

ARTICLE OPEN



Automation of diffusion database development in multicomponent alloys from large number of experimental composition profiles

Jing Zhong¹, Li Chen¹ and Lijun Zhang^{1,2}✉

Nowadays, the urgency for the high-quality interdiffusion coefficients and atomic mobilities with quantified uncertainties in multicomponent/multi-principal element alloys, which are indispensable for comprehensive understanding of the diffusion-controlled processes during their preparation and service periods, is merging as a momentous trending in materials community. However, the traditional exploration approach for database development relies heavily on expertise and labor-intensive computation, and is thus intractable for complex systems. In this paper, we augmented the HitDIC (high-throughput determination of interdiffusion coefficients, <https://hitdic.com>) software into a computation framework for automatic and efficient extraction of interdiffusion coefficients and development of atomic mobility database directly from large number of experimental composition profiles. Such an efficient framework proceeds in a workflow of automation concerning techniques of data-cleaning, feature engineering, regularization, uncertainty quantification and parallelism, for sake of agilely establishing high-quality kinetic database for target alloy. Demonstration of the developed infrastructures was finally conducted in fcc CoCrFeMnNi high-entropy alloys with a dataset of 170 diffusion couples and 34,000 composition points for verifying their reliability and efficiency. Thorough investigation over the obtained kinetic descriptions indicated that the sluggish diffusion is merely unilateral interpretation over specific composition and temperature ranges affiliated to limited dataset. It is inferred that data-mining over large number of experimental data with the combinatorial infrastructures are superior to reveal extremely complex composition- and temperature-dependent thermal-physical properties.

npj Computational Materials (2021)7:35; <https://doi.org/10.1038/s41524-021-00500-0>

INTRODUCTION

Interdiffusion involves in a variety of materials processes in metallic solids, for instance, solidification¹, solid solution², aging³, corrosion⁴, mutual interaction between coatings and matrix⁵, and so on. Diffusion represents the random motion of particles, i.e., atoms, molecules or ions, and is affected by the change of chemical potential of the systems, namely the interdiffusion or chemical diffusion. The movement of an individual particle is stochastic; however, it is governed by the thermodynamic and kinetic state of the system. In a complex system, the interactions between different types of particles are not identical. Such non-ideal situation becomes serious as the number of components increases in systems with multiple components. For example, HEAs (high-entropy alloys), where the multiple components are presented as principal constituents, nowadays serve as the alternatives for many traditional alloy systems, where only one or two components are presented as principal constituents. An interesting hotspot arises from the intriguing interactions, where diffusion rates seem to be rather low among the composition space around the equal atomic composition space. Such distinct diffusion behavior is reported to show the significant effect on the good mechanical properties and service performance of HEAs³. Regarding the complex essence of diffusion, extensive efforts have been contributed to the investigation of the diffusion behaviors of HEAs for the very recent years^{6–13}. Debates about such sluggish diffusion effect still continue due to the incomplete overview of the desired systems^{14–17}. However, a comprehensive insight into

diffusion behavior cannot be overwhelmingly supported without quantitative description of diffusion coefficients.

Determination of interdiffusion coefficients and later development of kinetic database have long been impeded by the incomplete techniques and toolsets. Matano-based methods are the most popular solutions for the determination of interdiffusion coefficient over the last several decades¹⁸. Such historical methods and related tools are oriented for simple systems, i.e., binary^{19–21}, ternary^{22,23}, and a portion of higher-order systems^{24–26}, and thus, are inadequate to meet up with the interest of industrial research, where the systems are complex with multi-components, e.g., solders, nickel-based alloys, HEAs and so on. Moreover, the size of the dataset is growing ever larger benefiting from advanced techniques for the preparation of diffusion couples^{27,28} and measurement of composition profiles^{29,30}. Assessment of diffusion description used to be based on 1–20 diffusion couples are shifting to hundreds, i.e., for Ni-based alloys and HEAs. The amount of work towards data-mining over such large dataset is challenging for the existing labor-intensive procedures and workflow.

To date, the most promising approach for unveiling the complex kinetic interactions among multicomponent alloy systems has recently been described as the numerical inverse method^{18,31–36}. The essence of the numerical inverse method rests with revealing the inverse problem, that is reasoning the interdiffusion coefficients from the experimental composition profiles. The numerical inverse method dominates the Matano-based ones for proceeding reasoning without considering the

¹State Key Laboratory of Powder Metallurgy, Central South University, Changsha, Hunan, China. ²Shenzhen Research Institute, Central South University, Shenzhen, China. ✉email: lijun.zhang@csu.edu.cn

number of components, but requiring the initial/final states and evolution rules¹⁸. Currently, several efforts have been contributed to extract diffusion information from the specific diffusion dataset by some researchers independently^{31–37}. Unfortunately, the well-defined toolsets for numerical inverse methods are not yet enough to cope with dataset of large size for the complex systems. The primary issue is due to the naive implementation of numerical inverse method without revealing the deep essence of the inverse problem, where the tendency of being ill-posed is pressing. Situation gets worse as the size of dataset and the dimension of parameter space grows, turning into the large-scale inverse problem^{38–40}. Curse of dimensionality haunts especially for multicomponent systems, because the number of descriptors for the related diffusion behaviors is large. Complying to CALPHAD (CALculation of PHase Diagram) approach, the interaction parameters to be concerned for a quinary system can be up to 200 and even more. Assessments over such complex systems are difficult because the parameter space to be explored become myriad, while time expense for exploration become numerous. Such large-scale inverse problem becomes intractable as it is much worse conditioned and often not uniquely solvable. When the parameter space and size of dataset reach a large scale, i.e., about 200 and more, pace would be extremely hard to proceed with manual construction based on expertise⁴¹, which is neither agile for integral database development nor meeting up with features of high-throughput and automation^{42–45}.

Consequently, to bridge between the challenges and anxious anticipation, the infrastructures in automation fashion are thus in urgent need for the establishment of diffusion database, serving as the primary motivation of the present work. High-performance computing (HPC) comes into the prior position to help with large dataset and speedup the related algorithms. Dimensionality reduction technique is required to simplify the overall complexity of the concerned diffusion descriptions. Meanwhile, both the uniqueness and generality should be emphasized towards the calculation results. Uncertainty quantification should also be served as an important portion, which indicates the useful information about the reliability of the assessed results.

Integration of the proposed techniques are further in need of enabling a workflow of automation in practical applications and for accomplishing the thoughtful concerns above. Subsequently, we are going to report a successful demonstration of an automation computation framework for interdiffusivity evaluation and atomic mobility database development. It especially paves the way for settling the large-scale inverse diffusion property problems with multicomponent and/or multi-principal element alloys. Demonstration of the advanced infrastructure proposed in the present work is thus performed by conquering fcc CoCrFeMnNi HEAs from the point of view of its related diffusion behaviors.

RESULTS AND DISCUSSION

Framework and infrastructure

To be clear, the methods, strategies and codes are developed and bundled as HitDIC infrastructures, in the interest of realizing the interdiffusivity computation and atomic mobility dataset development in multicomponent/multi-principal element alloys in a manner of automation. Originally, HitDIC is designed to extract interdiffusion coefficients from composition profiles and it has been successfully applied to multiple alloy systems^{10,12,46,47}. Later, HitDIC is featured with the capability of uncertainty quantification cooperating with Bayesian inference. However, the toolkits are not yet capable of dealing with dataset of large size and parameter space of high dimension. The well-established data-mining techniques are, therefore, employed to levitate HitDIC to large-scale inverse diffusion property problem. Infrastructure is built so as to drive the data/information flow in a manner of automation, as illustrated in Fig. 1a.

To begin with, dataset of composition profiles is extensively collected and preprocessed to produce denoised sample dataset. Hand-out validation can be proceeded by splitting sample sets into the training dataset and validation dataset with a ratio of 80–20. Secondly, the training dataset is further utilized in an optimization process, which offers the functionality for joint parameter selection and estimation. The optimization processes can be executed several

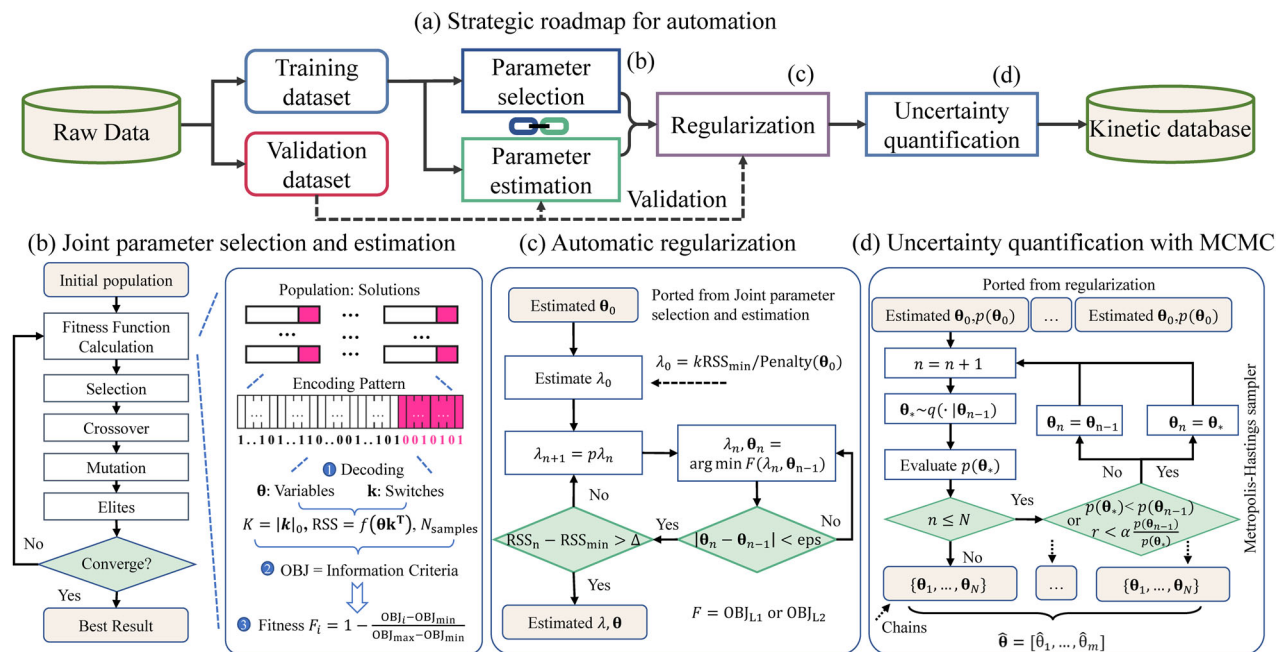


Fig. 1 Development roadmap of the HitDIC software. **a** Strategic workflow for automation; **b** Schematic illustration of variable-selection genetic algorithm and its encoding pattern of the variables and switches; **c** Illustration of the regularization algorithm for automatically tuning the regularization term, typical settings of which are $k = 1.5$ and $p = 10^{-3}$; **d** Illustration of the Metropolis–Hastings algorithm with multiple chains.

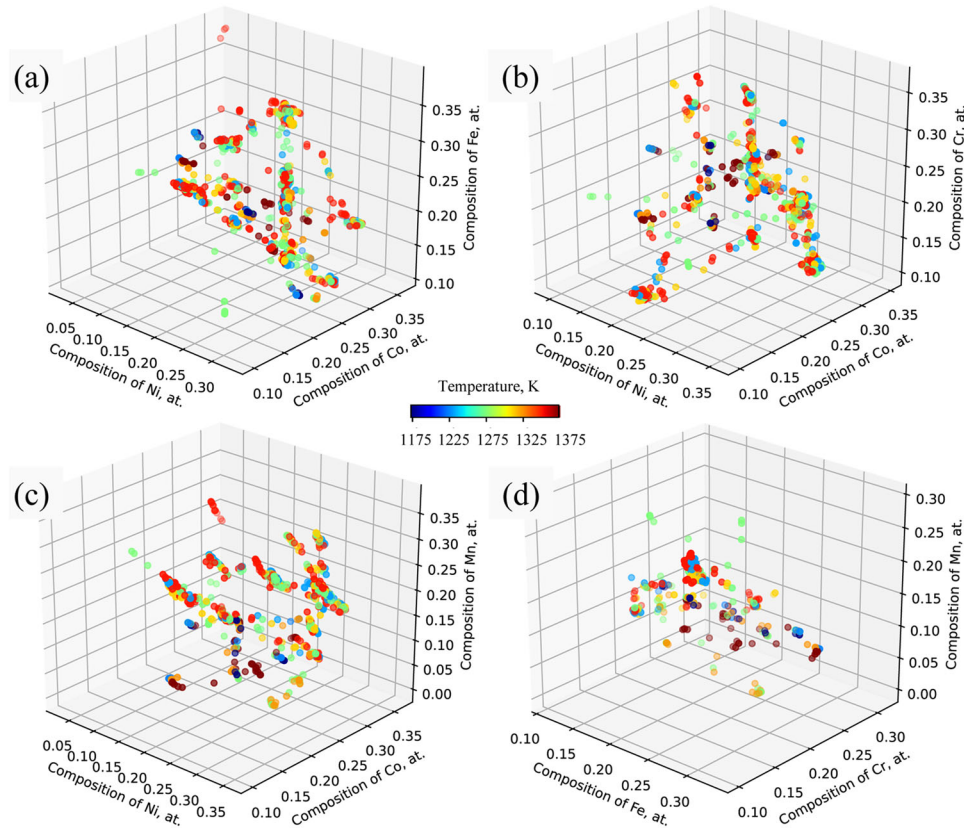


Fig. 2 Composition and temperature space of the dataset. Distributions of the composition points are viewed from composition coordinates of different components, i.e., **a** Ni, Co and Fe, **b** Ni, Co and Cr, **c** Ni, Co and Mn, and **d** Fe, Cr and Mn, among concerned temperature range for the dataset of fcc CoCrFeMnNi HEAs.

times, and it will be repeated until a staged model with subset of effective parameters of interest is determined. Thirdly, the staged model will then be further proceeded to a regularization process, where the regularization term can be automatically tuned. Generality and stability of the estimations are intended to be improved in this period. Finally, in case that uncertainty of parameters of interest is concerned, the Bayesian inference will be employed to estimate the posterior distributions of the concerned parameters. Once the estimations of the concerned parameters and their uncertainties are determined, the diffusion database for the concerned system is thus developed. All the optimizers, i.e., the optimizer for joint parameter selection and estimation, regularization optimizer and the samplers for Markov chain Monte Carlo (MCMC), are supported by the cost evaluator with parallelism (see the “Methods”). Workflow of the concerned modules, i.e., joint parameter selection and parameter estimation, regularization and uncertainty quantification, are integrated. As the high-performance computing is in need, such kernels are implemented with C++ for the consideration of efficiency. The pre-processing and post-processing are implemented with Python due to the plentiful toolkits for visualization and data manipulation, while the reports are generally presented in web view pages.

Data collection and preparation

To exemplify the capability of HitDIC infrastructure, HEAs are chosen as an example to demonstrate the capability of the developed infrastructure. The fcc CoCrFeMnNi alloys or Cantor alloys⁴⁸ are employed in the present work. Being attracted by the sluggish diffusion effect, extensive investigations over the Cantor alloy, as well as other HEA systems, have been conducted by various groups^{6–10}. However, the diffusion database for the

CoCrFeMnNi system is still considered incomplete. Therefore, Tsai et al.⁶ adopted the diffusion couple approach and a simplified Matano-based calculation method for studying diffusion coefficients on this system. Subsequent studies were carried out by Vaidya et al.^{9,49}, Kulkarni et al.⁵⁰, Verma et al.⁵¹, Chen and Zhang¹¹, Wang et al.⁵², Dąbrowa et al.⁸, and Kuczka et al.⁷ However, it is considered that those studies based on the limited dataset, i.e., <20 diffusion couples, are prone to be inconsistent with respect to the temperature or composition ranges. To produce trustworthy results, all reported data for the fcc CoCrFeMnNi system are therefore gathered, constituting a dataset of composition profiles with up to 170 groups of composition profiles over the temperature range of 1073–1373 K.

The collected composition profiles are further smoothed using the preset fitting functions, i.e., the logistic function or its superposition⁵³, the distribution functions and their superpositions⁵⁴ and so on. Noises are thus removed, and the smoothed composition profiles are produced. By means of the fitting process, as detailed in Supplementary Methods, the dataset is renewed by applying the prior assumption that noise should be suspended. The composition profiles are thus resampled with the fitted functions, while the overall dataset possess up to 34,000 composition points, as illustrated in Fig. 2. It is indicated the concerned system is stable among composition range from 0.05 at. to 0.35 at. and temperature range from 1175 to 1373 K. It is important to note that the composition points remain sparse over the composition space in Fig. 2. Considering the large composition ranges, establishment of diffusion database for the concerned system is more desirable for completing the view of diffusion behaviors rather than outspreading the measured composition space with many more expensive diffusion couple/multiple experiments.

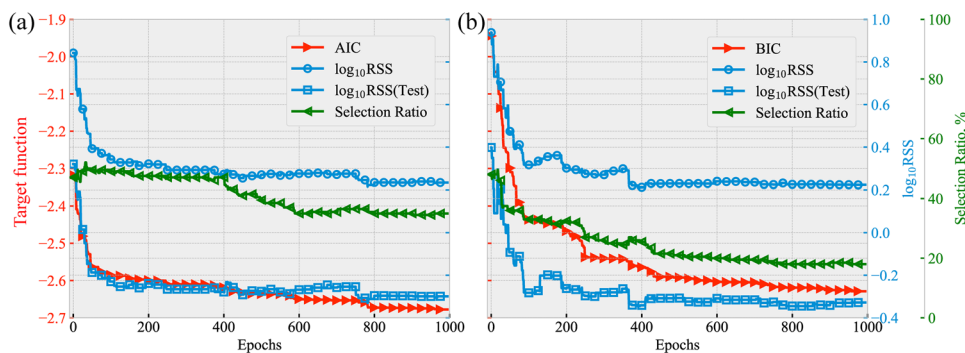


Fig. 3 Convergence sequences of the joint parameter selection and estimation processes. Results for joint parameter selection and estimation using variable-selection genetic algorithm based on **a** AIC and **b** BIC.

In this work, up to 34,000 composition points are employed, and the evaluated diffusion database is, therefore, considered generalizable among the mentioned composition ranges and temperatures ranges. What's more, before taking the successive procedures, the overall dataset is then split into the training dataset and the validation dataset, with a ratio of 80 to 20.

Joint parameter selection and estimation

Joint parameter selection and parameter estimation are proceeded with the variable-selection genetic algorithm. The default sets of atomic mobility parameters are assigned with the ones up to the first order for the binary systems and all those of ternary systems. For CoCrFeMnNi HEA, dimensionality of the concerned parameter space is large, e.g., roughly up to 200, which is about ten times more than that of binary or ternary systems. Redundant descriptors are likely to be introduced and desired to be sweep off due to convergency difficulty for optimizer towards problems of very high dimension. The variable-selection genetic algorithm is subsequently applied to identify the most appropriate subset of parameters and their related estimations.

Convergence sequence of the proposed algorithm is illustrated in Fig. 3, where both Akaike information criterion (AIC) and Bayesian information criterion (BIC) are tested. In Fig. 3a, the AIC, residual summation square (RSS) and selection ratios are superimposed. The sequence of selection ratio converges nicely, reaching about 35% out of 200 parameters. The AIC drops rapidly at the early stage, however, results in a long tail as the algorithm proceeds. Although AIC sequence evolves slowly at the latter stage, the competition between parameters remains intensely, as indicated in Supplementary Fig. 2. With respect to the training dataset, the RSS and selection ratios evolve consistently with AIC value, indicating the effectiveness of the proposed joint parameter selection and estimation process. Furthermore, the sequence of RSS for the training and testing sets behaves similarly, which implies the validity of the variable-selection genetic algorithm.

When it comes to the result based on BIC, more stringent selection efficiency is achieved, i.e., 18%, as shown in Fig. 3b. The potential reason for such distinct difference in selection ratios for the two criteria lies in that BIC imposes larger penalty on the number of concerned parameters. Comparing the optimized RSS for both criteria, the BIC succeeds to achieve by $RSS_{\text{BIC}} = 1.67$, which is slightly better than that of AIC, i.e., $RSS_{\text{AIC}} = 1.71$. That is when the BIC is employed, the result of variable-selection genetic algorithm is able to achieve a better fitting goodness with less descriptors. It is indicated that the BIC is more suitable when the dimension is similar to the size of dataset. Currently, the result based on BIC is accepted as the product of the joint parameter selection and estimation and used in subsequent investigation. For details about the selected effective parameters and the obsoleted ones, the readers can refer to Supplementary Table 2. It is worthy of mentioning that the AIC might be not effective for the

current size of the dataset; however, it might be effective for even larger dataset. Moreover, no matter what criterion is applied, the difficulty in proceeding the optimization remain intractable without well-designed cost evaluator accommodated with high-performance parallel computing resources.

Regularization

In the framework of inverse problem, the estimated model has a fixed but unknown probabilistic relation to the data space. In previous researches on numerical inverse method, the solution to the inverse problem is found to be sensitive to the size of dataset, when the overall size of the dataset is small^{53,55}. Such phenomenon accords well with the primary feature of inverse problem, i.e., being ill-posed, resulting in severe problem about uniqueness. Plainly, the optimization using merely the first term in Eqs. (11) or (12) is insufficient to guarantee uniqueness of the solutions to inverse problem. The feature of being ill-posed can be weakened by increasing the sample sizes, i.e., considering as many as observations in the inverse process. As shown in Fig. 3, the solutions to the inverse problem can be reduced to limited alternatives, as they are constrained by means of expertize and statistic criteria as mentioned. Unfortunately, the potential solutions to inverse problem are still massive, inferred from the convergence sequences with long tail as presented in Fig. 3.

Therefore, a technique to address the problem of non-uniqueness is taken into consideration. Regularization is one of the common techniques served for releasing the ill-posedness of the inverse problem. The key of this technique is to introduce the concept of conditional well-posedness and shifts from searching for stable methods to reaching approximate solution with prior assumption. In other words, the regularization is to apply prior assumption on the solutions to the inverse problem, and therefore the target solution can be reduced to a limited model and the parameter space of less freedom.

The most frequently used prior assumption for regularization is that the L_1 norm or L_2 norm of the solution to inverse problem is considerably small enough. In practice, regularization is fulfilled by solving an optimization problem penalized by the L_1 norm or L_2 norm of the concerned parameters, where a regularization term, i.e., λ , is introduced to rescale the penalty. Presently, a workflow is used to tune the regularization term online while improving the solutions to the inverse problem (see the "Methods"). With the proposed algorithm, the selected estimators are further investigated, and the convergence sequences are presented in Fig. 4. From the point of view in regularization, the value of the prior assumptions, i.e., L_1 norm or L_2 norm, represents the complexity of the model or parameters. The key to regularization algorithm is to figure out an estimation of parameters with least complexity, while ensuring that the fitting goodness does not significantly turn worse.

For the L_2 norm regularization, a significant increase of RSS value is observed when the regularization term grows larger than

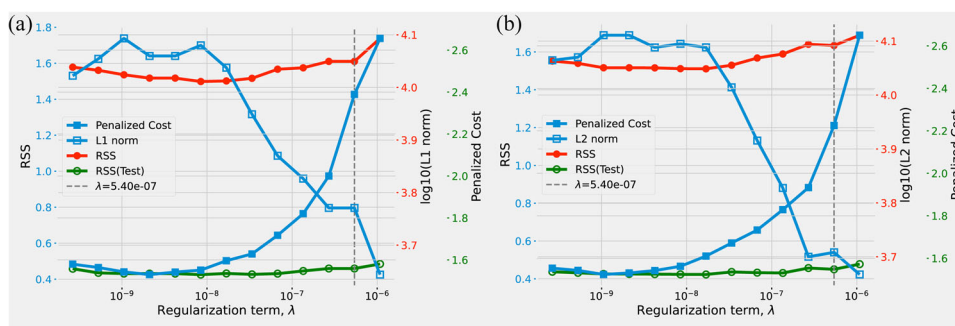


Fig. 4 Convergence sequences of the regularization strategy. Automatically tuning of regularization term for regularization processes using (a) L1 norm and (b) L2 norm.

5.40×10^{-7} , where the model complexity also decreases significantly comparing to the previous iterations, as indicated in Fig. 4b. Convergence sequence of the L₁ norm regularization behave similarly as the L₂ norm, and the turning point of the regularization term is the same, i.e., around 5.40×10^{-7} . Sequence of the RSS value for the validation dataset behaves consistently with that of training dataset, indicating the generality of estimations are good. Moreover, it also implies that the data distributions of the training set and validation set are similar, which generally exists when both datasets are sufficiently large. Such conclusion applies for both the L₁ norm and L₂ norm regularizations, as shown in Fig. 4a, b. As the regularizations are modeled based on the prior assumptions, both regularized results are considered reasonable and applicable for generalization and interpolation, as listed in the Supplementary Table 3.

Uncertainty quantification

Beyond the deterministic optimization procedures above, Bayesian inference is a suitable alternative for measuring the uncertainty of the solution to inverse problems. To further quantify the uncertainty of the concerned parameters, MCMC method, or more specifically Metropolis–Hastings algorithm, is currently used to infer their posterior means and variances (as demonstrated in methods). Considering acceptable computation cost, the multiple short but independent chains are employed currently. Twelve chains are proceeded with 50,000 iterations individually. Convergence diagnosis (see Supplementary Methods) is carried on the accepted sample points produced by the Metropolis–Hastings (M-H) sampler. As shown in Fig. 5a, within sequence variances change significantly at the early period, while it achieves quite stable level as the sequences evolve. A better illustration of the stability of the Markov chains is indicated in Fig. 5b, where the between sequence variances are flat for different parameters. Unfortunately, the potential scale reduction factors for different parameters hardly reach 1, as illustrated in Fig. 5c, indicating that the individual chains are not yet reaching the stationary state. However, considering that the between sequence converge satisfactorily as shown in Fig. 5b, the current assessment is deemed as reasonable.

The proposed M-H sampler totally draws 240,000 sample points with an accepted ratio of 40%. Posterior distributions of parameters are subsequently described as shown in Fig. 5d, where the mean and the related bounds of concerned parameters are imposed. It should be noted that the bounds are determined by a quantile of [0.2, 0.8], which covers about 60% sample points. Though only histograms of the posterior distributions of individual parameters are presented, they are actually subject to an integral joint posterior distribution. As shown in Fig. 5d, the posterior distributions of most parameters present in bell shapes, while the mean estimations reasonably rest around the high-density region.

Estimated parameters with quantified uncertainties are important precursors to the quantification of uncertainties underpinning prediction and decision-making. For instance, the bounds of the predicted composition profiles can be retrieved as the uncertainty propagates through the forward problem of diffusion. As an example, Fig. 6a–d illustrates the fitted composition profiles of four selected diffusion couples of fcc CoCrFeMnNi HEAs, as denoted with dash lines and markers. The model-predicted composition profiles are also imposed in Fig. 6, where the bounds of composition profiles rather than the exact optimal ones are presented. It has to be noted that the bounds are determined according to the parameters with uncertainties, which firstly propagate into the interdiffusion, and secondly to the composition profiles via the forward simulation. The overall goodness of the prediction to the experimental or fitted composition profiles are satisfactory, though parts of fitted composition profiles rest out of the bands of the predicted bounds, i.e., Fig. 6a. As the experimental procedures are taken as random events, the deviations outside the bounds infer the inconsistency with the assessed model. From the abroad comparison between the fitted and the model-predicted composition profiles, the generality ability of the selected parameters and evaluated uncertainties are reasonable. For complete view of fitting goodness towards the whole dataset, the readers are referred to Supplementary Figs 10–31.

Remarks on optimization techniques

In the present work, the joint parameter selection and estimation have been succeeded to significantly reduce the dimensionality of the proposed problem. What's more, the regularization is used to reduce the overall complexity of the selected model, while reserving the fitting goodness of the diffusion description. The two deterministic optimization strategies are feasible to come up with estimations, i.e., θ_{MLE} , θ_{ReL1} and θ_{ReL2} , with limited iterations. Besides, both techniques are constructed based on the genetic algorithm in the present work, which is famous for its robustness and promising ability for global optimal⁵⁶. The MCMC method is a statistic inference method, which is also capable of offering estimations, i.e., θ_{MAP} , that produce the promising fitting goodness to the observations. However, it is generally more time-consuming due to the considerable number of iterations. In most scenarios, the joint parameter selection and estimation followed with regularization are qualified, while the MCMC is superior to understand uncertainties of related model and parameters. Overall, the model-predicted composition profiles to observations are similar for most samples or diffusion couples, as illustrated in Fig. 6e–h. For intuitive comparison between the observations and predictions, the readers are referred to the composition profiles presented in Supplementary Figs 32–53 for more detail.

However, it should be clarified that different optimization strategies actually perform differently according to their optimization criteria. Evidence lies in the statistics over the deviations

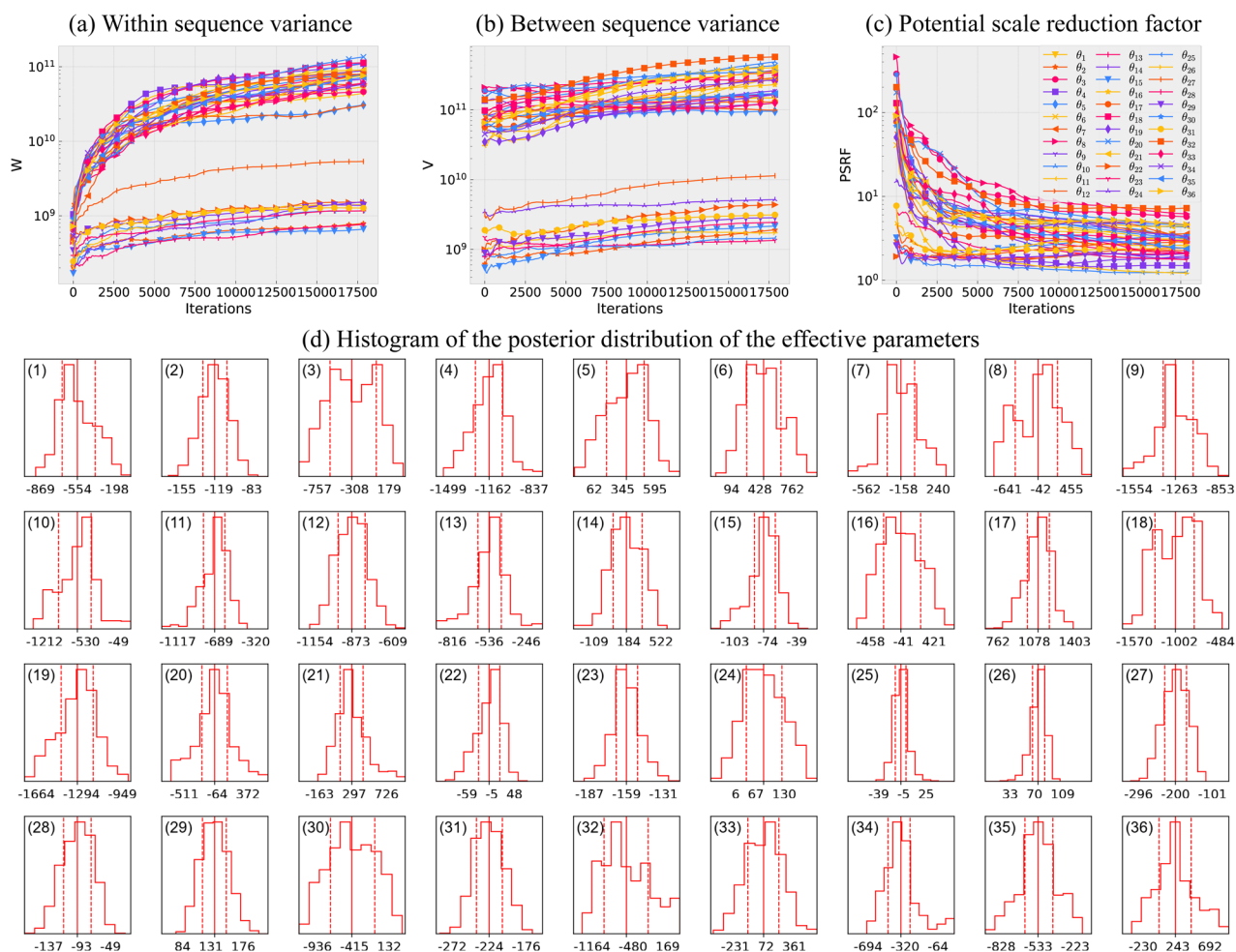


Fig. 5 Statistics over the sequences of samples. Samples are drawn by means of the Metropolis–Hastings algorithm with 12 independent chains. Statistics are **a** Within sequence variance, **b** Between sequence variance and **c** Potential scale reduction factor of the MCMC sequences. **d** Histogram of the posterior distribution of the effective parameters, while all estimations have been divided by 10^3 and specific notions of the parameters are indexed in Supplementary Table 3.

between the observations and the predictions, i.e., as illustrated in Fig. 7. They are the distributions of the prediction biases due to variable-selection genetic algorithm, L_1/L_2 norm regularization and MCMC. The histograms of distributions are very similar to each other, though the estimations of the algorithms are quite different, as listed in Supplementary Table 3. Variable-selection genetic algorithm with AIC and BIC, L_1/L_2 norm regularization achieve similar mean square errors as denoted as star in Fig. 7. Among the three strategies, the variable-selection genetic algorithm performs slightly better, as the prediction biases tend to concentrate more obvious around zero than the others. When it comes to the estimations produced by MCMC, the related fitting goodness is much better than the above three strategies as the related RSS is smaller. In trade of better overall fitting goodness, the prediction biases to the observations concentrate less significantly around 0 regarding the maximum a-posterior.

Reason for difference in performance of different algorithms also lies in optimization criteria of the algorithms. Joint parameter selection and estimation is subjected to the information criteria, where the fitting goodness or RSS is partly concerned in the cost function. When most appropriate subset of parameters are selected, the estimated parameters are closed to the maximum likelihood estimation θ_{MLE} . Regularization is expected to reduce the complexity of the concerned model by means of imposing additional penalty in the cost function. With the proposed regularization algorithm, results of regularization, i.e., θ_{ReL1} and

θ_{ReL2} , that do not significantly make worse prediction are taken. Therefore, the prediction performance of θ_{ReL1} and θ_{ReL2} is similar to that of θ_{MLE} , however, the L_1/L_2 norm shrinks distinctly. The goal of MCMC is to profile the joint posterior distribution of the concerned model and parameters. When the posterior distribution is reasonably drawn, the mean estimation is bound to the maximum a-posterior θ_{MAP} . The maximum a-posterior tends to achieve better fitting goodness towards the entire dataset, though it behaves slightly worse towards the specific samples. The difference between θ_{MLE} and θ_{MAP} is partly due to the great convergence difficulty of the variable-selection genetic algorithm in high dimension. Therefore, the maximum a-posterior or the regularized estimations are more convincingly accountable estimations.

Nevertheless, MCMC starting from random initial proposals might take much longer iterations to reach the stationary state, when the initial estimations are far from the high-density region. Situation gets worse when the dimension is extremely high and the dataset is especially large. Concerning the same problem, the expense of exploring the high-density region for the deterministic algorithms is significant lower, which is superior to providing promising initial proposals for the MCMC algorithm. However, despite of all the pros and cons above, results of MCMC are indispensable as the information about the uncertainties can be provided. Analysis over the potential influence on the model-predicted properties is therefore feasible. In more abroad

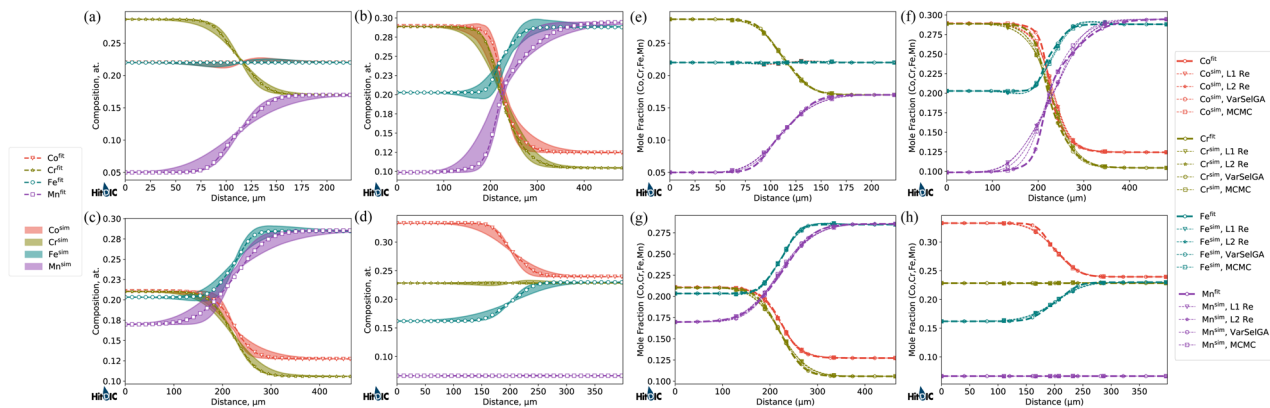


Fig. 6 Examples of fitted and model-predicted composition profiles. Composition profiles for the fcc CoCrFeMnNi high-entropy alloys: **a, b, c, and d** concern bounds of the composition profiles due to uncertainty of the diffusion database; **e, f, g, and h** concern model-predicted composition profiles due to optimization results of different algorithms, i.e., variable-selection genetic algorithm (VarSelGA), L1 norm regularization (L1 Re), L2 norm regularization (L2 Re), and Markov chain Monte Carlo (MCMC). The composition profiles of the fitted data (noted as fit) are ported from Tsai et al.⁶ for (a) and (e), Kucza et al.⁷ for (b) and (f), Dąbrowa et al.⁸ for (c) and (g), and Chen and Zhang¹¹ for (d) and (h).

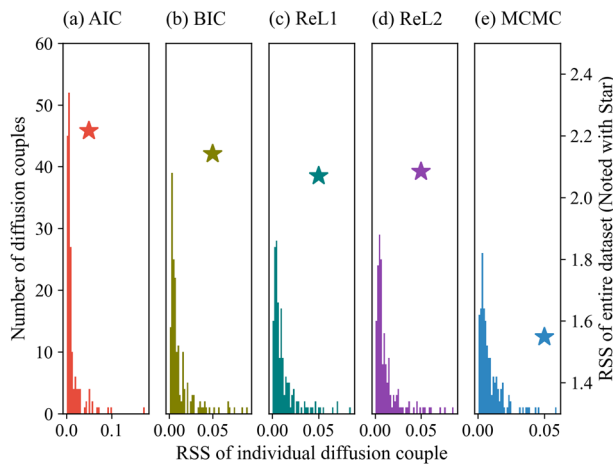


Fig. 7 Goodness-of-fit statistics for the proposed algorithms. The histograms are counts of diffusion couples with different intervals of mean square error concerning optimization algorithms, i.e., **a** variable-selection genetic algorithm based on AIC, **b** variable-selection genetic algorithm based on BIC, **c** L1 norm regularization, **d** L2 norm regularization, and **e** Markov chain Monte Carlo, while the marker “star” denotes RSS due to the entire dataset.

applications, the uncertainty is favorable to territories of material design by offering decision-making insights^{57,58}.

Remarks on efficiency of parallelism

Among all the optimizers, the cost evaluator is fundamental to carry out predictions and evaluate the residual summation of square between the predictions and observations. For dataset of large size, evaluation of RSS is expensive, especially for optimizer demanded on great number of iterations, i.e., MCMC samplers. Parallelism is therefore mandatory while efficiency of parallelism is considerable. The scalability of a parallelism scheme is a measure of its ability to effectively exploit an increasing number of cores or processors on HPC clusters. Scalability analysis is usually designed for the most satisfactory algorithm-architecture combination for a problem under different constraints on the growth of the problem size and the number of processors from the performance on fewer processors.

Efficiency test of the devised cost evaluator with parallelism is shown in Fig. 8b, where the case with 2 nodes (48 cores/threads on Intel Xeon CPU E5-2692) is set as the base. Comparison

between the Intel Xeon CPU E5-2692 and AMD EPYC 7452 are considered, where the identical compilers, as well as the corresponding compiler options, are adopted. The cost evaluator scales nicely when a large dataset is considered for both computing resources, where 576 or more cores can be utilized. When the number of nodes is small, i.e., 4 nodes for E5-2692 and 1 node for EPYC 7452, the efficiency of the two types of computing resources are similar. However, the speedup of EPYC 7452 hits about 11 times with 9 nodes (576 processors) comparing to 8 times with 24 nodes (576 processors) with respect to E5-2692. Such a result benefits from a fact that the machine with higher efficiency, i.e., high CPU clock cycle, works faster. Moreover, less message passing interface (MPI) communications between nodes would also benefit the efficiency for machines with more cores/threads on a single node. Currently, the optimization processes are carried on HPC clusters of AMD EPYC 7452.

Remarks on sluggish diffusion effect

For the HEAs, one of the most attracting topics is the existence of sluggish diffusion effect, which remains as a mystery in the past decade^{6,7,9–12,15,46,49}. Yeh⁵⁹ originally proposed that kinetics of diffusion is hindered in comparison to pure metals and conventional alloys, resulting in smaller values of diffusion coefficients. It is inferred that the potential deduction of the diffusion rate of the high-entropy alloy is due to the increase of entropy. To calibrate the influence of entropy on the diffusion rates, correlation between configurational entropy and different kinds of diffusion rates are examined. For sake of clarity, the thermodynamic description for the fcc CoCrFeMnNi system is considered identical to that of ideal solution phase, where only the configurational entropy rather than the excess interactions contributes to the thermodynamic factors.

To begin with, the effective tracer diffusion coefficients of pure metals or alloys of equal atomic compositions are compared. For pure metals or alloys of equal atomic compositions, larger configurational entropy is relatedly bound for higher-order system. Lacking in physical thermodynamic factors, the tracer diffusion coefficients are deemed as the effective ones, while the evaluated values using the assessed estimations θ_{MAP} are presented in Fig. 9a. The averages of the effective tracer diffusion coefficients of the quinary system are evaluated at various temperatures and taken as the base line for comparison, denoted as the dash lines in Fig. 9a. Considering the quaternary systems, the tracer diffusion coefficients fluctuate around the base line, which indicates that the reduction of entropy does not imply a

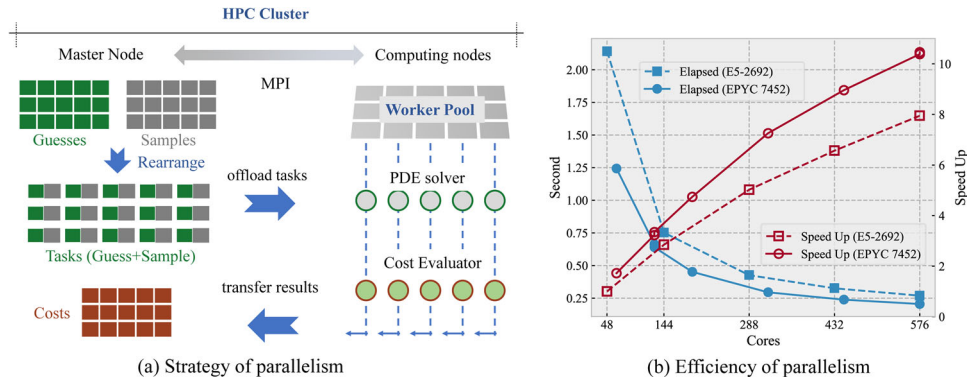


Fig. 8 Parallelism scheme of the cost evaluation kernel. **a** Schematic illustration of the parallelism mode for evaluating the cost of multiple guesses over large dataset with the support of MPI and HPC; **b** Parallelism efficiency of the cost evaluator with parallelism for a single batch compared between 2×12 Intel Xeon CPU E5-2692 v2 @ 2.20 GHz and 2×32 AMD EPYC 7452 @ 2.35 GHz, where 136 diffusion couples are consumed in a single batch using 1 MPI task per core.

firm tendency towards the acceleration of diffusion. Regarding the lower order systems, i.e., unary, binary, and ternary systems, the base plane rests around the middle among related tracer diffusion coefficients. It is thus concluded the entropy of configuration does not play a significant role in either hindering or accelerating the diffusion rates. In addition, the effect of averaging is rather obvious, as the deviation of the tracer diffusion coefficients shrinks as the number of components increases.

Owing to the absence of physical thermodynamic description, the effective tracer diffusion coefficients are further examined by comparing with the experimental results. The evaluated effective tracer diffusion coefficients are qualitatively compared to the ones measured by Tsai et al.⁶ and Vaidya et al.^{9,49}, as illustrated in Supplementary Figure 3. Among the tracer diffusion coefficients, D_{Mn}^* is dominant. The obtained effective tracer diffusion coefficient, i.e., D_{Mn}^* , accords well with the results from Tsai et al.⁶ and Vaidya et al.^{9,49} However, results measured by Tsai et al. tends to underestimate D_{Mn}^* especially among the range of lower temperature. D_{Cr}^* measured by Tsai et al. and the present work are similar to each other, both of which are larger than the experimental results by Vaidya et al. As for D_{Fe}^* , the results from Tsai et al. and Vaidya et al. agree well with each other, while the one measured in this work tends to be larger. Similar tendency applies to D_{Ni}^* with respect to D_{Fe}^* . When it comes to D_{Co}^* , Vaidya et al. come up with results that are smaller than the other two.

For the numerical inverse method, D_{Fe}^* and D_{Cr}^* are very similar, while D_{Co}^* and D_{Ni}^* are also similar. What's more, difference among D_{Fe}^* , D_{Co}^* , D_{Cr}^* and D_{Ni}^* measured by Vaidya et al. are trivial, especially for the lower temperature range. The results measured by Tsai et al. for D_{Fe}^* , D_{Cr}^* , and D_{Ni}^* are also trivial with respect to each other. Overall, dependency of tracer diffusion coefficients on temperature are rather similar for all the components, indicating the similarity of their thermodynamic and kinetic behaviors. Unfortunately, the prior assumption of being the ideal solution is not fully applicable to the thermodynamic description of the fcc CoCrFeMnNi HEAs. From the point view of numerical inverse method, more profound thermodynamic description is, therefore, expected for the desired tracer diffusion coefficients from the research community, for sake of producing generalizable tracer diffusion coefficients of physical reliability among large composition ranges. Focusing on the correlation between diffusion rates and configuration entropy, the effective tracer diffusion coefficients are qualitatively reasonable notwithstanding its physical validity.

For more convincing evidence, the interdiffusion coefficients are adopted for characterizing the diffusion behaviors of fcc CoCrFeMnNi system and its related subsystems. Despite the number of components, different constituents for systems with the same components also contribute to the variation of the

configurational entropy. To demonstrate the contribution of various constituents, the interdiffusion coefficients of fcc CoCr-FeMnNi systems projected over various composition coordinates are evaluated.

Taking Fig. 10a, f, k and p as an example, main terms of interdiffusion coefficient matrices at 1073 K, i.e., \bar{D}_{CoCo}^{Ni} , \bar{D}_{CrCr}^{Ni} , \bar{D}_{FeFe}^{Ni} , and \bar{D}_{MnMn}^{Ni} , are projected over the composition coordinate of Co ranging from 0 to 0.2, according to the first column of subfigures in Fig. 10. When $x_{Co} = 0$, the matrix denotes fcc CrFeMnNi alloy; while $x_{Co} = 0.2$, the matrix denotes fcc CoCrFeMnNi alloy. The related entropy of ideal mixing is also imposed on the x-axis on the bottom, which can be evaluated according to

$$S^{cnf} = -R \left[x_E \log(x_E) + (1 - x_E) \log\left(\frac{1 - x_E}{4}\right) \right] \quad (1)$$

Among Fig. 10a, f, k and p, x_E stands for the composition of component Co. With the increase of x_{Co} , the configurational entropy also increases. As shown, against the increment of entropy, \bar{D}_{CoCo}^{Ni} , \bar{D}_{CrCr}^{Ni} , \bar{D}_{FeFe}^{Ni} , and \bar{D}_{MnMn}^{Ni} decreases significantly, implying a trend of being hindered by entropy.

When the projection falls to the composition coordinate of Cr, the response of the tracer diffusion coefficients to the variation of entropies are flat considering \bar{D}_{CrCr}^{Ni} , shown in Fig. 10g. However, \bar{D}_{CoCo}^{Ni} , \bar{D}_{FeFe}^{Ni} , and \bar{D}_{MnMn}^{Ni} drops as the related configurational entropy increases. As for the projection over the composition coordinate of Fe, \bar{D}_{CoCo}^{Ni} , \bar{D}_{CrCr}^{Ni} , \bar{D}_{FeFe}^{Ni} , and \bar{D}_{MnMn}^{Ni} shows limited variety against the related configurational entropy, though implying a tendency of reduction presented in Fig. 10c, h, m and r. From the point view of Mn composition coordinate, all interdiffusion coefficients are roused up as the configurational entropy presents a tendency of rising, though the magnitude of increment for \bar{D}_{MnMn}^{Ni} is less significant, as illustrated in Fig. 10d, i, n and s. When it comes to the composition coordinate of Ni, the trending of \bar{D}_{CrCr}^{Ni} and \bar{D}_{FeFe}^{Ni} are rather flat with respect to the variation of entropy, referring to Fig. 10j and o. Moreover, controversial trends are observed for \bar{D}_{CoCo}^{Ni} and \bar{D}_{MnMn}^{Ni} , as the former drops while the latter rises up.

The composition ranges of quaternary systems are covered by current training dataset, and thus, the interpolated interdiffusion coefficients for the quaternary systems are considered properly generalized. Considering correlation for temperatures other than 1073 K, similar conclusions can be drawn according to Supplementary Figs 4–6. As tendency of various interdiffusion coefficients against various projection coordinates remains similar to that of 1073 K. Regarding temperatures other than the specific ones, the related diffusion rates can be inferred from the Arrhenius relation of different components and systems, as listed in Supplementary Table 1. However, among the concerned

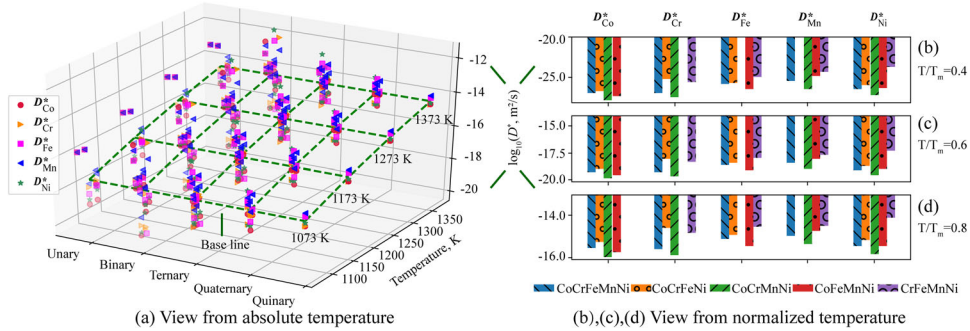


Fig. 9 Relation between diffusion rates and variation of configurational entropy. **a** View from absolute temperature: Deviation of the effective tracer diffusion coefficients over various matrices against the quinary system at different temperatures; **b** View from normalized temperature: Tracer diffusion coefficients of various components on the selected matrices at normalized temperatures.

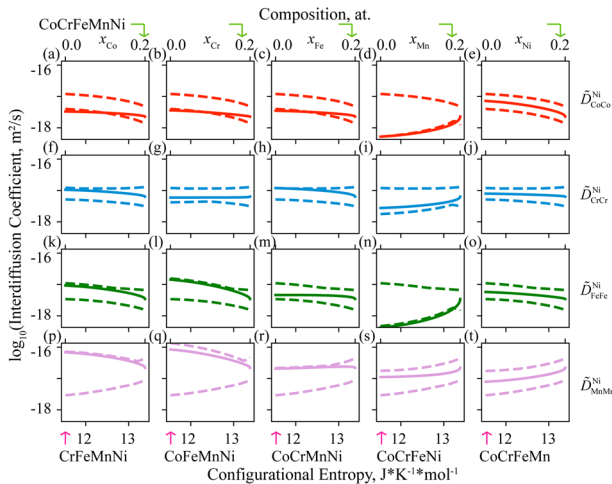


Fig. 10 Projections of interdiffusion coefficients along different composition coordinates. The dash lines refer to the bounds of the related interdiffusion coefficients (\tilde{D}_{CoCo}^{Ni} , \tilde{D}_{CrCr}^{Ni} , \tilde{D}_{FeFe}^{Ni} , \tilde{D}_{MnMn}^{Ni}) at 1073 K, while the solid lines are related to interdiffusion coefficients due to the maximum a-posteriori.

temperature range, the interdiffusion coefficients do not perceive the tendency of sluggish diffusion. That is, the observations above do not launch a firm correlation between the variation of entropy and interdiffusion coefficients. Unfortunately, without a comprehensive overview of the diffusion rates, it is prone to unilaterally attribute the deduction of diffusion rates to the increment of entropy, i.e., \tilde{D}_{CoCo}^{Ni} and \tilde{D}_{MnMn}^{Ni} in Fig. 10e and t respectively.

A portion of the previous investigations over the diffusion behavior of HEAs reported that under the normalized temperature scale, the diffusion rates of the systems with higher entropy would be smaller⁶. To further examine such a hypothesis, the effective tracer diffusion coefficients for various systems with equal atomic constituents at different normalized temperatures are evaluated for direct comparison. Assuming that alloys serve under $0.4T_m$ of the related systems, as shown in Fig. 9b, CoCrMnNi alloy is dominantly smaller than the rest, i.e., CoCrFeNi, CoFeMnNi, CrFeMnNi and CoCrFeMnNi alloys. Moreover, the effective tracer diffusion coefficients over the CoCrFeMnNi alloy rank beyond those of CoCrMnNi alloy, though the former is deemed as the one with higher entropy. Similar tendencies are found among the normalized temperatures, i.e., $T/T_m = 0.6$ and $T/T_m = 0.8$. Again, from the point of view in normalized temperature with respect to the melting point, the comparison result does not earn credit for the existence of sluggish diffusion.

Referring to averaging effect in Fig. 9a, the diffusion rates of various systems remain the same level of the fcc CoCrMnNi alloy. Among the concerned alloys, fcc CoCrMnNi alloy has the lowest melting temperature, i.e., 1500.82 K⁸, while the melting temperatures for the others alloys are CoCrFeNi (1711 K⁸), CoFeMnNi (1543 K⁸), CrFeMnNi(1620 K⁶⁰) and CoCrFeMnNi(1572 K⁸). It seems that the fcc CoCrMnNi system achieves the lowest diffusion rates at various normalized temperatures because of the lower melting point. That is, the normalized temperature, i.e., $0.4T_{CoCrMnNi}^m = 600$ K, is significantly smaller than that of fcc CoCrFeMnNi alloy, i.e., $0.4T_{CoCrFeMnNi}^m = 628$ K. As diffusion rate is subjected to the Arrhenius relation, it is not surprising that the alloy with lower melting point achieves lower diffusion rate with respect to the normalized temperature.

With the assessed diffusion descriptions of fcc CoCrFeMnNi system and its related subsystems, the concerned diffusion behaviors are able to be fully demonstrated by quantitative mathematical relations. As a conclusion due to rigor comparison, the sluggish diffusion of the fcc CoCrFeMnNi high-entropy alloy remains no more than thermo-physical state functions instead of mystery, which can be quantitatively evaluated with credible diffusion database found on large amount of experimental information.

METHODS

Numerical inverse method

Concerning the diffusion processes, with ad-hoc thermo-kinetic description, the predictions to diffusion behaviors of mass can be revealed by solving diffusion equations. To fulfill such ambitions, the thermo-kinetic description is indispensable. The idea of tuning the most suitable kinetic description that accounting for observations, i.e., the experimental composition profiles, lies with the inverse problem, namely the numerical inverse method. The inverse problem of kinetic description can be generally casted into the framework of the partial differential equation constrained optimization problem,

$$\boldsymbol{\theta} = \arg \min_{\boldsymbol{\theta}} \sum_i \|\mathbf{x}_i - \tilde{\mathbf{x}}_i(\boldsymbol{\theta})\|_2^2 \quad (2)$$

where \mathbf{x}_i denotes the experimental composition profile, while $\tilde{\mathbf{x}}_i$ the prediction to the composition profiles using the extended Fick's second law. In Eq. (2), the prediction needs to be produced by means of solving the diffusion equations, i.e.,

$$\frac{\partial x_k}{\partial t} = \nabla \cdot \left(\sum_{j=1}^{n-1} \tilde{D}_{kj}^n(\boldsymbol{\theta}) \nabla x_j \right) \quad (3)$$

for k ranging from 1 to n , where n is the number of concerned components in the system. Foremost, modeling of the interdiffusion coefficients is essential for construction of the forward problem and inverse problem. Currently, modeling of the interdiffusion coefficients is subject to the CALPHAD convention⁶¹, as detailed in Supplementary Methods.

With the numerical inverse method, the interdiffusion coefficients of the concerned systems can be retrieved. Atomic mobility parameters are also

available because the interdiffusion coefficients are parameterized following the CALPHAD convention. In the previous applications of numerical inverse method, both the interdiffusion coefficients and atomic mobility parameters are accessible, although the number of the diffusion couples involved in the calculation is less than 20. What's more, when interdiffusion coefficients serves as the target of calculation, HitDIC performs nicely recovering the interdiffusion coefficients for the lower systems. Regarding to the growing number of diffusion couples in concerned dataset, advanced techniques and strategies are introduced in the present work for pursuing diffusion database of high-quality.

Variable-selection genetic algorithm

We assume K as the competing parameters in the total parameter space \mathcal{M} and a subset of them generates the observations, noted as \mathbf{x} . Associated with all the parameters, there is a vector of parameters $\boldsymbol{\theta}$, i.e., $\theta_1, \theta_2, \dots, \theta_K$. We can introduce an additional vector of selection parameters \mathbf{k} , i.e., k_1, k_2, \dots, k_K . The objective is to identify the true subset as well as to estimate the parameters associated with the subset,

$$\boldsymbol{\theta}_{\text{MLE}} \mathbf{k}^T = \arg \max_{\mathbf{k}} p(\mathbf{k}, \boldsymbol{\theta} | \mathbf{x}) \quad (4)$$

where $p(\mathbf{k}, \boldsymbol{\theta} | \mathbf{x})$ is the posterior probability distribution. Each parameter of \mathbf{k} is the indicator that takes the value 1 when the associated parameter comes in to force and is 0 otherwise. According to the Bayes' rule, Eq. (2) is equivalent to Eq. (4), as

$$p(\mathbf{x} | \mathbf{k}, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\|\mathbf{x}_i - \bar{\mathbf{x}}_i(\boldsymbol{\theta}, \mathbf{k})\|_2^2}{2\sigma^2}\right) \quad (5)$$

and

$$p(\mathbf{k}, \boldsymbol{\theta} | \mathbf{x}) = p(\mathbf{k}, \boldsymbol{\theta}) p(\mathbf{x} | \boldsymbol{\theta}, \mathbf{k}) / p(\mathbf{x}) \quad (6)$$

where $p(\mathbf{x} | \mathbf{k}, \boldsymbol{\theta})$ is the likelihood function. Here, σ is the variance of the residual between the predictions and observations. Generally, $p(\mathbf{x} | \boldsymbol{\theta}, \mathbf{k}) / p(\boldsymbol{\theta}, \mathbf{k})$ is taken as constant, though $p(\mathbf{x} | \boldsymbol{\theta}, \mathbf{k})$ and $p(\boldsymbol{\theta}, \mathbf{k})$ are not explicitly accessible.

In most genetic algorithms, only two main components are of problem dependence⁶², i.e., problem encoding and evaluation function. In order to accommodate the problem of parameter selection and parameter estimation for automatic evaluation of interdiffusion coefficients and atomic mobilities, the binary encoding is adopted in the present work,

$$\text{binary string} = H(\boldsymbol{\theta}, \mathbf{k}) \quad (7)$$

where H is the proposed encoding function. The proposed encoding pattern is schematically illustrated in Fig. 1b, where the selection parameters are laid out after the model parameters. Each selection parameter takes only one *allele*, while the number of bits occupied by individual model parameter may vary according to desired precision. The laid-out of the encoding string has potential influence on the effects of the operators of the genetic algorithm⁶³, i.e., mutation and crossover. The crossover operator includes different schemata, i.e., single-point, two-point, uniform and arithmetic crossover, while the single-point/two-point crossover operator is adopted in the present work to retain the robustness of the selection and optimization processes. To endow sufficient ability for evolution, the mutation is uniformly considered for all bits via bit inversion with a ratio of mutation about 1%.

The canonical genetic algorithm is responsible for parameter estimation regarding to the evaluation and fitness of the problem. To fulfill the target of parameter selection, the evaluation or the objective function must be surrogated by introducing penalty on the number of effective model parameters. As one of the fancy evolutionary algorithms, the genetic algorithm possesses the features of being scalable and being flexible to consider many criteria in the optimization processes. Referring to solution with exhaustive selection and scoring scheme, the potential criteria, i.e., F-test, information criteria and regularization, are the potential options. In general, the fitness function is used for genetic algorithm for measuring the driving force for evolution. Fitness is different from the objective functions, noted as $\text{OBJ}(\boldsymbol{\theta})$. The RSS, i.e., Eq. (2), is one of most popular options for the objective function, where the least RSS is generally pursued. Fitness function represents the probability of survival for the population of the solutions, therefore, larger fitness values are more desirable for the selection operators. Proper conversion is thus in need between the objective functions and the fitness function. For a population

with P individuals, the fitness function, F , can be defined as

$$F_i = 1 - \frac{\text{OBJ}_i - \text{OBJ}_{\min}}{\text{OBJ}_{\max} - \text{OBJ}_{\min}} \quad (8)$$

With the fitness function, the selection operator for genetic algorithm can therefore be conducted using the roulette over the fitness sequence.

One of the simplest and most convenient objective functions regarding the parameter selection is the information criterion, which concerns the model complexity and fitting goodness simultaneous, i.e., the AIC or BIC,

$$\text{OBJ}_{\text{AIC}} = 2K - 2\ln(\hat{L}) \quad (9)$$

or

$$\text{OBJ}_{\text{BIC}} = K\ln(N) - 2\ln(\hat{L}) \quad (10)$$

where K is the number of the effective parameters, N is the number of observations and \hat{L} is maximum value of the likelihood function. Generally, \hat{L} cannot be directly accessible, though it can be related to RSS as detailed in Supplementary Methods. Meanwhile, a study case for benchmarking is available in Supplementary Discussion.

Regularization optimizer with automatic hyper-parameter tuning

The regularization is generally served as a powerful tool for preventing the overfitting while improving the generality of the assessed model and also estimated parameters^{64,65}. It is an important concept in the inverse problem, machine learning and so on. The most common strategy for regularization is to construct a surrogated objective function by introducing penalty on L_1/L_2 norm to the original objective function, i.e., Eq. (2). With a regularization term λ , the objective function can be reformulated as

$$\text{OBJ}_{L_1} = \text{RSS} + \lambda \|\boldsymbol{\theta}\| \quad (11)$$

for L_1 regularization, and as

$$\text{OBJ}_{L_2} = \text{RSS} + \lambda \|\boldsymbol{\theta}\|_2^2 \quad (12)$$

for L_2 regularization. By determining an appropriate scale for the regularization term, the results that balance the extrapolation and explanatory capability of the proposed parameters/models are to be acquired, when the arguments of the minimum of the surrogated objective functions are resolved.

However, the regularization term is a tricky hyper-parameter, which deserves meticulous tuning^{64,66,67}. An algorithm for automatically tuning the regularization term is proposed, as shown in Fig. 1c. Firstly, the estimated effective parameters and their estimations are imported from the variable-selection genetic algorithm. The least RSS ported from variable-selection GA is then used to estimate initial regularization term. The workflow then proceeds into a subroutine where the most appropriate estimations, i.e., $\boldsymbol{\theta}_{\text{ReL1}}$ or $\boldsymbol{\theta}_{\text{ReL2}}$, are pursued until no significant change takes place between subsequent iterations. The RSS value will be double-checked to verify whether there is a significant increase in RSS of the training dataset. The workflow will be terminated once the current RSS value surpasses the least RSS at a certain degree. Or the regularization term will be increased, and the subroutine to determine new alternative estimations would be repeated. During the iterations, the least RSS will be updated as if a new alternative with smaller RSS occurs.

Metropolis–Hastings sampler with multiple independent chains

For the nonlinear inverse problem, the Bayesian inference might be the only tool available for quantifying the uncertainty of the concerned model and parameters. MCMC is one of the important interference tools based on the Bayes' rule⁶⁸,

$$p(\boldsymbol{\theta} | \mathbf{x}) = p(\boldsymbol{\theta}) p(\mathbf{x} | \boldsymbol{\theta}) / p(\mathbf{x}) \quad (13)$$

The posterior distribution, i.e., $p(\boldsymbol{\theta} | \mathbf{x})$, is generally not able to produce estimations, i.e., θ , directly, however, it can be employed in a reversible Markov process during the Monte Carlo simulation. The samples are then drawn from the target distribution, while the posterior estimations can, therefore, be obtained. However, the Bayesian inference with the MCMC method is generally challenging for model with large parameter space and dataset of large sample size, considering the cost of time and computing expenses⁶⁹. This is partly due to the computational cost of such methods, since the evaluation of the objective function, i.e., Eq. (5) or Eq. (13), is

generally computing expensive. Sufficient random walks are expected such that the obtained posterior distribution reaches a stationary state, which might require tens or hundreds of times of iterations more than the optimization or regularization processes.

In the naive Metropolis–Hastings algorithm, the posterior distribution is evaluated with respect to all the samples in the dataset, assuming that all samples, i.e., \mathbf{x} , are independently measured. In addition, the multiple independent chains, i.e., M , are employed in the present work to draw samples from the posterior distribution of m parameters. One very long chain for MCMC is not applicable for a tall dataset due to time efficiency, and thus increasing the number of chains and running the chains in parallelism would be a promising alternative. The overall workflow of the developed sampling kernel is illustrated in Fig. 1d. It is worthy of mentioning that the initial estimations for the MCMC sequences are ported from those produced by regularization processes in order to let the MCMC chains locate around the high probability regions.

Implicit solver for multicomponent diffusion equations

In the inverse problem, the solver to the forward problem is extremely essential for the inverse process. To ensure the stability of the forward simulation process, a fully implicit finite difference scheme is applied to relax the stable condition constrained by step sizes of space and time. Prior to the demonstration of the proposed scheme, benchmarks are available in Supplementary Methods.

On a one-dimension domain representing a diffusion couple, taking the i -th grid node as an example, the conjugated grids are h_{i-1} and h_{i+1} , and h_i^- and h_i^+ are the spacings before and behind the current grid node. Imposing finite difference scheme on Eq. (3), the recursive formula for the p -th component of a system with M components can be formulated

$$\begin{aligned} & -c_{p,j-1}^{t+1} \left[\frac{2D_{pp,j-1/2}^{t+1}}{h_i^-(h_i^-+h_i^+)} \right] + c_{p,j}^{t+1} \left[\frac{1}{\Delta t} + \frac{2D_{pp,j-1/2}^{t+1}}{h_i^-(h_i^-+h_i^+)} + \frac{2D_{pp,j-1/2}^{t+1}}{h_i^-(h_i^-+h_i^+)} \right] - c_{p,j+1}^{t+1} \left[\frac{2D_{pp,j+1/2}^{t+1}}{h_i^+(h_i^++h_i^+)} \right] \\ & = \frac{1}{\Delta t} c_{p,j}^t + \sum_{q \neq p}^{M-1} \left[\frac{2D_{pq,j-1/2}^{t+1}}{h_i^-(h_i^-+h_i^+)} c_{q,j-1}^{t+1} - \left(\frac{2D_{pq,j-1/2}^{t+1}}{h_i^-(h_i^-+h_i^+)} - \frac{2D_{pq,j-1/2}^{t+1}}{h_i^-(h_i^-+h_i^+)} \right) \right. \\ & \quad \left. c_{q,j}^{t+1} + \frac{2D_{pq,j+1/2}^{t+1}}{h_i^+(h_i^++h_i^+)} c_{q,j+1}^{t+1} \right] \end{aligned} \quad (14)$$

where Δt is the time step. To fulfill numerical simulation, the coefficient terms can be rewritten in the form of matrix and vector,

$$A_{pp,j,i}^{t+1} = \frac{2D_{pp,j+1/2}^{t+1}}{h_i^+(h_i^++h_i^+)} + \frac{2D_{pp,j-1/2}^{t+1}}{h_i^-(h_i^-+h_i^+)} \quad (15)$$

$$A_{pp,j,i-1}^{t+1} = \frac{2D_{pp,j-1/2}^{t+1}}{h_i^-(h_i^-+h_i^+)} \quad (16)$$

$$A_{pp,j,i+1}^{t+1} = \frac{2D_{pp,j+1/2}^{t+1}}{h_i^+(h_i^++h_i^+)} \quad (17)$$

$$\begin{aligned} b_{p,j}^{t+1} &= \frac{1}{\Delta t} c_{p,j}^t + \sum_{q \neq p}^{M-1} \left[\frac{2D_{pq,j-1/2}^{t+1}}{h_i^-(h_i^-+h_i^+)} c_{q,j-1}^{t+1} - \left(\frac{2D_{pq,j-1/2}^{t+1}}{h_i^-(h_i^-+h_i^+)} - \frac{2D_{pq,j-1/2}^{t+1}}{h_i^-(h_i^-+h_i^+)} \right) \right. \\ & \quad \left. c_{q,j}^{t+1} + \frac{2D_{pq,j+1/2}^{t+1}}{h_i^+(h_i^++h_i^+)} c_{q,j+1}^{t+1} \right] \end{aligned} \quad (18)$$

where \mathbf{A}_{pp}^{t+1} is the coefficient matrix on the left-hand side of Eq.(14), \mathbf{b}_p^{t+1} is the right-hand side of Eq.(14). For zero-flux boundary condition,

$$A_{pq,0,0}^{t+1} = 1, A_{ij,0,1}^{t+1} = -1, b_{p,0}^t = 0 \quad (19)$$

and

$$A_{pq,N-1,N-1}^{t+1} = 1, A_{pq,N-1,N-2}^{t+1} = -1, b_{p,N-1}^t = 0 \quad (20)$$

Equations (14)–(20) subject to a set of linear equations, where the coefficient matrices are in the form of tri-diagonal matrices, i.e.,

$$\mathbf{A}_{pp}^{t+1} \mathbf{c}_p^{t+1} = \mathbf{b}_p^{t+1} \quad (21)$$

which can be solved easily with tri-diagonal matrix algorithm. In case that \mathbf{A}_{pp}^{t+1} and \mathbf{b}_p^{t+1} are implicit, additional operations are required to estimate the two terms for each time step as illustrated in the Algorithm 1 of Supplementary Methods.

Parallelism of cost evaluator

The bottleneck of the efficiency for the numerical inverse method lies in the time-consuming process of evaluating the objective function. In the framework of HitDIC software, a cost evaluator is responsible for calling the solver to diffusion equations to produce predictions, and calculating the deviation between the predictions and observations for the output of the objective function. For the sample dataset of large size, the evaluation of objective function is computationally expensive, which may be unfeasible without HPC. To assess the computing resources on HPC, the parallelism with MPI technique is adopted for the cost evaluator, as illustrated in Fig. 8a. The proposed parallelism mechanism behind the cost evaluator is suitable for genetic algorithm, MCMC with multiple chains and so on. Taking the genetic algorithm as an example, each iteration is required to evaluate cost for multiple guesses in a population, while the cost evaluator for each guess relies on the predictions to multiple samples in the training dataset. More specifically, the guesses and the samples are firstly rearranged to form the sequence of tasks in the master node. Once tasks are ready, the signal from the master node will be sent to the workers in the worker pool, while the workers will then offload tasks from master node repeatedly. In an active worker process, the cost evaluator would be executed in sequence, while the value of objective function will then be transferred back to the master. State of the worker will then be flushed and it will wait in the pool for the remaining tasks. The master node is responsible for offloading tasks and collecting the results to/from the computing nodes, and returning the results to different optimization solvers.

DATA AVAILABILITY

The key data that support the findings within this paper can be found at the GitHub address https://github.com/zhongjingjogy/fcc_CoCrFeMnNi, and other data are available from the corresponding author upon reasonable request.

CODE AVAILABILITY

Related algorithms are bundled in the HitDIC software, the latest version of which can be accessible from <https://hitdic.com>.

Received: 24 August 2020; Accepted: 14 January 2021;

Published online: 01 March 2021

REFERENCES

1. Takaki, T. et al. Primary arm array during directional solidification of a single-crystal binary alloy: large-scale phase-field study. *Acta Mater.* **118**, 230–243 (2016).
2. Reed, R. C. *The Superalloys: Fundamentals and Applications*. (Cambridge University Press, 2006).
3. Yeh, J.-W. et al. Nanostructured high-entropy alloys with multiple principal elements: novel alloy design concepts and outcomes. *Adv. Eng. Mater.* **6**, 299–303 (2004).
4. Ta, N., Zhang, L., Li, Q., Lu, Z. & Lin, Y. High-temperature oxidation of pure Al: kinetic modeling supported by experimental characterization. *Corros. Sci.* **139**, 355–369 (2018).
5. Clarke, D. R., Oechsner, M. & Padture, N. P. Thermal-barrier coatings for more efficient gas-turbine engines. *MRS Bull.* **37**, 891–898 (2012).
6. Tsai, K.-Y., Tsai, M.-H. & Yeh, J.-W. Sluggish diffusion in Co–Cr–Fe–Mn–Ni high-entropy alloys. *Acta Mater.* **61**, 4887–4897 (2013).
7. Kucza, W. et al. Studies of “sluggish diffusion” effect in Co–Cr–Fe–Mn–Ni, Co–Cr–Fe–Ni and Co–Fe–Mn–Ni high entropy alloys; determination of tracer diffusivities by combinatorial approach. *J. Alloy. Compd.* **731**, 920–928 (2018).
8. Dąbrowa, J. et al. Demystifying the sluggish diffusion effect in high entropy alloys. *J. Alloy. Compd.* **783**, 193–207 (2019).
9. Vaidya, M., Pradeep, K. G., Murty, B. S., Wilde, G. & Divinski, S. V. Bulk tracer diffusion in CoCrFeNi and CoCrFeMnNi high entropy alloys. *Acta Mater.* **146**, 211–224 (2018).
10. Chen, S., Li, Q., Zhong, J., Xing, F. & Zhang, L. On diffusion behaviors in face centered cubic phase of Al–Co–Cr–Fe–Ni–Ti high-entropy superalloys. *J. Alloy. Compd.* **791**, 255–264 (2019).
11. Chen, W. & Zhang, L. High-throughput determination of interdiffusion coefficients for Co–Cr–Fe–Mn–Ni high-entropy alloys. *J. Phase Equilib. Diffus.* **38**, 457–465 (2017).
12. Wang, R., Chen, W., Zhong, J. & Zhang, L. Experimental and numerical studies on the sluggish diffusion in face centered cubic Co–Cr–Cu–Fe–Ni high-entropy alloys. *J. Mater. Sci. Technol.* **34**, 1791–1798 (2018).

13. Choi, W.-M., Jo, Y. H., Sohn, S. S., Lee, S. & Lee, B.-J. Understanding the physical metallurgy of the CoCrFeMnNi high-entropy alloy: an atomistic simulation study. *NPJ Comput. Mater.* **4**, 1 (2018).
14. Dąbrowa, J. & Danielewski, M. State-of-the-art diffusion studies in the high entropy alloys. *Metals* **10**, 347 (2020).
15. Divinski, S. V., Pokoev, A. V., Esakiraja, N. & Paul, A. A mystery of “sluggish diffusion” in high-entropy alloys: the truth or a myth? *Diffus. Found.* **17**, 69–104 (2018).
16. Zhang, C. et al. Understanding of the elemental diffusion behavior in concentrated solid solution alloys. *J. Phase Equilib. Diffus.* **38**, 434–444 (2017).
17. Beke, D. & Erdélyi, G. On the diffusion in high-entropy alloys. *Mater. Lett.* **164**, 111–113 (2016).
18. Zhong, J., Chen, L. & Zhang, L. High-throughput determination of high-quality interdiffusion coefficients in metallic solids: a review. *J. Mater. Sci.* **55**, 10303–10338 (2020).
19. Matano, C. On the relation between the diffusion-coefficients and concentrations of solid metals. *Jpn. J. Appl. Phys.* **8**, 109–113 (1933).
20. Wagner, C. The evaluation of data obtained with diffusion couples of binary single-phase and multiphase systems. *Acta Metall.* **17**, 99–107 (1969).
21. Sauer, F. & Freise, V. Diffusion in binären Gemischen mit Volumenänderung. *Z. für Elektrochemie, Ber. der Bunsenges. für. Physikalische Chem.* **66**, 353–362 (1962).
22. Kirkaldy, J. S. & Young, D. J. *Diffusion in the Condensed State*. (Institute of Metals, London, 1987).
23. Whittle, D. & Green, A. The measurement of diffusion coefficients in ternary systems. *Scr. Metall.* **8**, 883–884 (1974).
24. Paul, A. A pseudobinary approach to study interdiffusion and the Kirkendall effect in multicomponent systems. *Philos. Mag.* **93**, 2297–2315 (2013).
25. Esakiraja, N. & Paul, A. A novel concept of pseudo ternary diffusion couple for the estimation of diffusion coefficients in multicomponent systems. *Scr. Mater.* **147**, 79–82 (2018).
26. Esakiraja, N., Pandey, K., Dash, A. & Paul, A. Pseudo-binary and pseudo-ternary diffusion couple methods for estimation of the diffusion coefficients in multicomponent systems and high entropy alloys. *Philos. Mag.* **99**, 2236–2264 (2019).
27. Zhao, J.-C., Zheng, X. & Cahill, D. G. High-throughput diffusion multiples. *Mater. Today* **8**, 28–37 (2005).
28. Xu, H. et al. Determination of accurate interdiffusion coefficients in fcc Ag-In and Ag-Cu-In alloys: a comparative study on the Matano method with distribution function and the numerical inverse method with HitDIC. *J. Alloy. Compd.* **798**, 26–34 (2019).
29. Kodentsov, A. A., Bastin, G. F. & van Loo, F. J. J. in *Methods for Phase Diagram Determination* 222–245 (Elsevier, 2007).
30. Kodentsov, A. & Paul, A. in *Handbook of Solid State Diffusion*, Vol 2 207–275 (Elsevier, 2017).
31. Chen, W., Zhang, L., Du, Y., Tang, C. & Huang, B. A pragmatic method to determine the composition-dependent interdiffusivities in ternary systems by using a single diffusion couple. *Scr. Mater.* **90–91**, 53–56 (2014).
32. Chen, W., Zhong, J. & Zhang, L. An augmented numerical inverse method for determining the composition-dependent interdiffusivities in alloy systems by using a single diffusion couple. *MRS Commun.* **6**, 295–300 (2016).
33. Kuczka, W. A combinatorial approach for extracting thermo-kinetic parameters from diffusion profiles. *Scr. Mater.* **66**, 151–154 (2012).
34. Bouchet, R. & Mevrel, R. A numerical inverse method for calculating the interdiffusion coefficients along a diffusion path in ternary systems. *Acta Mater.* **50**, 4887–4900 (2002).
35. Chen, Z., Zhang, Q. & Zhao, J.-C. pydiffusion: A Python library for diffusion simulation and data analysis. *J. Open Res. Softw.* **7**, 13 (2019).
36. Gaertner, D. et al. Concentration-dependent atomic mobilities in FCC CoCr-FeMnNi high-entropy alloys. *Acta Mater.* **166**, 357–370 (2019).
37. Zhang, Q. & Zhao, J.-C. Extracting interdiffusion coefficients from binary diffusion couples using traditional methods and a forward-simulation method. *Intermetallics* **34**, 132–141 (2013).
38. Biegler, L. et al., eds. *Large-Scale Inverse Problems and Quantification of Uncertainty*. (John Wiley & Sons, 2011).
39. Chung, J., Knepper, S. & Nagy, J. G. in *Handbook of Mathematical Methods in Imaging* 47–90 (Springer New York, 2015).
40. Cullen, M., Freitag, M. A., Kindermann, S., & Scheichl, R. eds. *Large Scale Inverse Problems: Computational Methods and Applications in the Earth Sciences*. (De Gruyter, 2013).
41. Zhang, L. & Chen, Q. in *Handbook of Solid State Diffusion*, Vol.1. 321–362 (Elsevier, 2017).
42. Olson, G. B. & Kuehmann, C. J. Materials genomics: from CALPHAD to flight. *Scr. Mater.* **70**, 25–30 (2014).
43. National Research Council, Division on Engineering and Physical Sciences, National Materials Advisory Board & Committee on Integrated Computational Materials Engineering. *Integrated Computational Materials Engineering: A Transformational Discipline for Improved Competitiveness and National Security*. (National Academies Press, 2008).
44. Nikolaev, P. et al. Autonomy in materials research: a case study in carbon nanotube growth. *NPJ Comput. Mater.* **2**, 1–6 (2016).
45. Ozaki, Y. et al. Automated crystal structure analysis based on blackbox optimization. *NPJ Comput. Mater.* **6**, 75 (2020).
46. Li, Q. et al. On sluggish diffusion in Fcc Al–Co–Cr–Fe–Ni high-entropy alloys: an experimental and numerical study. *Metals* **8**, 16 (2017).
47. Chen, J. & Zhang, L. Composition-dependent interdiffusivity matrices in face centered cubic Ni–Al–X (X = Rh and W) alloys at 1423, 1473 and 1523 K: A high-throughput experimental measurement. *Calphad* **60**, 106–115 (2018).
48. Cantor, B., Chang, I. T. H., Knight, P. & Vincent, A. J. B. Microstructural development in equiatomic multicomponent alloys. *Mater. Sci. Eng. A* **375–377**, 213–218 (2004).
49. Vaidya, M., Trubel, S., Murty, B. S., Wilde, G. & Divinski, S. V. Ni tracer diffusion in CoCrFeNi and CoCrFeMnNi high entropy alloys. *J. Alloy. Compd.* **688**, 994–1001 (2016).
50. Kulkarni, K. & Chauhan, G. P. S. Investigations of quaternary interdiffusion in a constituent system of high entropy alloys. *AIP Adv.* **5**, 097162 (2015).
51. Verma, V., Tripathi, A. & Kulkarni, K. N. On interdiffusion in FeNiCoCrMn high entropy alloy. *J. Phase Equilib. Diffus.* **38**, 445–456 (2017).
52. Wang, R. *On the Determination of Diffusion Coefficients and Sluggish Diffusion Effect of Face-centered Cubic Co-Cr-Fe-Ni-X(X=Mn,Cu) High Entropy Alloys*. (Central South University, 2018).
53. Zhong, J., Chen, W. & Zhang, L. HitDIC: a free-accessible code for high-throughput determination of interdiffusion coefficients in single solution phase. *Calphad* **60**, 177–190 (2018).
54. Wei, M. & Zhang, L. Application of distribution functions in accurate determination of interdiffusion coefficients. *Sci. Rep.* **8**, 5071 (2018).
55. Zhong, J., Zhang, L., Wu, X., Chen, L. & Deng, C. A novel computational framework for establishment of atomic mobility database directly from composition profiles and its uncertainty quantification. *J. Mater. Sci. Technol.* **48**, 163–174 (2020).
56. McCall, J. Genetic algorithms for modelling and optimisation. *J. Comput. Appl. Math.* **184**, 205–222 (2005).
57. Kochenderfer, M. J. *Decision Making Under Uncertainty: Theory and Application*. (MIT Press, 2015).
58. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *NPJ Comput. Mater.* **5**, 21 (2019).
59. Yeh, J.-W. Recent progress in high entropy alloys. *Ann. Chim. Sci. Mat.* **31**, 633–648 (2006).
60. Lederer, Y., Toher, C., Vecchio, K. S. & Curtarolo, S. The search for high entropy alloys: a high-throughput ab-initio approach. *Acta Mater.* **159**, 364–383 (2018).
61. Andersson, J. & Ågren, J. Models for numerical treatment of multicomponent diffusion in simple phases. *J. Appl. Phys.* **72**, 1350–1355 (1992).
62. Whitley, D. A genetic algorithm tutorial. *Stat. Comput.* **4**, 65–85 (1994).
63. Bhandari, D., Murthy, C. & Pal, S. K. Genetic algorithm with elitist model and its convergence. *Int. J. Pattern Recogn.* **10**, 731–747 (1996).
64. Poggio, T., Torre, V. & Koch, C. Computational vision and regularization theory. *Nature* **317**, 314–319 (1985).
65. Girosi, F., Jones, M. & Poggio, T. Regularization theory and neural networks architectures. *Neural Comput.* **7**, 219–269 (1995).
66. Hansen, P. C. & O’Leary, D. P. The use of the L-curve in the regularization of discrete ill-posed problems. *SIAM J. Sci. Comput.* **14**, 1487–1503 (1993).
67. Engl, H. W., Hanke, M. & Neubauer, A. *Regularization of Inverse Problems*. Vol. 375 (Springer Science & Business Media, 1996).
68. Hewson, P. Statistical rethinking: a Bayesian course with examples in R and Stan. *J. R. Stat. Soc. A Stat.* **179**, 1131 (2016).
69. Robert, C. P., Elvira, V., Tawn, N. & Wu, C. Accelerating MCMC algorithms. *WIREs Comput. Stat.* **10**, e1435 (2018).

ACKNOWLEDGEMENTS

The financial support from the National MCF Energy R&D Program of China (Grant No. 2018YFE0306100), the Guangdong Province Key-Area Research and Development Program of China (2019B010943001), the Youth Talent Project of Innovation-driven Plan at Central South University (Grant No. 2282019SYLB026), and the Hunan Provincial Science and Technology Program of China (Grant No. 2017RS3002)-Huxiang Youth Talent Plan is acknowledged. Jing Zhong acknowledges the support from the Fundamental Research Funds for the Central Universities of Central South University (Grant No. 2018zts129). This is also part of Dr. Jing Zhong’s post-doctoral research work at Central South University, China. Jing Zhong acknowledges Dr. Richard Otis for discussing the strategy for parameter selection during the 47th CALPHAD conference in Mexico.

AUTHOR CONTRIBUTIONS

L.Z. conceived the presented idea and provided necessary materials. J.Z. designed, developed, and maintained the HitDIC infrastructure, data management system. J.Z.

and L.Z. wrote the manuscript. All the authors, J.Z., L.C., and L.Z., discussed the results and commented on the manuscript.

COMPETING INTERESTS

HitDIC is a free-accessible software independently developed by current authors. The proposed algorithms and strategies are bundled with HitDIC, while the related results are computed with HitDIC and its related infrastructures. Therefore, the authors have no more competing interests to clarify.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-021-00500-0>.

Correspondence and requests for materials should be addressed to L.Z.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021