implemented thorough *scikit-learn*[47] library. Here, a computationally intensive tuning scheme, incorporating a broad range of hyperparameters, is employed to maximize the R^2 and minimize the MSE scores to ensure the optimum performance of model. For a comparative study of performance of models based on the algorithm chosen, we also included the training and testing results of SVR. In all models, training and testing were carried out in segmented data intervals in order to determine the ideal range of target data that provides maximum model performance with a larger number of training data. Furthermore, we conducted 10 random trials to assess the stability of the model's performance across different test-train sampling sizes while maintaining a test-train ratio of 20% and 80% throughout.

Model evaluation: As a standard regression problem, the following three evaluation criteria have been used to compare the performance of all models, including the Root Mean Squared Error (RMSE), and coefficient of determination (R^2) calculated as follows.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(D_i^{exp} - D_i^{pred})^2} \tag{3}$$

$$R^2 = \left(\frac{\sum(D_i^{exp} D_i^{pred}) - \frac{\sum(D_i^{exp} D_i^{pred})}{n}}{\sqrt{\left[\sum D_i^{exp^2} - \frac{\left(\sum D_i^{exp^2}\right)}{n}\right]\left[\sum D_i^{pred^2} - \frac{\left(\sum D_i^{pred^2}\right)}{n}\right]}} \right)^2 \quad (4)$$

In addition to error estimation, we have also performed an uncertainty analysis in our modeling. Our findings demonstrate that the errors generated by the *RF* algorithm implemented in *LoLo* (an RF-centered machine learning library in Scala) library are well-calibrated.[48] As mentioned in Ling et al.[49]'s work, an ideally calibrated uncertainty estimate should have a particular relationship with the errors of an ML model. Specifically, the distribution of $r(x)/\sigma(x)$ where $r(x)$ is the normalized residuals of the prediction given by $r(x) = \frac{\hat{f}(x) - f(x)}{\sigma(x)}$ (here, $\hat{f}(x) - f(x)$ is the difference between the predicted and actual value) and $\sigma(x)$is the uncertainty of the prediction given by Equation 5,

$$\sigma(x) = \sqrt{\left(\sum_{i=1}^{n} max[\sigma_i^2(x), \omega]\right) + \tilde{\sigma}^2(x)} \quad (5)$$

where $\sigma_i^2(x)$ is the sample-wise variance at test point x due to training point i, ω is the noise threshold in the sample-wise variance estimates, and $\tilde{\sigma}(x)$ is an explicit bias function. The noise threshold is set to $\omega = min_i \sigma^2(x_i)$, the magnitude of the minimum variance over the training data as suggested by Ling et al.[49]. If the uncertainty estimates were perfectly well-calibrated and the samples in the data set were independently distributed, then the normalized residuals would follow a Gaussian distribution with zero mean and unit standard deviation.

3.1 Machine Learned Models of Diffusion Coefficient

3.1.1 Impurity diffusion in impure metallic media

In the impurity diffusion mode comprising 2506 data points, correlation analysis revealed that among the 19 features of diffusing element and 35 features of diffusion media show

correlations below 90%. The heatmaps illustrating the correlation between features after removing highly correlated features are given in Figure 3. Figure 3a describes the feature space of diffusing element, characterized by a predominantly negative correlation trend compared to the feature space of diffusion media shown in Figure 3 b, where a more positive correlation trend is evident on average, denoted by the red squares. Merging these two feature spaces yields a combined feature space, excluding features with correlations exceeding 75%. Specifically, three features were eliminated from diffusing element's feature space, and four from diffusion media's. The correlation heatmap of the combined feature space, depicted in Figure 3 c, indicates an average feature correlation of less than 25% after the final correlation correction. Finally, the *temperature* feature is integrated into the combined feature space, leading to a training dataset dimension of 2497 × 49 (comprising 48 diffusion features and temperatures). Notably, some rows with missing variables in the feature space have been omitted, which accounts for the dataset's dimension of 2497.

After analyzing the distribution of the target variable, the diffusion coefficient, it is apparent that the log-transformed diffusion coefficient exhibits a more normal distribution compared to the pure diffusion coefficient. In this model, using the log of the diffusion coefficient as the target variable proves beneficial (with a better $R2$ score). The Figure S4a illustrates this shift, with the mean and standard deviation for the pure diffusion coefficient being 1.15×10^{-7} m^2/s and 2.64×10^{-6} m^2/s, respectively. After applying the logarithm transformation, these values become 31.00 and 7.06, respectively. The log-transformed diffusion coefficient is more suitable for modeling due to its normalized distribution characteristics. Similarly, the feature, *temperature* also exhibits a normal distribution, with a mean of 1010.39 K and a standard deviation of 450.69 K.

PC analysis (Figure 4a) reveals that 25 PCs can represent 90% of the variance in the data, whereas 39 PCs can represent 99% of the variance. Notably, the number of PCs needed to represent at least 90 % variance is remarkably higher, indicating the presence of complex relationships among the descriptors.[50] Figure 4b depicts the important features

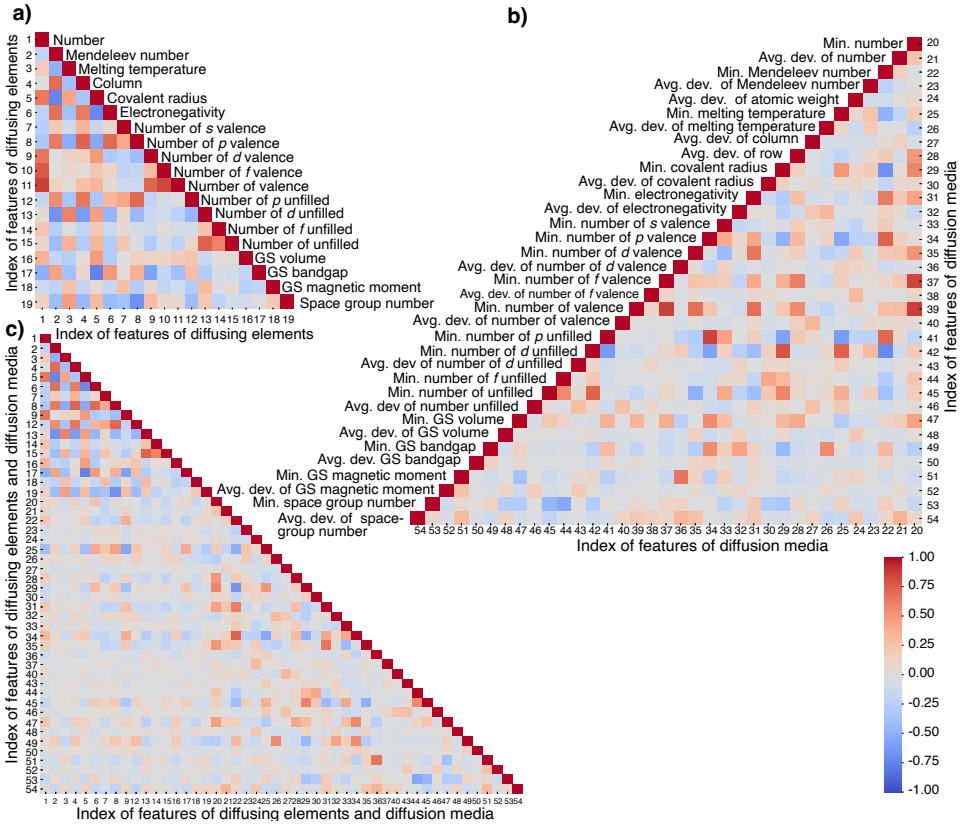

Figure 3: Feature spaces – impurity diffusion in IM media for a) diffusing elements, b) diffusion media, and c) the combined feature space used for training. Features of diffusing elements and diffusion media are indexed sequentially and the respective indices are marked in the combined feature space for reference. Common abbreviations such as 'min.' for minimum, 'avg.' for average, and 'dev.' for deviation are used.

when projecting eigenvectors of features in the PC space spanned by the first two PCs that together represent $\approx 24\%$ of variance in data. In this figure, red and blue arrows represent important features of diffusing elements and diffusion media, and other trivial features are marked by green arrows. Further, the features of diffusing element namely the *GS bandgap*, *electronegativity*, *number of p valance*, *number of p unfilled*, *Mendeleev number*, *column*, and *covalent radius* are identified as the most contributing features to the PC1 and PC2. In

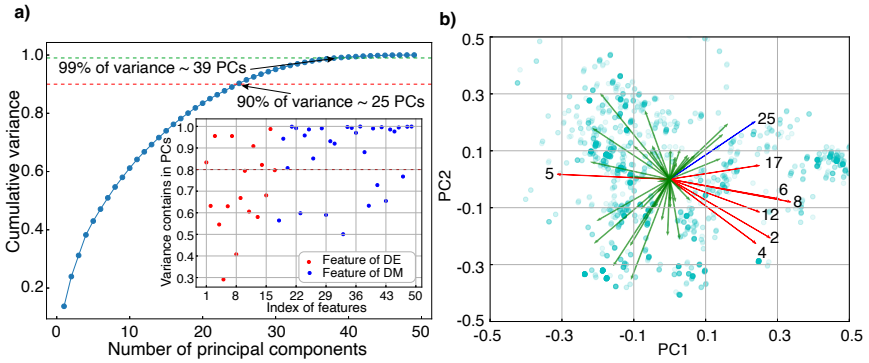

Figure 4: PC analysis; a) PCs that explain cumulative variance with inset graph that depicts variance of features represented by 25PCs and b) most contributed features in PC space spanned by first two PCs. The eigenvectors of features are projected as arrows. The important features of diffusing element are marked in red, diffusion media in blue, and other less contributing features in green color. The scatter plot in the background illustrates the scaled distribution of data points in PC space.

contrast, only one feature of diffusion media, the *min. melting temperature*, is identified as the most contributing feature in the same PC space. The decreased variance ($\approx 24\%$)of the data represented in PC 1 and 2 is the prime reason for this. However, when considering the magnitude of eigenvectors of features in 39-dimensional PC space, as given in the inset of Figure 4a, it can be seen that the majority of features exhibit a variance above 80%. Therefore, 39 PCs that explain 99% of the variance in data are used when training with PCs.

Figure 5a displays the model's performance when trained using feature space, yielding a R^2 score of 0.94; b, on the other hand, displays the model's performance when trained with PCs, yielding a R^2 value of 0.90. Although the feature-trained model outperforms PC-trained model in terms of the R^2 score, it is noteworthy that the PC-trained model exhibits a slightly lower MSE compared to the feature-trained model.

When examining the model's uncertainty as given in Figure S5a, both models show that the root mean square out-of-bag approach is not a well-calibrated metric; it significantly overestimates the error for a substantial portion of the data points. In contrast, employing

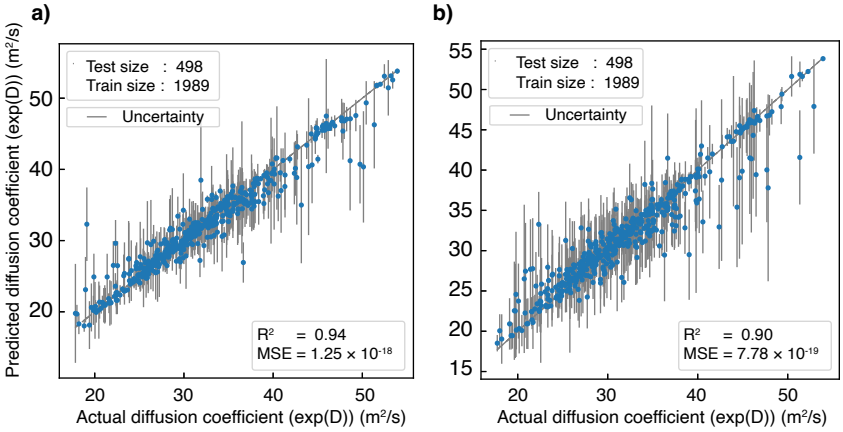

Figure 5: Performance of model – impurity diffusion in IM media: a) when trained with the entire feature space (dimension: 2497× 49) and b) with principal components of the entire feature space(dimension: 2497× 39). The gray solid lines depict the uncertainty associated with each data point.

the *Lolo* uncertainty approach results in comparatively well-calibrated uncertainty estimates (Figure S5b), indicating that the samples in the dataset are independently distributed. On comparing *Lolo* uncertainty estimates between feature-trained and PC-trained models as shown in Figure S5b, histograms are closer to normalized distributions of residuals in feature-trained rather than PC-trained models. Nevertheless, the *Lolo* uncertainty approach cannot comprehensively address all sources of uncertainty as seen from the small outgrowth of histogram in Figure S5b. This is particularly due to the uncertainties arising from factors not explained by the existing feature set—commonly referred to as 'unknown unknowns' as mentioned in Ling et al.[49]'s work. For example, the diffusion coefficient data is unavailable for certain temperature steps, leading to gaps in information. The absence of such information may slightly undermine the reliability of uncertainty estimates.

Upon examining the feature importance derived from the *RF* algorithm used during training, we noted that 16 out of the 49 features used hold significance based on their mean accuracy decrease score. These crucial features, along with their respective standard

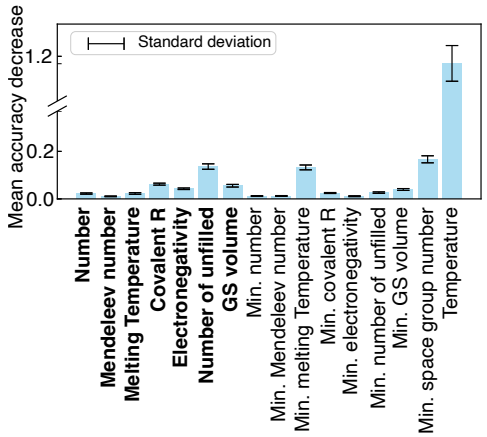

Figure 6: Feature importance – impurity diffusion in IM media: features of diffusing elements are highlighted in bold fonts and diffusion media are given in normal fonts.

deviations, are detailed in Figure 6. Notably, the *temperature* of the diffusion process emerges as the most pivotal feature among them. Furthermore, within the subset of seven important features pertaining to diffusion elements, the *number of unfilled orbitals* stands out as the most critical. Likewise, within the nine important features associated with diffusion media, both the *minimum melting temperature* and *GS volume* were identified as significant.

3.1.2 Self diffusion in impure metallic media

The dataset for self-diffusion in IM media has 2506 data points. 19 features of diffusing elements and 35 features of diffusion media were identified to have mutual correlations of less than 90% by correlation analysis. Figure 7a describes the feature space of diffusing element, characterized by negative and positive correlation trends in equal proportion compared to the feature space of diffusion media shown in Figure 7b, where a more positive correlation trend is evident on average, denoted by the reddish squares. After merging the feature spaces of diffusing element and diffusion media and removing features with correlations over 75%, one feature from diffusing element and sixteen features from diffusion media were eliminated.

The resulting correlation heatmap (Figure 7c) shows an average feature correlation below 25% after the final correction. Additionally, the *temperature* feature was incorporated into the combined space, resulting in training data of dimension 2506×30 (29 diffusion features and temperatures).

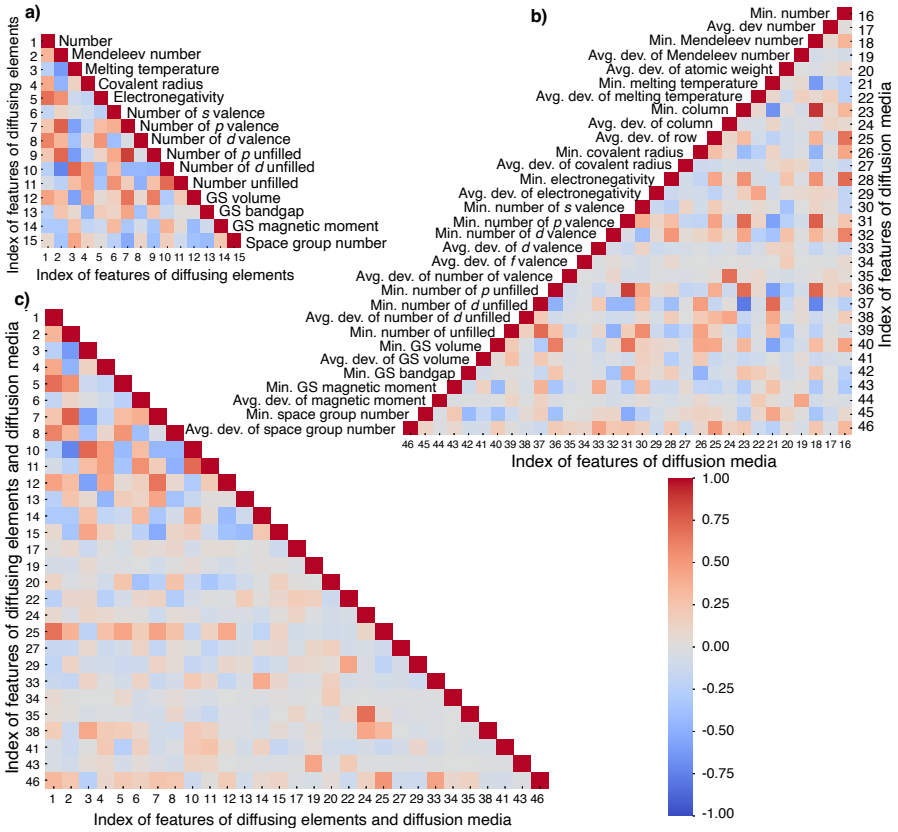

Figure 7: Feature spaces – self diffusion in IM media for a) diffusing elements, b) diffusion media, and c) the combined feature space used for training. Features of diffusing element and diffusion media are indexed sequentially and the respective indices are marked in the combined feature space for reference. Common abbreviations such as 'min.' for minimum, 'avg.' for average, and 'dev.' for deviation are used.

Statistical analysis reveals that the log-transformed diffusion coefficient displays a more normalized distribution compared to the original diffusion coefficient, leading to better mod-

eling performance (with an improved R^2 score). As illustrated in Figure S4b, the mean and standard deviation of the pure diffusion coefficient are 8.09×10^{-12} m^2/s and 7.68×10^{-11} m^2/s, respectively. Following the logarithmic transformation, these values shift to 34.90 m^2/s and 6.01 m^2/s. This transformation enhances modeling performance due to the normalized distribution trend. Similarly, the *temperature* feature also demonstrates a normal distribution, with a mean of 1027.98 K and a standard deviation of 500.71 K.

PC analysis (Figure 8 a) shows that 16 PCs capture 90% of the data's variance, while 24 PCs capture 99%. Notably, representing at least 90% variance requires significantly more PCs, indicating complex relationships among descriptors.[50] In Figure 8 b, important features are shown using eigenvectors projected onto the first two PCs, explaining 28% of the variance. Diffusing element's features such as *covalent radius*, *Mendeleev number*, *melting temperature*, *electronegativity*, *number of p*, and *d valence*, and *number d unfilled*, and *GS volume* contribute most to PC1 and PC2. Conversely, only one feature of diffusion medium,

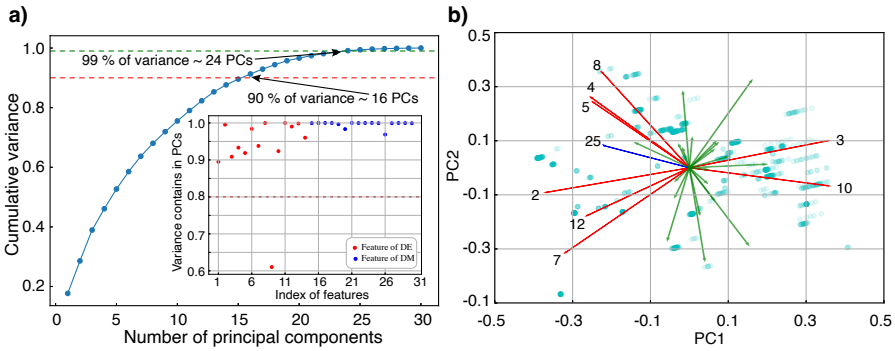

Figure 8: PC analysis; a) PCs that explain cumulative variance with inset graph that depicts variance of features represented by 24PCs and b) most contributed features in PC space spanned by first two PCs. The eigenvectors of features are projected as arrows. The important features of diffusing elements are marked in red, diffusion media in blue, and other less contributing features in green color. The scatter plot in the background illustrates the scaled distribution of data points in PC space.

the *average deviation of row*, is identified as the most contributing feature in the same PC space. The decreased variance of the data ($\approx 28\%$) represented by PC 1 and 2 is the prime

reason for this. However when considering the magnitude of eigenvectors of features in 24-dimensional PC space, as given in the inset of Figure 8a, it can be seen that the majority of features exhibit a variance above 80%. Therefore, 24 PCs that explain 99% of the variance in data are used when training with PCs.

Figure 9a shows the model's performance when trained with feature space, achieving an R^2 score of 0.95; in comparison, Figure 9b shows the model's performance when trained with PCs, resulting in an R^2 value of 0.91. Despite the feature-trained model outperforming the

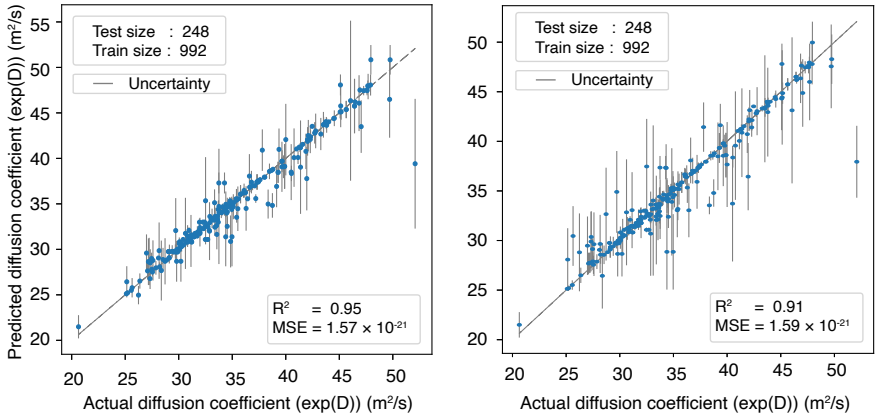

Figure 9: Performance of model – self diffusion in IM media: a) when trained with entire feature space (dimension: 2506 × 30) and b) with principal components of the entire feature space (dimension: 2506 × 24). The vertical gray solid lines depict the uncertainty associated with each data point.

PC-trained model in R^2 score, both models exhibit a very similar MSE score of approximately 1.5×10^{-21}.

When examining the model's uncertainty (Figure S5c), both models reveal that the root mean square out-of-bag approach is not well-calibrated; it tends to overestimate errors significantly for many data points. Conversely, using the *Lolo* uncertainty approach yields comparatively well-calibrated uncertainty estimates (Figure S5d), indicating independently distributed samples in the dataset. Comparing *Lolo* uncertainty estimates between feature-trained and PC-trained models (Figure S5d), histograms show closer-to-normalized distri-

butions of residuals in the feature-trained models than in the PC-trained ones. However, as mentioned in Sec. 3.1.1, the *Lolo* uncertainty approach cannot comprehensively address all sources of uncertainty, as evidenced by the histogram's outgrowth in Figure S5d. This is mainly due to the missing diffusion coefficient data for certain temperature steps, resulting in the overestimation of uncertainty.

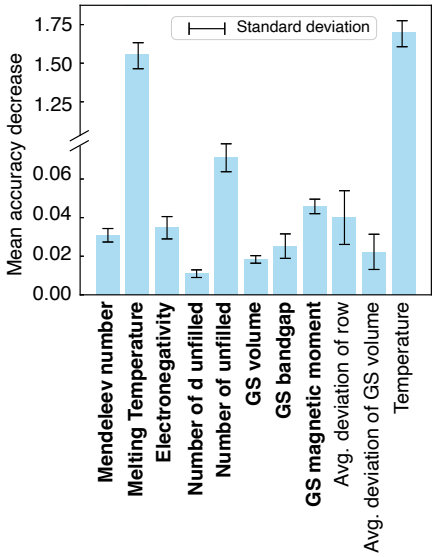

Figure 10: Feature importance – self diffusion in IM media: features of diffusing elements are highlighted in bold font and diffusion media are given in normal font.

When analyzing the feature importance derived from the *RF* algorithm used during training, we observed that 11 out of the 29 features used are significant based on their mean accuracy decrease score. These important features, along with their respective standard deviations, are detailed in Figure 10. As expected, the *temperature* of the diffusion process emerges as the most crucial feature among them. Furthermore, within the subset of eight important features related to diffusing elements, *melting temperature* and *number of unfilled orbitals* are highlighted as the most pivotal. Similarly, within the two important features

associated with diffusion media, both *average deviation of row* and *average deviation in GS volume* were identified as highly significant.

3.1.3 Self diffusion in multi-component alloys

In the context of self diffusion mode, with 3431 data points, 18 features of diffusing elements and 87 features of diffusion media have been identified that correlate less than 90%. The heatmap presented in Figure 11 visualizes the feature space following the post-correlation filtration. In Figure 11 a, diffusing element's feature space shows predominantly positive correlations, similar to diffusion medium's feature space depicted in Figure 11 b, where a positive correlation trend is also noticeable on average, highlighted by reddish squares. Then, the combined feature space was formed by excluding highly correlated features (over 75%), resulting in the elimination of 6 features from diffusing elements and 47 from diffusion media. The correlation heatmap of the combined feature space (Figure 11 c) indicates an average feature correlation of less than 25%. Finally, incorporating *temperature* into the combined feature space expanded the dataset to dimensions of 3431 $\times$ 54, comprising 53 diffusion features and *temperature*.

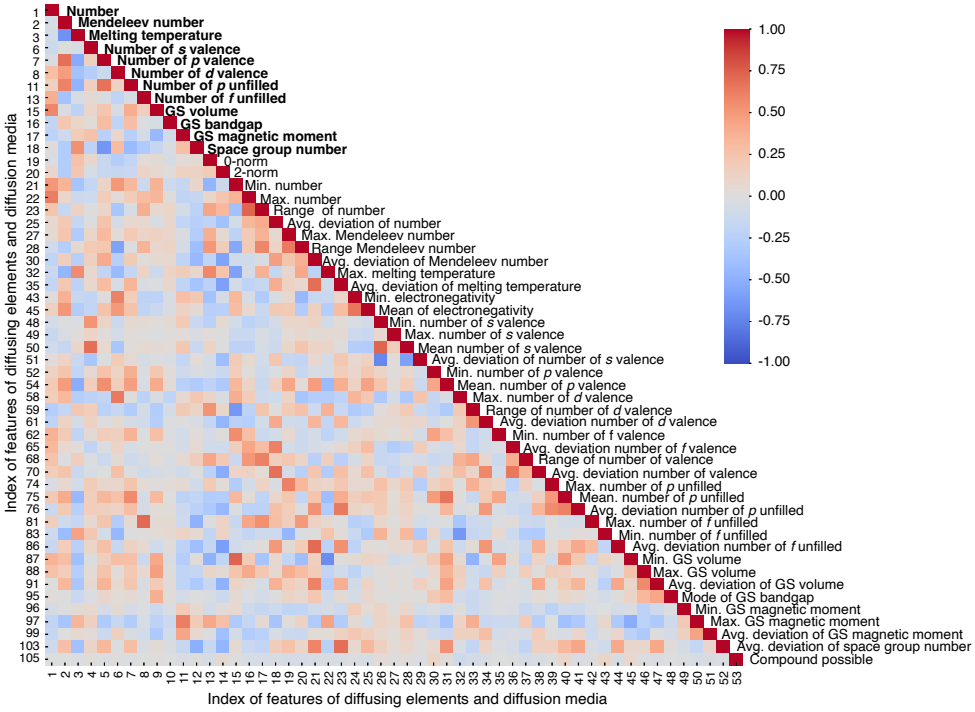

Figure 11: Feature space – self diffusion in MCA media: combined feature space of diffusing element and diffusion media used for training. Features of diffusing elements are highlighted in bold fonts and diffusion media in regular fonts(Individual feature space of diffusing element and diffusion media for this model are given in Figure S6a and b respectively). Common abbreviations such as 'min.' for minimum, 'avg.' for average, and 'dev.' for deviation are used.

A statistical examination of the distribution of the diffusion coefficient, reveals that log-transformed diffusion coefficient follows a more normal distribution compared to the normal diffusion coefficient. Employing the log of the diffusion coefficient as the target variable in this model results in a better R^2 score. Figure S4c visually depicts this transformation, with the mean and standard deviation for the normal diffusion coefficient being 3.27×10^{-12} m^2/s and 2.89×10^{-11} m^2/s, respectively. After the logarithm transformation, these values shift to 33.80 m^2/s and 5.37 m^2/s, respectively. The log-transformed diffusion coefficient

is favored for modeling because it closely aligns with normal distribution. Likewise, the *temperature* feature also exhibits a normal distribution, with a mean of 1234.19 K and a standard deviation of 374.52 K.

PC analysis (Figure 12 a) indicates that 21 PCs capture 90% of variance of data, while 40 PCs capture 99%. This disparity suggests complex relationships among descriptors, as more PCs are needed to represent at least 90% variance. In Figure 12 b, important features are displayed using eigenvectors projected onto the basis of first two PCs, together explaining approximately 28% of the total variance in data. diffusing element's features such as *Mendeleev*

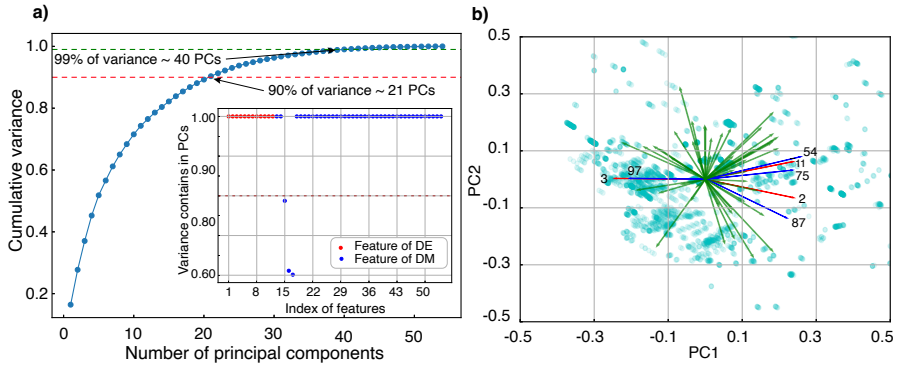

Figure 12: PC analysis – self diffusion in MCA media; a) PCs that explain cumulative variance with inset graph that depicts variance of features represented by 40 PCs and b) most contributed features in PC space spanned by first two PCs. The eigenvectors of features are projected as arrows. The important features of diffusing elements are marked in red, diffusion media in blue and other less contributing features in green color. The scatter plot in the background illustrates the scaled distribution of data points in PC space.

number, *melting temperature*, and *number of p unfilled orbitals* are significant contributors mainly to PC1. In the same PC space, important diffusion medium's features include *mean number of p valence*, *minimum GS volume*, and *maximum GS magnetic moment*. Given the fact that PC1 and PC2 collectively represent approximately 28% of the variance in the data, considering a higher-dimensional PC space that incorporates greater variance is recommended for training. The scatter plot given in the inset of Figure 12a suggests that training with 40 PCs is beneficial as most features capture above 90% variance.

This model exhibits overfitting when trained with the feature space using RF. This is primarily due to the high diversity in diffusion media within the self diffusion dataset, which comprises 191 distinct types of diffusion media, as given in Table 1. Therefore we choose *MLPRegressor* to train the model using feature space that gives R^2 score of 0.93 with MSE 2.48×10^{-22}. Training was also performed using 40 PCs that captured 99% of variance in training data. Figure 13b illustrates the model's performance when trained using the PC space with *RF*, achieving an R^2 score of 0.92 with MSE 1.65×10^{-22}. These results are very similar to those obtained from the *MLPRegressor* using the feature space. This indicates that despite the high diversity in the feature space leading to *RF* overfitting, projecting the features to PC space and then using PCs for *RF* training yields accurate results and precise uncertainty estimates.

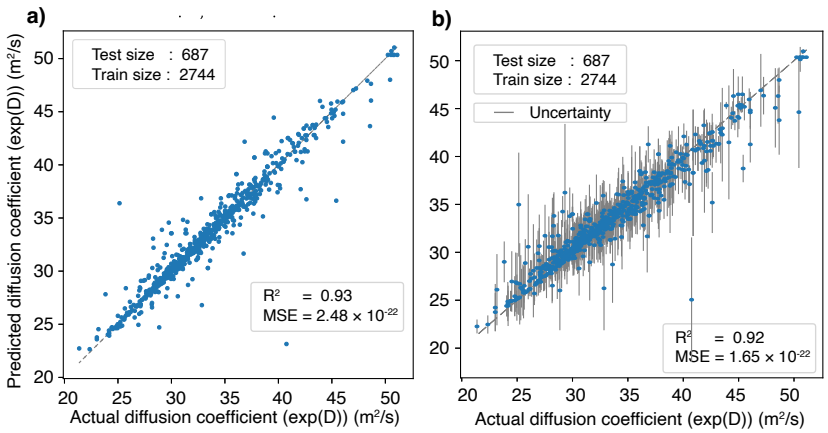

Figure 13: Performance of model – self diffusion in MCA media when trained with a) the entire feature space (dimension: 3431 × 54) and with b) the principal components of the entire feature space (dimension: 3431 × 40). The gray solid lines depict the uncertainty associated with each test data point.

On comparing out-of-bag uncertainty estimates calculated using feature space (trained with *MLPRegressor* since *RF* is overfitting) and PC space (trained with *RF*) as given in Figure S5e it can be seen that both training schemes overestimates errors for many data points. However using *Lolo* approach, as given in Figure S5f the uncertainty estimates are

more normally distributed for most of the data points.

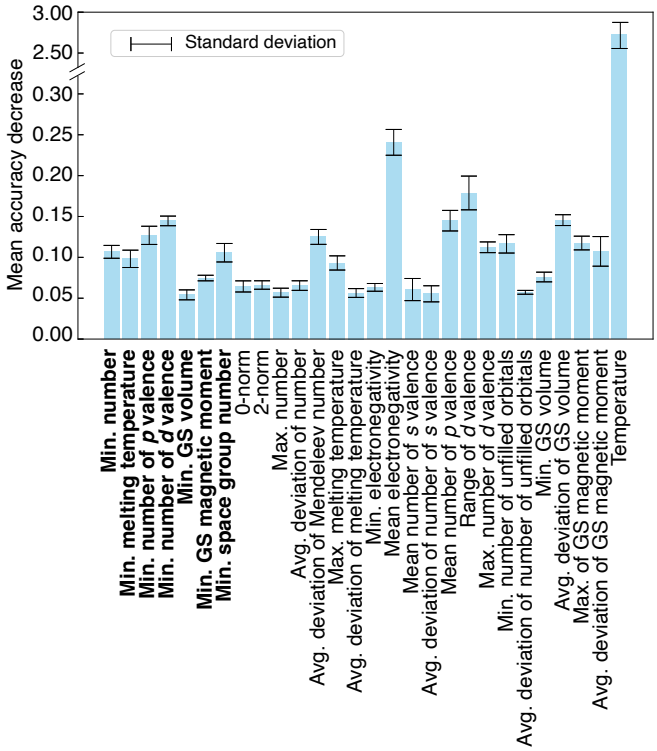

Figure 14: Feature importance – self diffusion in MCA media: *MLPRegressor* were used with feature permutation scheme. Features of diffusing elements are highlighted in bold font and diffusion media are given in normal font.

Figure 14 summarizes the feature importance analysis calculated using *MLPRegressor* based on feature permutation. Among the six key features of diffusing elements, *minimum number of p* and *d valence* were identified as most important. For diffusion media, out of twenty important features, *mean electronegativity* and *range of d valence* of the composition elements were highlighted as most important. Furthermore, *temperature* of diffusion process stands out as the most important feature among all other features.

3.1.4 Impurity diffusion in multi-component alloys

In the impurity diffusion mode, encompassing 2567 data points, 19 features of diffusing elements and 85 features of diffusion media were identified to correlate less than 90%. As

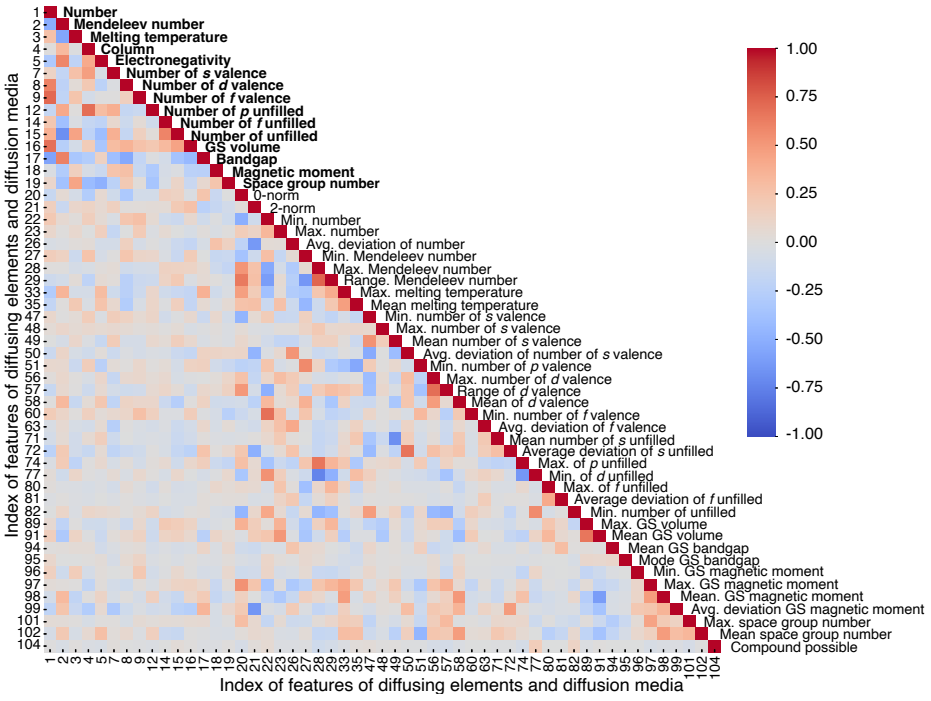

Figure 15: Feature space – impurity diffusion in MCA media: combined feature space used for training. Features of diffusing elements are highlighted in bold characters and diffusion media in regular fonts(Individual feature space of diffusing element and diffusion media for this model are given in Figure S7a and b respectively). Common abbreviations such as 'min.' for minimum, 'avg.' for average, and 'dev.' for deviation are used.

depicted in Figure S7a, diffusing element's feature space predominantly exhibits negative correlations, contrasting with diffusion media's feature space shown in Figure S7b, where a positive correlation trend is evident on average, highlighted by reddish squares. Subsequently, the combined feature space was created by excluding highly correlated features (over 75%), resulting in the removal of 4 features from diffusing elements and 47 from diffusion media.

The correlation heatmap of the combined feature space (Figure 15c) indicates an average feature correlation of less than 25%. Finally, integrating *temperature* into the combined feature space expanded the dataset to dimensions of 2567 × 49, comprising 50 diffusion features and temperatures.

To enhance the model performance during training, the diffusion coefficient is transformed to a log scale, resulting in a more normal distribution and a better R^2 score. The mean and standard deviation for the original diffusion coefficient are 1.66×10^{-9} m^2/s and 1.98×10^{-6} m^2/s, respectively, while after the logarithm transformation, these values shift to 32.90 m^2/s and 8.53 m^2/s, respectively as shown in Figure S4d. Similarly, the *temperature* feature exhibits a normal distribution with a mean of 1049.33 K and a standard deviation of 422.75 K.

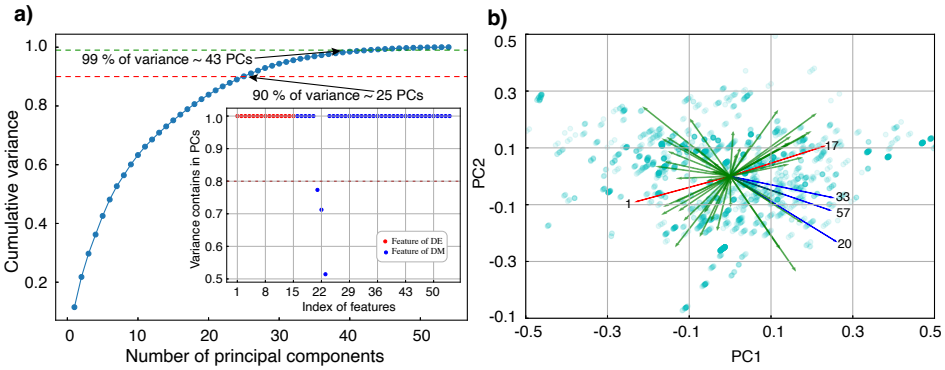

Figure 16: PC analysis – impurity diffusion in MCA media; a) PCs that explain cumulative variance with inset graph that depicts variance of features represented by 43PCs and b) most contributed features in PC space spanned by first two PCs. The eigenvectors of features are projected as arrows. The important features of diffusing elements are marked in red, diffusion media in blue, and other less contributing features in green color. The scatter plot in the background illustrates the scaled distribution of data points in PC space.

PC analysis (Figure 16a) reveals that 25 PCs capture 90% of the data's variance, while 43 PCs capture 99%. This discrepancy indicates complex relationships among descriptors, requiring more PCs to represent at least 90% variance. In Figure 16b, important features are displayed using eigenvectors projected onto the first two PCs, explaining approximately

22% of the total variance in data. Diffusing element's features such as *number* and *band gap* significantly contribute to PC1. Similarly, notable diffusion media's features include *0-norm*, *maximum melting temperature*, and *range of d valence* in the same PC space. Considering that PC1 and PC2 collectively represent only 22% of the variance in the data, we selected a 43-dimensional PC space that captures 99% of the variance for PC-based training, as shown in the inset of Figure 16a. In this space, most features contribute over 80% variance ensuring the effective inclusion of all of the features.

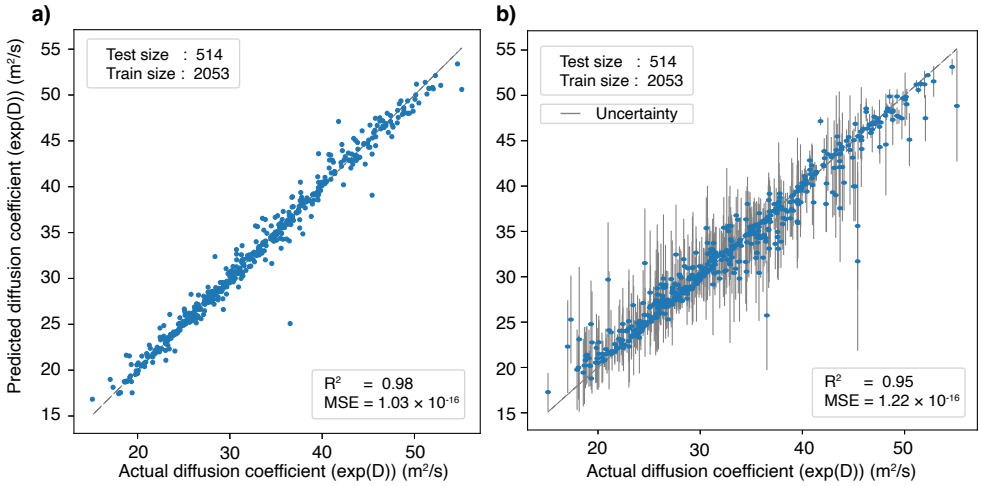

Figure 17: Performance of model – impurity diffusion in MCA media when trained with a) the entire feature space (dimension: 2567 × 49) and b) principal components of the entire feature space (dimension: 2567 × 43). The vertical gray solid lines depict the uncertainty associated with each data point.

Similar to the self diffusion model in MCA media described in Sec. 3.1.3, this model also displays overfitting when trained with the feature space, mainly because of the highly diverse diffusion media (271 types) present in the data. For this reason, we used the *MLPRegressor* for training with the feature space, achieving a higher R^2 score of 0.98, as shown in Figure 17a. However, *RF* performs well when trained using 43 PCs but a comparatively lower R^2 score of 0.95 as depicted in Figure 17b. When examining the MSE of both training schemes, the

differences are minimal, with MSE values on the order of 10^{-16}.

When comparing out-of-bag uncertainty estimates calculated using the feature space (trained with *MLPRegressor* since *RF* is overfitting) and the PC space (trained with *RF*) as shown in Figure S5g, it is evident that both training methods tend to overestimate errors for many data points. In contrast, the *Lolo* approach, depicted in Figure S5h, provides uncertainty estimates that are more normally distributed across most data points.

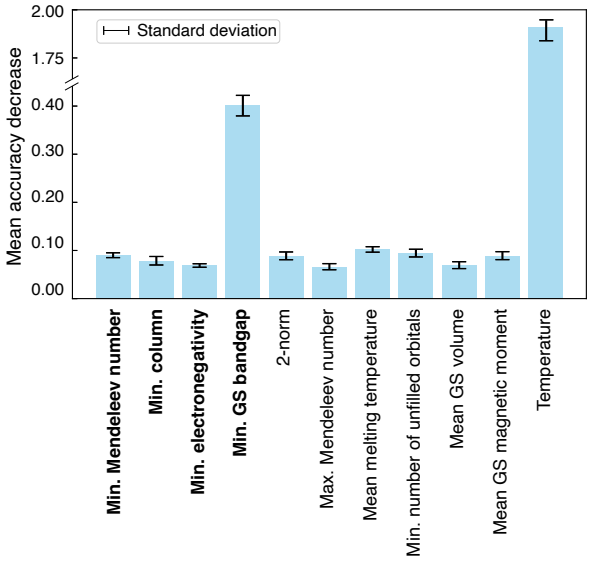

Figure 18: Feature importance – impurity diffusion in MCA media: *MLPRegressor* were used with feature permutation scheme. Features of diffusing elements are highlighted in bold font and diffusion media are given in normal font.

Figure 18 summarizes the feature importance analysis calculated using *MLPRegressor* based on feature permutation. Among the four key features of the diffusing elements, *minimum GS bandgap* was identified as the most significant. For diffusion media, out of six important features, *mean melting temperature* and *minimum number of unfilled orbitals* of composition elements were highlighted as the most important. Further, *temperature* of diffusion process was identified as the most influential feature overall.

3.1.5 Chemical diffusion in alloys

The dataset for self-diffusion in IM media comprises 483 data points. From correlation analysis, 9 features of diffusing elements and 45 features of diffusion media were identified with correlations below 90%. Figure S8a depicts the feature space of diffusing elements, showcasing mainly positive correlation trends. Similarly, Figure S8b illustrates the feature space of diffusion media, also with predominantly positive correlations, indicated by the reddish squares. Upon merging these feature spaces and filtering out features with correlations exceeding 75%, 5 features from diffusing elements and 30 features from diffusion media were removed. The resulting correlation heatmap (Figure 19) exhibits an average feature correlation below 25% after these adjustments. Lastly, incorporating the *temperature* feature into

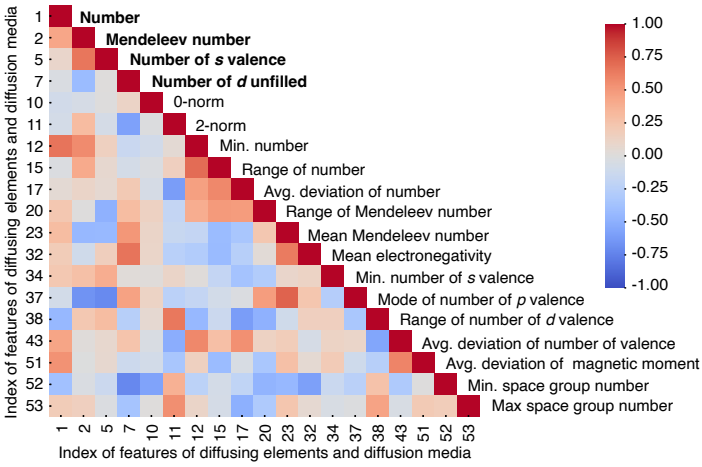

Figure 19: Feature space – chemical diffusion in MCA media: combined feature space used for training. Features of diffusing elements are highlighted in bold characters and diffusion media in regular fonts(Individual feature space of diffusing element and diffusion media for this model are given in Figure S8a and b respectively). Common abbreviations such as 'min.' for minimum, 'avg.' for average, and 'dev.' for deviation are used.

the combined space led to training data dimensions of 483 × 20 (19 diffusion features and *temperature*).

The statistical analysis indicates that the log-transformed diffusion coefficient exhibits

a more normalized distribution compared to the original diffusion coefficient. This normalization leads to improved modeling performance, reflected in an enhanced R^2 score. In Figure S4e, the log-transformed standard deviation graph displays two bell-shaped curves, indicating higher data variability, mostly symmetric around the mean. Specifically, for the pure diffusion coefficient shown in Figure S4e, the mean and standard deviation are 5.83×10^{-11} m^2/s and 1.17×10^{-10} m^2/s, respectively. After logarithmic transformation, these values shift to 30.01 m^2/s and 5.59 $m^2 s$, respectively. This transformation contributes to improved modeling performance due to the normalized distribution trend. Additionally, the *temperature* feature also demonstrates a normal distribution, with a mean of 1264.02 K and a standard deviation of 426.21 K.

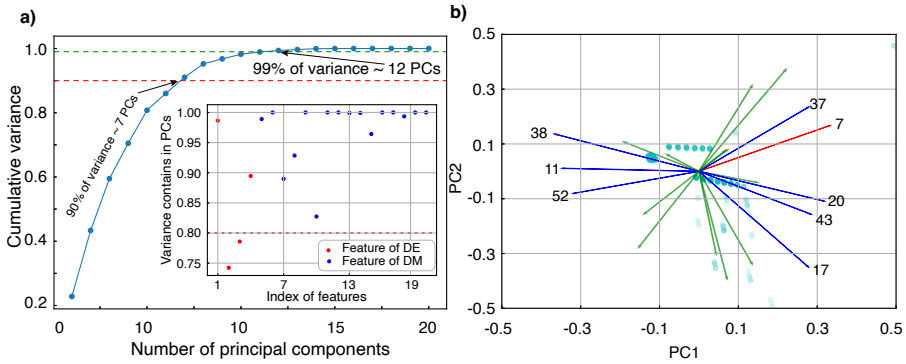

Figure 20: PC analysis; a) PCs that explain cumulative variance with inset graph that depicts variance of features represented by 12PCs and b) most contributed features in PC space spanned by first two PCs. The eigenvectors of features are projected as arrows. The important features of diffusing elements are marked in red, diffusion media in blue, and other less contributing features in green color. The scatter plot in the background illustrates the scaled distribution of data points in PC space.

The PC analysis (Figure 20 a) reveals that 7 PCs capture 90% of the data's variance, while 12 PCs capture 99%. It is noteworthy that achieving a representation of at least 90% variance necessitates a considerably higher number of PCs, indicating complex relationships among descriptors. In Figure 20 b, important features are depicted using eigenvectors projected onto the first two PCs, explaining 43% of the variance. Notably, a single feature of diffusing

element, the *number of d unfilled orbitals*, appears as the strongest feature which is more leaned to PC1. And, features of diffusion media, such as the *2-norm*, *average deviation of number*, *range of Mendeleev number*, *mode of the number of p valence*, *range of the number of p valence*, *average deviation of valence*, and *minimum space group number*, are identified as the most contributing features in the same PC space. However, when considering the magnitude of eigenvectors of features in the 12-dimensional PC space that includes higher variance of data, as shown in the inset of Figure 20a, it is visible that the majority of features exhibit a variance above 80%. Therefore, 24 PCs that explain 99% of the variance in the data are used during training with PCs.

Figure 21a shows the model's performance when trained using the feature space, resulting in an R^2 score of 0.94. Conversely, Figure 21b displays the model's performance when trained with PCs, achieving an R^2 value of 0.96. Comparing these models, the PC-trained model slightly outperforms the feature-trained model, despite similar MSE scores.

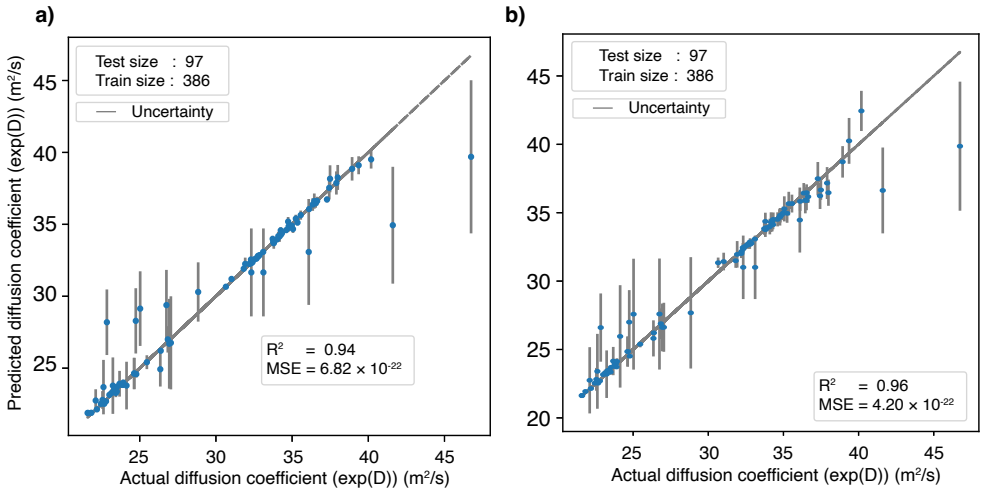

Figure 21: Performance of model – chemical diffusion in MCA media: a) when trained with the entire feature space (dimension: 483 × 20) and b) with principal components of the entire feature space (dimension: 483 × 12). The gray solid lines depict the uncertainty associated with each data point.

When examining the model's uncertainty (Figure S5k), both models demonstrate that the root mean square out-of-bag approach is poorly calibrated, tending to overestimate errors significantly for many data points. In contrast, using the *Lolo* uncertainty approach yields comparatively well-calibrated uncertainty estimates (Figure S5l), indicating independently distributed samples in the dataset. Comparing *Lolo* uncertainty estimates between feature-trained and PC-trained models (Figure S5l), histograms show closer-to-normalized distributions of residuals in the feature-trained models than in the PC-trained ones. However, as mentioned in Sec. 3.1.1 and Sec. 3.1.2, the *Lolo* uncertainty approach cannot comprehensively address all sources of uncertainty, as evidenced by the histogram's outgrowth in Figure S5l. This is primarily due to missing diffusion coefficient data for certain temperature steps.

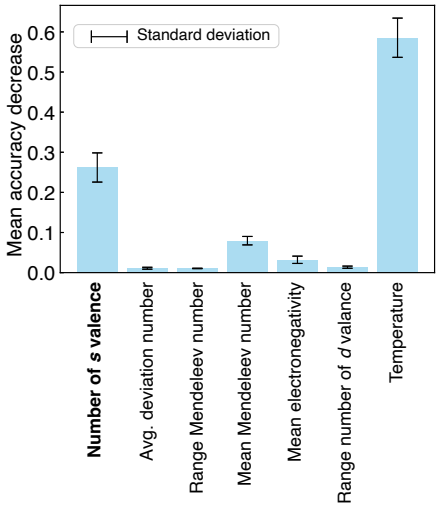

Figure 22: Feature importance – chemical diffusion in MCA media: Features of diffusing elements are highlighted in bold font and diffusion media are given in normal font.

When analyzing the feature importance derived from the *RF* used during training, we observed that 7 out of the 20 features used for training are significant based on their mean accuracy decrease score. These important features and their respective standard deviations

are detailed in Figure 22. Notably, the *temperature* of the diffusion process emerges as the most crucial feature among them. However, only one feature related to diffusing elements, the *number of s valence* is identified as important. Similarly, within the subset of important features associated with diffusion media, both *mean Mendeleev number* and *mean electronegativity* were identified as significant.

4 Performance Analysis of Models

This section includes a comparative analysis of performance of models trained with RF, DNN and SVR algorithms. All models are trained with the methodology described in section 3 with corresponding feature space as described in sections 3.1.1 – 3.1.5. When evaluating the model's performance using R^2 and MSE scores, the results indicate that RF and DNN trained models outperform SVR models. Specifically, RF models demonstrate optimal performance in predicting impurity and self-diffusion in IM media and DNN models for impurity, self and chemical diffusion in MCA media, as given in Table 2. The same table also highlights the overfitting of RF models as the number of features increases in the case of impurity and self diffusion in MCA media. Since the highly correlated features of the DE and DM were

Table 2: Performance overview of all models.

Diffusion Medium	Model	Number of features	R^2			MSE ($\times 10^{-19}$)		
			RF	DNN	SVR	RF	DNN	SVR
Impure metallic	Impurity	49	0.94	0.93	0.88	12.50	387.90	15.40
	Self	30	0.95	0.92	0.87	0.015	0.0480	0.047
Multi-component alloy	Impurity	54	*	0.98	0.97	*	1030.1	1363.4
	Self	54	*	0.96	0.90	*	0.0025	0.0021
	Chemical	20	0.94	0.96	0.93	0.0068	0.0092	0.0047

* Model is overfitting.

removed during the feature engineering step, none of the features from the final feature space removed to negotiate hyperparameter tuning time. This approach maximizes the model's

predictive performance, as all features in the final feature space contribute, either directly or indirectly, to the model's efficiency.

The prediction bias and limitations of a ML model can be understand through the nature and diversity of the training data along with R^2, MAE score and uncertainty estimates. Figure S1 and Figure S2 provide a comprehensive overview of the nature of the diffusion data used for training. The pie charts in Figure S1 clearly illustrate that nearly 40% of DEs are from the d-block for both IM and MCA media, which explains the model's higher efficiency in predicting d-block elements as DE. Similarly, the average percentage contributions from the s, p, and f blocks are 15%, 32%, and 11%, respectively, for DE, leading to a corresponding bias in prediction accuracy. Furthermore, as shown in Figure S2, the majority of DM is also composed of d-block elements (frequency of specific elements is also given in the same figure), which results in the model being more inclined towards the accurate predictions of diffusion coefficient for DM composed of d-block elements.

5 Conclusions

The exploration of five ML diffusion models presented in this article underscores the paramount importance of fundamental atomic feature of diffusing element and diffusion medium in governing the diffusion process. While traditional empirical approaches focus on relating flux gradients and temperatures to estimate diffusion coefficients, our study illuminates the potential of a holistic modeling scheme that meticulously considers the atomic environment of the diffusion process.

This ML-framework incorporates most of the crucial elements from the periodic table, either as diffusion elements or diffusing media. This enhances the diversity and inclusion of atomic species, resulting in a versatile model capable of accurately predicting diffusion coefficients for a wide range of diffusing elements and media. By employing two training approaches–using feature space and principal component space–we observed that the variance

in atomic descriptors influences the prediction of diffusion coefficients.

This study also emphasizes the benefits of applying a deep neural network for training, particularly when the feature space exhibits higher diversity. This is evidenced by models for self-diffusion and impurity diffusion in multi-component alloy media. The complexities arising from diverse multi-component alloys necessitate a sophisticated feature space, which challenges the efficiency of the *Random Forest* algorithm and advocates for the reliance on neural networks. However, employing principal component analysis is an effective alternative when dealing with a complex feature space. PC analysis captures the variance in the feature space through a reduced number of components, enabling the RF algorithm to perform better, although not as efficiently as neural networks.

When looking into the important features that evolved from the five diffusion models, we observed that features of diffusing elements mainly, *atomic radius*, *atomic volume*, *number of unfilled orbitals*, *electronegativity*, *band gap*, *magnetic moment* and *melting temperature* are significant features. On the other hand, when it comes to diffusion media, *mean Mendeleev number*, *minimum space group number*, *maximum number of s valence*, *mode of GS bandgap*, *minimum GS magnetic moment* and *minimum electronegativity* of elements present in the composition of alloys are significant. In addition to this, *temperature* of the diffusion process stands out as one of the important features in all of the models.

In conclusion, the diffusion framework, ML-DiCE, comprising models for impurity ($R^2 = 0.94$) and self diffusion ($R^2 = 0.95$) modes in IM media, as well as self ($R^2 = 0.96$), impurity ($R^2 = 0.98$), and chemical ($R^2 = 0.96$) diffusion modes in MCA media, demonstrates superior performance with optimum R^2 scores and minimal prediction error and uncertainty. We also reaffirm that when the feature space becomes excessively large and exhibits complex relationships, employing PC analysis can effectively capture variance thus enabling training without overfitting in tree-based regression models. Our framework, alongside our user-friendly open-source code, can be accessed at https://github.com/yanqingsu/ML-DiCE. This resource facilitates the estimation of diffusion coefficients and fine-tuning of the material

properties of diffusion media and elements to achieve the desired diffusion coefficient when designing alloys.

6 Supplementary Material

See the supplementary material for further details on model feasibility, data preprocessing steps, feature information, standard deviation of target variable and temperature, uncertainty analysis, and correlation analysis.

Acknowledgement

This work used Bridges-2 at the Pittsburgh Supercomputing Center through allocation MAT220034 from the Advanced Cyberinfrastructure Coordination Ecosystem: Services & Support (ACCESS) program, which is supported by National Science Foundation grants #2138259, #2138286, #2138307, #2137603, and #2138296. YR acknowledges support from the National Science Foundation under grant #2045084. The support and resources from the Center for High-Performance Computing at the University of Utah are gratefully acknowledged. ASK and YS thank the National Science Foundation under Grant No. 2345709. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Conflict of interest

The authors declare no competing financial interest.

References

(1) Fick, A. Ueber Diffusion. *Ann. Phys.* **1855**, *170*, 59–86.

(2) Junhua, K.; Lin, Z.; Bin, G.; Pinghe, L.; Aihua, W.; Changsheng, X. Influence of Mo content on microstructure and mechanical properties of high strength pipeline steel. *Mater. Des.* **2004**, *25*, 723–728.

(3) Garrett, R. P.; Lin, J.; Dean, T. A. An investigation of the effects of solution heat treatment on mechanical properties for AA 6xxx alloys: experimentation and modelling. *Int. J. Plast.* **2005**, *21*, 1640–1657.

(4) Zhong, W.; Zhao, J.-C. First experimental measurement of calcium diffusion in magnesium using novel liquid-solid diffusion couples and forward-simulation analysis. *Scr. Mater.* **2017**, *127*, 92–96.

(5) Kulkarni, N. S.; Bruce Warmack, R. J.; Radhakrishnan, B.; Hunter, J. L.; Sohn, Y.; Coffey, K. R.; Murch, G. E.; Belova, I. V. Overview of SIMS-Based Experimental Studies of Tracer Diffusion in Solids and Application to Mg Self-Diffusion. *J. Phase Equilib. Diffus.* **2014**, *35*, 762–778.

(6) O'Leary, M. H. Measurement of the isotope fractionation associated with diffusion of carbon dioxide in aqueous solution. *The Journal of Physical Chemistry* **1984**, *88*, 823–825.

(7) Matano, C. On the Relation between the Diffusion-Coefficients and Concentrations of Solid Metals. *Japan. J. Phys* **1933**, *8*, 109.

(8) Kirkaldy, J. S. DIFFUSION IN MULTICOMPONENT METALLIC SYSTEMS. *Can. J. Phys.* **2011**,

(9) Onsager, L. Reciprocal Relations in Irreversible Processes. I. *Phys. Rev.* **1931**, *37*, 405–426.

(10) Onsager, L. Reciprocal Relations in Irreversible Processes. II. *Phys. Rev.* **1931**, *38*, 2265–2279.

(11) Muralikrishna, G. M.; Tas, B.; Esakkiraja, N.; Esin, V. A.; Kumar, K. C. H.; Golovin, I. S.; Belova, I. V.; Murch, G. E.; Paul, A.; Divinski, S. V. Composition dependence of tracer diffusion coefficients in Fe–Ga alloys: A case study by a tracer-diffusion couple method. *Acta Mater.* **2021**, *203*, 116446.

(12) Classical And Quantum Dynamics In Condensed Phase Simulations: Proceedings ... - Google Books. 2024; https://www.google.com/books/edition/Classical_And_Quantum_Dynamics_In_Conden/UoTVCgAAQBAJ?hl=en&gbpv=1&pg=PR5&printsec=frontcover, [Online; accessed 9. Apr. 2024].

(13) Henkelman, G.; Uberuaga, B. P.; Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *J. Chem. Phys.* **2000**, *113*, 9901–9904.

(14) Vineyard, G. H. Frequency factors and isotope effects in solid state rate processes. *J. Phys. Chem. Solids* **1957**, *3*, 121–127.

(15) Voter, A. F.; Doll, J. D. Transition state theory description of surface self-diffusion: Comparison with classical trajectory results. *J. Chem. Phys.* **1984**, *80*, 5832–5838.

(16) Arnaldsson, A. Calculation of quantum mechanical rate constants directly from ab initio atomic forces. *University of Washington* **2007**,

(17) Kulathuvayal, A. S.; Su, Y. Ionic Transport through the Solid Electrolyte Interphase in Lithium-Ion Batteries: A Review from First-Principles Perspectives. *ACS Applied Energy Materials* **2023**, *6*, 5628–5645.

(18) Frenkel, D.; Smit, B. *Understanding molecular simulation : from algorithms to applications. 2nd ed*; 1996; Vol. 50.

(19) Einstein, A. Über die von der molekularkinetischen Theorie der Wärme geforderte Be-

wegung von in ruhenden Flüssigkeiten suspendierten Teilchen. *Ann. Phys.* **1905**, *322*, 549–560.

(20) Wang, X.; Ramírez-Hinestrosa, S.; Dobnikar, J.; Frenkel, D. The Lennard-Jones potential: when (not) to use it. *Phys. Chem. Chem. Phys.* **2020**, *22*, 10624–10633.

(21) Lü, Y. Z.; Wang, Q. D.; Zeng, X. Q.; Zhu, Y. P.; Ding, W. J. Behavior of Mg–6Al–xSi alloys during solution heat treatment at 420°C. *Mater. Sci. Eng., A* **2001**, *301*, 255–258.

(22) Huang, S.; Zhang, J.; Ma, Y.; Zhang, S.; Youssef, S. S.; Qi, M.; Wang, H.; Qiu, J.; Xu, D.; Lei, J.; Yang, R. Influence of thermal treatment on element partitioning in α+β titanium alloy. *J. Alloys Compd.* **2019**, *791*, 575–585.

(23) Li, Y.; Yang, W.; Dong, R.; Hu, J. Mlatticeabc: Generic Lattice Constant Prediction of Crystal Materials Using Machine Learning. *ACS Omega* **2021**, *6*, 11585–11594.

(24) Peng, C.; Tran, P.; Nguyen-Xuan, H.; Ferreira, A. J. M. Mechanical performance and fatigue life prediction of lattice structures: Parametric computational approach. *Compos. Struct.* **2020**, *235*, 111821.

(25) Lee, S.; Zhang, Z.; Gu, G. X. Generative machine learning algorithm for lattice structures with superior mechanical properties. *Mater. Horiz.* **2022**, *9*, 952–960.

(26) Wang, T.; Zhang, C.; Snoussi, H.; Zhang, G. Machine Learning Approaches for Thermoelectric Materials Research. *Adv. Funct. Mater.* **2020**, *30*, 1906041.

(27) Wei, H.; Bao, H.; Ruan, X. Perspective: Predicting and optimizing thermal transport properties with machine learning methods. *Energy and AI* **2022**, *8*, 100153.

(28) Tran, M.-K.; Panchal, S.; Chauhan, V.; Brahmbhatt, N.; Mevawalla, A.; Fraser, R.; Fowler, M. Python-based scikit-learn machine learning models for thermal and electrical performance prediction of high-capacity lithium-ion battery. *Int. J. Energy Res.* **2022**, *46*, 786–794.

(29) Ren, Y.; Zhang, K.; Zhou, Y.; Cao, Y. Phase-Field Simulation and Machine Learning Study of the Effects of Elastic and Plastic Properties of Electrodes and Solid Polymer Electrolytes on the Suppression of Li Dendrite Growth. *ACS Appl. Mater. Interfaces* **2022**, *14*, 30658–30671.

(30) Lv, C.; Zhou, X.; Zhong, L.; Yan, C.; Srinivasan, M.; Seh, Z. W.; Liu, C.; Pan, H.; Li, S.; Wen, Y.; Yan, Q. Machine Learning: An Advanced Platform for Materials Development and State Prediction in Lithium-Ion Batteries. *Adv. Mater.* **2022**, *34*, 2101474.

(31) Zhao, X.; Luo, T.; Jin, H. Predicting Diffusion Coefficients of Binary and Ternary Supercritical Water Mixtures via Machine and Transfer Learning with Deep Neural Network. *Ind. Eng. Chem. Res.* **2022**, *61*, 8542–8550.

(32) Allers, J. P.; Priest, C. W.; Greathouse, J. A.; Alam, T. M. Using Computationally-Determined Properties for Machine Learning Prediction of Self-Diffusion Coefficients in Pure Liquids. *J. Phys. Chem. B* **2021**, *125*, 12990–13002.

(33) Leverant, C. J.; Harvey, J. A.; Alam, T. M.; Greathouse, J. A. Machine Learning Self-Diffusion Prediction for Lennard-Jones Fluids in Pores. *J. Phys. Chem. C* **2021**, *125*, 25898–25906.

(34) Guo, S.; Huang, X.; Situ, Y.; Huang, Q.; Guan, K.; Huang, J.; Wang, W.; Bai, X.; Liu, Z.; Wu, Y.; Qiao, Z. Interpretable Machine-Learning and Big Data Mining to Predict Gas Diffusivity in Metal-Organic Frameworks. *Adv. Sci.* **2023**, *10*, 2301461.

(35) Jirasek, F.; Hasse, H. Combining Machine Learning with Physical Knowledge in Thermodynamic Modeling of Fluid Mixtures. *Annual Review of Chemical and Biomolecular Engineering* **2023**, 31–51.

(36) Olumegbon, I. A.; Alade, I. O.; Oyedeji, M. O.; Qahtan, T. F.; Bagudu, A. Development of machine learning models for the prediction of binary diffusion coefficients of gases. *Eng. Appl. Artif. Intell.* **2023**, *123*, 106279.

(37) Li, G.; Zhu, Y.; Guo, Y.; Mabuchi, T.; Li, D.; Huang, S.; Wang, S.; Sun, H.; Tokumasu, T. Deep Learning to Reveal the Distribution and Diffusion of Water Molecules in Fuel Cell Catalyst Layers. *ACS Appl. Mater. Interfaces* **2023**, *15*, 5099–5108.

(38) Matsuo, M.; Yamazaki, M. Diffusion database kakusan. 2023; https://samurai.nims.go.jp/misc_reports/21abef80-9b37-4d6e-8220-e025841f83e5?locale=en, [Online; accessed 20. Jul. 2023].

(39) Yamazaki, M.; Xu, Y.; Murata, M.; Tanaka, H.; Kamihira, K.; Kimura, K. NIMS structural materials databases and cross search engine œ MatNavi. *BALTICA VII* **2007**, 193.

(40) Ogata, T.; Yamazaki, M. New stage of MatNavi, materials database at NIMS. **2012**,

(41) Ward, L. et al. Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **2018**, *152*, 60–69.

(42) Ward, L.; Agrawal, A.; Choudhary, A.; Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *npj Comput. Mater.* **2016**, *2*, 1–7.

(43) Benesty, J.; Chen, J.; Huang, Y.; Cohen, I. *Noise Reduction in Speech Processing*; Springer: Berlin, Germany, 2009; pp 1–4.

(44) pandas development team, T. pandas-dev/pandas: Pandas. 2020; https://doi.org/10.5281/zenodo.3509134.

(45) Kaiser, H. F. The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement* **1960**, *20*, 141–151.

(46) Fabrigar, L. R.; Wegener, D. T.; MacCallum, R. C.; Strahan, E. J. Evaluating the use of exploratory factor analysis in psychological research. *Psychological methods* **1999**, *4*, 272.

(47) Pedregosa, F. et al. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* **2011**, *12*, 2825–2830.

(48) CitrineInformatics lolo. 2024; https://github.com/CitrineInformatics/lolo, [Online; accessed 23. Apr. 2024].

(49) Ling, J.; Hutchinson, M.; Antono, E.; Paradiso, S.; Meredig, B. High-Dimensional Materials and Process Optimization Using Data-Driven Experimental Design with Well-Calibrated Uncertainty Estimates. *Integr. Mater. Manuf. Innov.* **2017**, *6*, 207–217.

(50) Jolliffe, I. T.; Cadima, J. Principal component analysis: a review and recent developments. *Philos. Trans. Royal Soc. A* **2016**, *374*.