CoSMoR: Decoding decision-making process along continuous composition pathways in machine learning models trained for material properties

Dishant Beniwal* and Pratik K. Ray
Department of Metallurgical and Materials Engineering, Indian Institute of Technology Ropar, Rupnagar 140001, Punjab, India

(Received 22 November 2022; revised 17 March 2023; accepted 5 April 2023; published 21 April 2023)

A key challenge in materials informatics is to decode the decision-making process of machine learning (ML) models that have been trained to predict material properties. The existing methods usually rank alloy features based on importance metrics and do not provide material-specific fundamental insights. Here, we present the Compositional Stimulus and Model Response (CoSMoR) framework that can be applied to any composition-based ML model (irrespective of the algorithm used) to calculate the exact contribution of each feature towards the manifestation of target material property along a continuous compositional pathway. CoSMoR utilizes the local partial dependencies of target property with respect to each feature and combines it with feature variations associated with discretized compositional variations to measure exact feature contributions. We showcase the importance of CoSMoR through implementation on phase-selection problem in multiprincipal element alloys (MPEAs), wherein it leads to physical insights into phase transitions. A detailed overview of the framework, along with the codes and step-by-step implementation of the algorithm, has been provided to enable extension to new or preexistent models.

DOI: 10.1103/PhysRevMaterials.7.043802

I. INTRODUCTION

The application of machine learning (ML) in materials science has seen tremendous growth in recent years [1–6]. Compositional ML models, which use a combination of composition and elemental properties as input features, have been used extensively for the prediction of a wide variety of materials phenomena such as phase selection [7–11], mechanical properties [12–14], oxidation behavior [15–19], structure stability [20,21], and alloy discovery [22–28]. Given the success of these ML models even on unseen compositional space, there is a strong possibility that these models are capturing the underlying physical principles using the input material descriptors, even though these decisions are often hidden due to the complicated form assumed by ML models (especially deep neural networks). This introduces exciting avenues for uncovering physical insights from the trained ML models through the decoding of their decision-making process. Here, we have developed the **Co**mpositional **S**timulus and **Mo**del **R**esponse (CoSMoR) framework that discretizes the compositional space and calculates the exact feature contributions along any given compositional pathway. This is done by combining the partial-local dependence (PLD) of ML model (with respect to each feature) with the sensitivity of that feature (with respect to the composition) at each composition step.

With the growing applicability and reliability of ML for materials science, the interpretability of these models has appeared at the forefront in recent years [29–33]. As discussed eloquently by Lipton [34] and Oviedo *et al.* [31], the term "interpretability," and the expectations surrounding it, can be quite subjective since it is associated with a great deal of technical jargon such as explainability, simulability, decomposability, algorithmic transparency, understandability, etc. Thus, it is imperative to define *a priori* the expectations for any framework that aims at decoding the decision-making process of ML models. In this regard, we describe here the four attributes of interpretation obtained from CoSMoR framework:

(i) *Type of explanation*: Suppose we have a base alloy, B. If we start adding another component, A, to this base alloy, i.e., we move along A_xB_{1-x} composition pathway, then CoSMoR provides the exact contribution of each feature towards the changes in ML predicted property with respect to the base alloy composition.

(ii) *Correctness of model explanation:* The feature contributions calculated by CoSMoR along a composition pathway are not relative or indicative metrics of feature importance, but are instead quantitatively exact with respect to the model decision-making process. For example, suppose we have a hardness ML model that predicts an increase of 100 HV when 10 at.% Al is added to CrFeNi base alloy; then, implementation of CoSMoR will tell exactly how much each feature contributed towards this overall increase of 100-HV hardness as predicted by the ML model.

(iii) *Causation for model understanding*: The existing interpretation methodologies such as partial-dependence plots, accumulated local effects, Local interpretable model-agnostic explanations, and SHapley Additive exPlanations (SHAP) provide an understanding of how important each feature is towards the overall model decision-making. But, from an alloy design perspective, it is much more meaningful to understand the model decision-making with respect to compositional variations because: (1) the direct point of control in alloy design is the elemental composition, not the feature values; (2) the features cannot be varied independently since any alloying

*Corresponding author: dishant.21mmz0005@iitrpr.ac.in

addition will affect all features; and (3) the features cannot be changed by any arbitrary amount or to any arbitrary set of values since the elements have a fixed set of properties and we have direct control over only the alloy composition. Since CoSMoR probes feature contributions as a function of composition, the causality for model understanding obtained from CoSMoR is rooted in compositional variations rather than arbitrary feature variations.

(iv) *Scope of explanation (local/global):* In the context of compositional ML models, the scope of explanation from any interpretation framework may be defined as (1) local if the understanding obtained is at a single composition value for example, calculating the partial dependence of model output on each feature for a given alloy (similar to SHAP values), and (2) global if the understanding obtained reflects in general how the feature manifests in the model decision-making, for example, formulation of surrogate models with simplified and interpretable mathematical forms. While local explanations tend to be more accurate with respect to the model decision-making process, they are valid only for a fixed composition value. On the other hand, the global explanations can span a much more expansive compositional space, but the understanding becomes generalized, though approximate. CoSMoR aims at combining the best of these two approaches to provide material-specific insights. Fundamentally, the explanation obtained from CoSMoR is local in nature as it calculates feature contributions for each composition step at a time. But, the scale of understanding is not limited to a single composition point, and instead a complete continuous composition pathway can be probed. Moreover, composition pathways defined as A_xB_{1-x} can span a considerable compositional space since components A and B can be either elements or any stoichiometric combination of elements.

II. METHODOLOGY FOR DEVELOPMENT OF CoSMoR

In CoSMoR framework, the compositional space is represented by atomic fraction (and not weight fraction). Further, it is discretized using a composition step-size variable denoted by Δc. This is a user input parameter; for e.g., in Sec. IV (where CoSMoR has been applied to phase selection in MPEAs), a composition step-size value of 0.01 has been used that represents 0.01 at. fraction (i.e., 1 at.%). Consider a hypothetical compositional ML model that takes F number of input features (denoted as $X_i,\ i \in [1, 2, 3, \ldots F]$) to predict a single target parameter Y. At any composition c, we have a set of all feature values. CoSMoR is built on the hypothesis that if the composition changes by a discrete amount Δc, then the resultant change in prediction Y can be represented exactly as an accumulation of contributions from all features, i.e.,

$$[\Delta Y]_{c\to c+\Delta c} = \sum_{i=1}^{F} [\Delta Y(X_i)]_{c\to c+\Delta c}. \tag{1}$$

The partial dependence (PD) of ML model on any feature is reflected by the sensitivity of the ML model to an independent change in that feature. But, this PD value depends on the value of all the other features also, i.e., at different compositions the PD values will be different. Thus, the term partial-local dependence is used to denote the fact that any PD calculation is valid only locally (i.e., at a particular composition value only). As shown schematically in Fig. 1(a), PLD of the model with respect to feature X_i at any composition c is calculated by changing the value of feature X_i by a small amount δX_i, while keeping all the other features the same, and measuring the change in model output δY. Thus, the PLD (denoted as $m_c^{X_i}$) with respect to feature X_i at composition c is calculated as $m_c^{X_i} = (\frac{\delta Y}{\delta X_i})_c$. Here, δX_i is a user-input parameter whose value will depend on the scale of each feature. If all features were normalized to a uniform scale (as recommended for ML model development), the same value of δX_i may be used for all features; for, e.g., in Sec. IV, we used a value of 0.02 for δX_i since all features were normalized to [0, 1] scale.

Once we know the PLD values with respect to each feature at composition c, we can calculate the contribution of each feature for any discrete composition change. Suppose we change the composition by one step size, i.e., we move from $c \to c + \Delta c$. For this composition step, as shown in Fig. 1(b), the contribution of any feature X_i (denoted as $[\Delta Y]^{X_i}_{c\to c+\Delta c}$) towards the overall change in target Y can be calculated as the product of PLD and feature change, i.e.,

$$[\Delta Y]^{X_i}_{c\to c+\Delta c} = m_c^{X_i}\ [\Delta X_i]_{c\to c+\Delta c}, \tag{2}$$

where $[\Delta X_i]_{c\to c+\Delta c}$ is the change in value of feature X_i as composition changes from $c \to c + \Delta c$.

With this, now we can explore the exact contribution of each feature along a continuous composition pathway, as shown schematically in Fig. 1(c). Suppose we take an initial baseline composition c_0 and start increasing the concentration in steps of Δc. After N steps, the concentration is $c_0 + N\Delta c$, and the cumulative contribution of any feature X_i towards the overall change in Y prediction is calculated as

$$[\Delta Y]^{X_i}_{c_0\to c_0+N\Delta c} = \sum_{n=1}^{N} \left[m^{X_i}_{c_0+(n-1)\Delta c}\ [\Delta X_i]_{c_0+(n-1)\Delta c\to c_0+n\Delta c}\right]. \tag{3}$$

CoSMoR returns both the stepwise as well as the cumulative feature contributions along a composition pathway, as shown in Fig. 1(c). While the user is afforded the flexibility to use any or both of these, in most use cases, cumulative contributions may be the preferred choice since plotting these provides a clear and intuitive visualization of both the local changes (represented by variations in slope) as well as cumulative effect (with respect to baseline composition) for feature contributions. Plotting these cumulative feature contributions as a function of compositional variation provides two key insights. Firstly, quantitively accurate relative contributions of features can be ascertained, which can in turn be associated with the underlying physical phenomena represented by the features. Secondly, different composition pathways can be compared and the similarities or differences in the underlying physics can be understood by comparing these feature contributions. While the CoSMoR methodology to extract exact feature contributions will remain same for all ML models, the interpretation of these contributions will be problem specific. As an example, in Sec. IV, we have presented the application of CoSMoR to phase-selection problem in MPEAs.

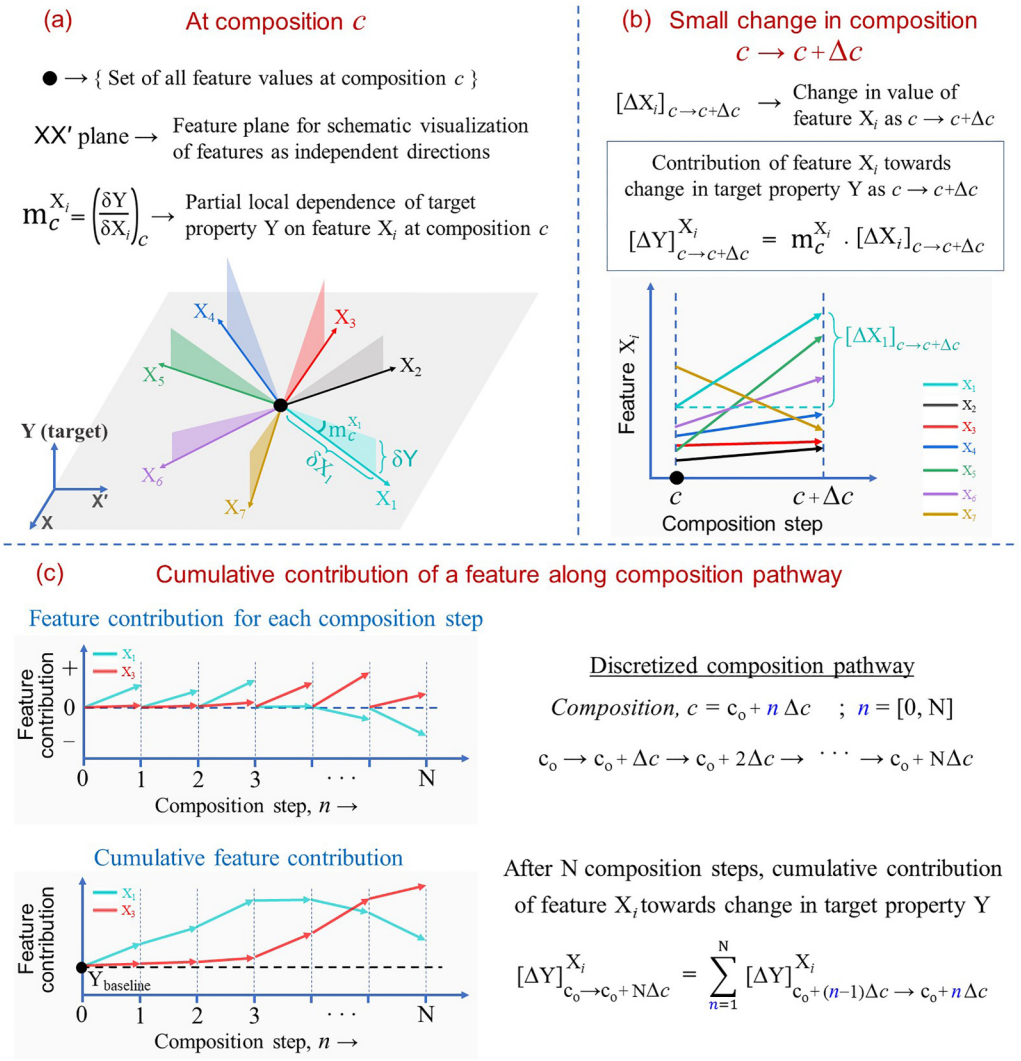

FIG. 1. Methodology used in CoSMoR to extract exact feature contributions along a compositional pathway. (a) Calculation of local partial dependence ($\delta Y/\delta X_i$) of target property Y with respect to feature X_i at any composition c. (b) Calculation of individual feature contributions towards change in phase probability for each one composition step, i.e., as composition changes from $c \to c + \Delta c$. (c) Calculation of cumulative feature contributions along a continuous composition pathway wherein the initial concentration is treated as baseline.

III. IMPLEMENTATION OF CoSMoR

While it is possible to implement CoSMoR with any programming language (as long as the concerned ML model can be loaded into it), in our work, we have developed it using PYTHON. The code, along with detailed instructions, examples, and video guides, will be made available as open source so that it can be easily integrated with diverse workflows (refer to the code availability statement preceding the Acknowledgments for details).

The workflow starts with the creation of "cosmor" class and Fig. 2 shows the flowchart depicting the implementation of CoSMoR methodology. We first go over the user inputs required for cosmor class. A composition pathway is represented as a pseudobinary system (A_xB_{1-x}), where components A and B can be either elements or stoichiometric combination of elements, e.g., Al, AlTi, Al2Ti, and AlTi2Ni are all valid component inputs. The concentration of component A (i.e., x) is used as the independent variable and thus, the user must specify that component as A whose concentration has to be varied. For example, if effect of Al addition to "Cu2NiTi" has to be studied, then component A will be Al and component B will be Cu2NiTi. The composition step-size (Δc) input is specified in atomic fraction (typically 0.01) and is used to discretize the composition space. The upper and lower bounds for the composition pathway are also required in the form of "start concentration of component A" and "end concentration of component A." These are specified in atomic fractions and must lie between [0, 1] with the additional constraint that start composition is less than end composition. Finally, feature step-size (δX_i) input is required for the calculation of PLD values of model with respect to each feature. The current code assumes the same value of δX_i for all features and is thus suitable for models where normalized features are used. This functionality will be expanded in future updates to support the use of non-normalized features.

The implementation requires six core functions, as shown in Fig. 2. Two of these, viz., "*create_features*" and

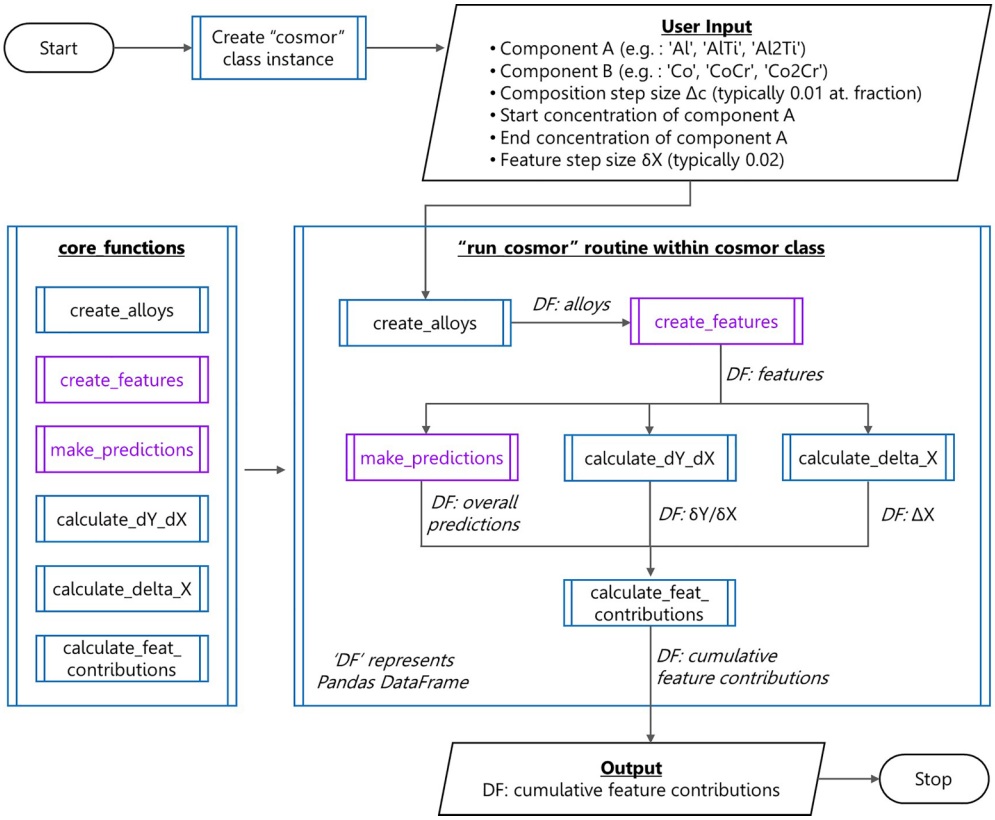

FIG. 2. Flowchart showing implementation of CoSMoR. Function blocks with purple outline and text (namely, create_features and make_predictions) are the user-defined functions whereas rest of the functions are in-built in the cosmor code.

"*make_predictions*," would be specific to each model and thus have to be defined by the user, which the user would have created during the development of ML model. The *create_features* function takes the compositions (generated by "*create_alloys*" function) as an input and returns feature values for each composition. The *make_predictions* function takes these feature values as input and generates the ML prediction for each composition. More information into the development of these functions, along with examples and unit tests, has been provided in the code repository (refer to code availability statement). The rest of the functions are model agnostic and thus do not require any user modification.

IV. PHASE SELECTION IN MULTIPRINCIPAL-ELEMENT ALLOYS

To demonstrate the application and importance of CoSMoR framework, we have applied it to probe the phase selection in MPEAs.

A. Development of ML models

The ML models used here have been reported in our previous work [35]. The training dataset for these models was compiled from the database reported by Borg *et al.* [36] that contains 426 MPEAs with experimentally observed phase information in as-cast condition at room temperature. While the dataset is small, we have shown in our previous works [7,35] that the model learning aligns well with the experimental results in a wide variety of alloy systems. Three separate artificial neural network models were trained for predicting the probability of occurrence of phases. The first model predicts only the probability of fcc phase, i.e., P(fcc), the second model predicts only the probability of bcc phase, i.e., P(bcc), and the third model predicts only the probability of IM-phase P(IM), where IM stands for intermetallic. Since these probabilities are independent of each other, they can individually vary between [0, 1] and their summation does not have to be equal to 1. While predicting the presence or absence of a phase, a probability threshold of 0.5 has been used, i.e., the model predicts the presence of a phase if the predicted $P(\text{phase}) \geqslant 0.5$. To elaborate, $P(\text{IM}) = 0.6$ means that the probability for IM phase to occur is 0.6 and we would conclude that it is present based on the threshold of 0.5. But, it gives no information on whether fcc and bcc phases are present or absent. For that, we would need to look separately at P(fcc) and P(bcc) values. These models were driven by physics-based features that have been shown to be correlated with phase stabilities in MPEAs [7] and a cross-validation accuracy of 91, 95, and 76% was obtained for bcc, fcc, and IM phase, respectively. The feature set comprised seven features:{Metallic radius asymmetry (δ_{met}), Valence electron count (VEC), Covalent radius asymmetry (δ_{cov}), Elastic modulus asymmetry (δ_{E}), Average cohesive energy (E^{coh}), Chemical enthalpy of mixing (ΔH^{chem}), Elastic enthalpy of mixing (ΔH^{el})}. Since the focus here is on the implementation of CoSMoR to gain physical insights, more details related to the training, validation, and performance of these models can be obtained from previous work [35].

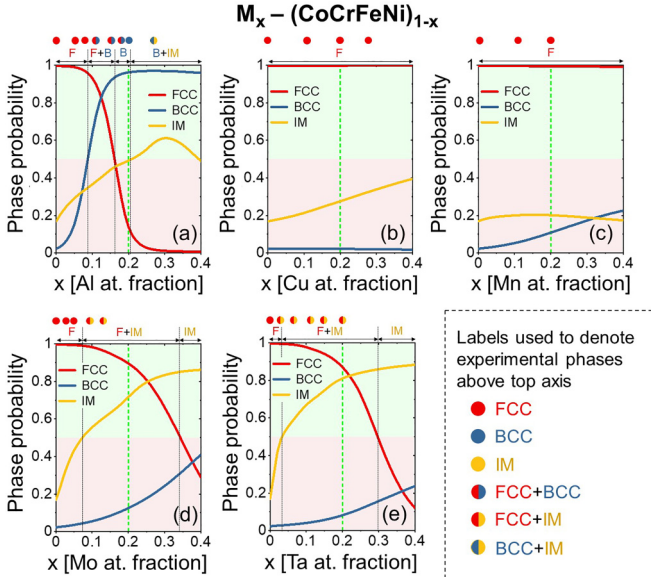

FIG. 3. Probability of occurrence of fcc, bcc and IM phases as predicted by machine learning models in M_x-(CoCrFeNi)$_{1-x}$ alloy systems. Phase probabilities as a function of (a) Al, (b) Cu, (c) Mn, (d) Mo, and (e) Ta addition in CoCrFeNi alloy. A phase probability (P) threshold of 0.5 has been used to indicate the presence of a particular phase, and correspondingly each plot has been divided into green and red regions representing $P \geqslant 0.5$ (i.e., phase presence) and $P \leqslant 0.5$ (i.e., phase absence), respectively. Predicted phase-transition boundaries have also been highlighted above the top axis of each plot, along with the experimentally observed phases at some discrete compositions. The green dotted line ($x = 0.2$) in each plot represents the equiatomic quinary composition.

B. Phase variations in M_x-(CoCrFeNi)$_{1-x}$

Since CoSMoR brings forth the decision-making process of the ML model along continuous composition pathways, it is pertinent to ascertain whether the ML model is even capable of capturing the continuous variations in the first place. For this purpose, we probed the effect of various element additions (Al, Cu, Mn, Mo, and Ta) on phase evolution in CoCrFeNi. Equiatomic quaternary CoCrFeNi is a medium-entropy reduced Cantor alloy that exhibits fcc solid-solution phase in as-cast condition [37,38]. We selected CoCrFeNi as the base alloy since the effect of different alloying elements has been studied experimentally for this system [38–51]; thus, these compositions can be used as validation checkpoint for the phase mappings predicted through ML model. Figure 3 shows the predicted occurrence probabilities of fcc, bcc, and IM phases as a function of different alloying additions in CoCrFeNi quaternary along with the experimentally observed phases at some discrete compositions. With the addition of Cu and Mn, only fcc phase is predicted throughout the composition range of 0–0.4 at. fraction, as seen in Figs. 3(b) and 3(c). This matches the experimental observations for Cu [44,47,50] and Mn [39,46] variations.

Addition of Al to CoCrFeNi increases the occurrence probability of bcc phase, as seen in Fig. 3(a), and the model predicts formation of stable bcc phase beyond 9 at.% Al; this aligns closely with the experimental observations made by Chou *et al.* [51] wherein the alloy structure transitioned from single-phase fcc to dual-phase (fcc+bcc) as Al concentration was increased from 8.57 to 11.1 at.%. The model predicts a decrease in occurrence probability of fcc phase as Al increases and the structure transitions completely to bcc phase above 16 at.% Al; this complete transition from fcc → bcc phase has also been observed experimentally by Cieslak *et al.* [45] and Chou *et al.* [51] above 15.7 and 20 at.% Al respectively. The model also predicts existence of IM phases above 20 at.% Al, which aligns with the formation of $B2$ phase above 20 at.% Al, as observed experimentally [45]. Thus, the effect of continuous Al variation has been captured accurately by the ML model as the learned transition boundaries align very closely with experimental observations.

Addition of Mo or Ta to (CoCrFeNi) induces a strong IM formation tendency, as seen in Figs. 3(d) and 3(e), and the model predicts a transition from fcc phase to a dual-phase (fcc+IM) structure above 7 at.% Mo and 3 at.% Ta. This aligns with the experimental studies for varying Mo [42] and Ta [41,43] concentration, which showed the formation of IM phase at 9 at.% Mo and 2.5 at.% Ta. Also, the predicted occurrence probability of bcc phase remains comfortably below the 0.5 threshold and thus no bcc phase formation is predicted with Mo or Ta addition in (CoCrFeNi).

C. Decoding the decision-making process in M_x-(CoCrFeNi)$_{1-x}$

To understand the decision-making process used by the ML model for predicting phase probabilities, we implemented CoSMoR to calculate exact contribution of individual features towards fcc, bcc, and IM phase probabilities in M_x-(CoCrFeNi)$_{1-x}$ alloy systems probed in Sec. IV B. The feature contributions in these alloy systems, along with normalized feature values, have been plotted in Figs. 4 and 5, wherein $x = 0$ has been used as the baseline composition. A quick observation of these trends shows that (a) the feature contributions vary nonlinearly, sometimes even nonmonotonically, with respect to feature values; (b) same features contribute differently towards occurrence probability of different phases, viz., fcc, bcc, and IM; and (c) the relative importance of features changes as we move from one system to another. These observations indicate that the decision-making process of the ML model is not purely statistical, and that it updates dynamically as the alloy system changes, thereby highlighting the cognizance of the model to underlying physics that drives the phase-selection process in MPEAs. Here, we discuss the findings that stand out on closer inspection.

VEC and δ_{met} contributions dominate the occurrence probability of fcc and bcc phases, as seen with respect to addition of Al, Mo, or Ta to (CoCrFeNi) in Figs. 4(a) and 4(b), 5(a) and 5(b), and 5(e) and 5(f), respectively. As Al, Mo, or Ta concentration increases, VEC decreases whereas δ_{met} increases, but both VEC and δ_{met} contribute strongly towards a decrease in fcc and increase in bcc phase probability. Also, the magnitude of feature contributions changes sharply in the fcc → bcc transition domain (9–16 at.% Al) observed from experimental studies [45,51]. This aligns with the previous observations wherein VEC has been shown to be strongly associated with the stability of fcc and bcc phases [35,52–54]. Contrary to these systems, the addition of Cu or Mn to (CoCrFeNi)

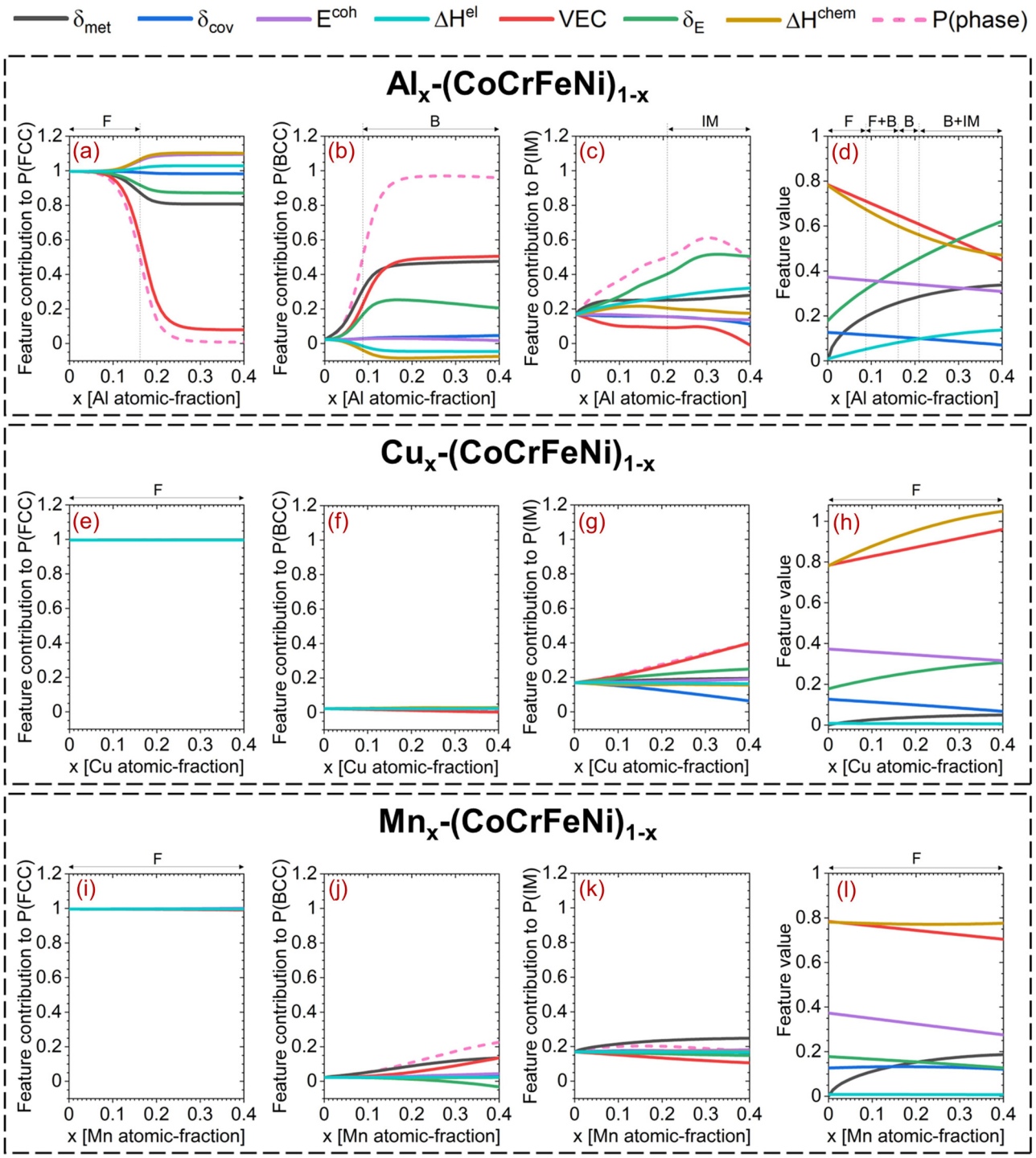

FIG. 4. Interpreting the decision-making process of phase-selection ML models using CoSMoR. Cumulative contribution of each feature towards the overall phase probability P(phase) in (a)–(c) Al_x-$(CoCrFeNi)_{1-x}$, (e)–(g) Cu_x-$(CoCrFeNi)_{1-x}$ and (i)–(k) Mn_x-$(CoCrFeNi)_{1-x}$ alloy systems. Normalized feature values as a function of (d) Al, (h) Cu, and (l) Mn concentration. feature contributions here are cumulative contributions along the composition pathway with respect to baseline composition of $x = 0$.

does not induce any significant change in the occurrence probability of fcc or bcc phase, as seen in Figs. 4(e) and 4(f) and Figs. 4(i) and 4(j), respectively. Even though δ_{met} increases with addition of Mn (Fig. 4(l)), its contribution towards fcc and bcc phase probability is negligible and marginal, respectively. This aligns with previous observations [35] that the VEC acts as a classifier wherein the importance of not only VEC, but also that of the other features, towards occurrence of fcc/bcc phases is dictated by the VEC value. As seen with Al addition, the contributions of all features towards P(fcc) and P(bcc) see sudden changes beyond a threshold VEC value. Thus, it appears that the model has successfully learned these physics-based relationships to establish the stability of fcc and bcc phases in MPEAs.

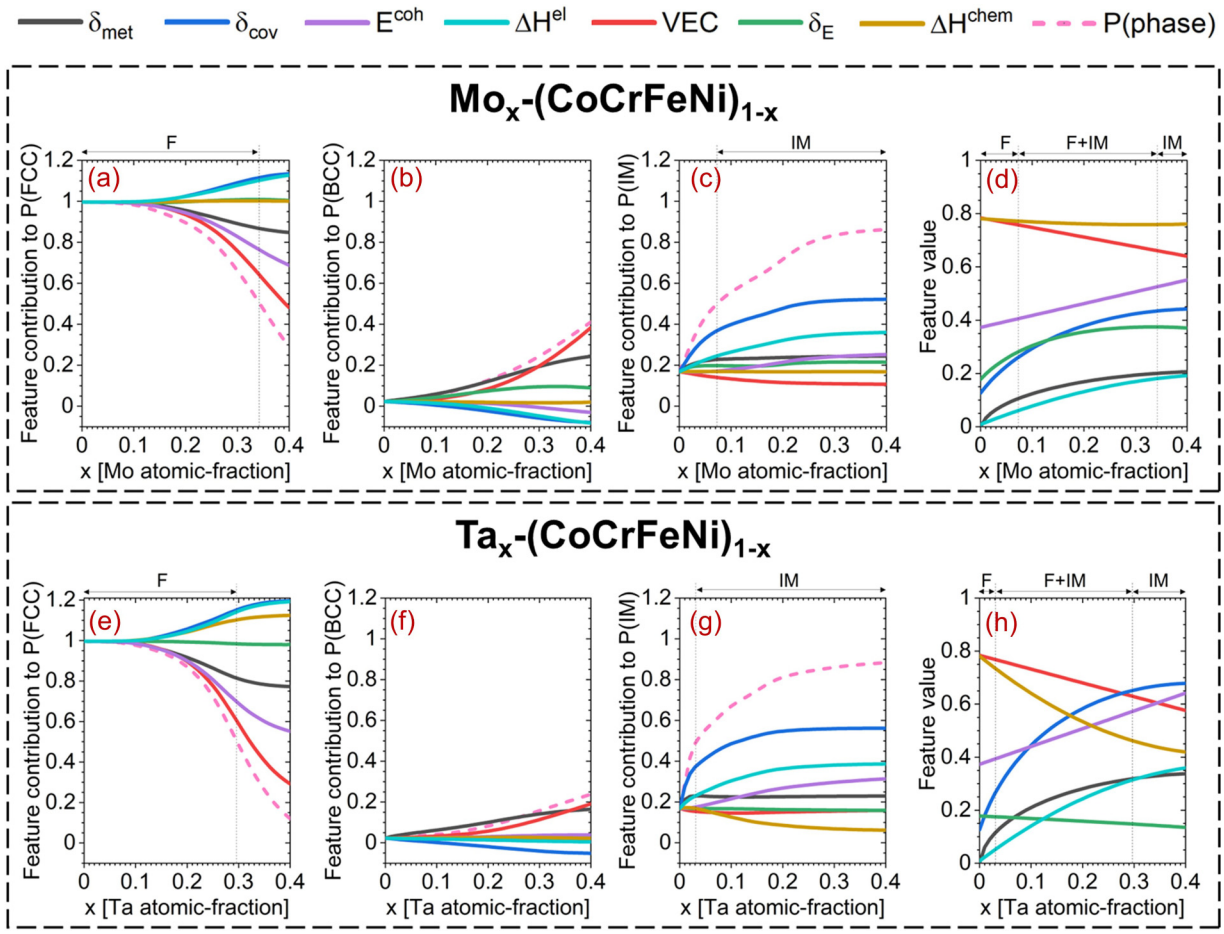

FIG. 5. Interpreting the decision-making process of phase-selection ML models using CoSMoR. Cumulative contribution of each feature towards the overall phase probability P(phase) in (a)–(c) Mo_x-$(CoCrFeNi)_{1-x}$ and (e)–(g) Ta_x-$(CoCrFeNi)_{1-x}$ alloy systems. Normalized feature values as a function of (d) Mo and (h) Ta concentration. Feature contributions here are cumulative contributions along the composition pathway with respect to baseline composition of $x = 0$.

While fcc and bcc phase occurrence is dominated by VEC and δ_{met}, the IM phase probability in M_x-$(CoCrFeNi)_{1-x}[M = \{Al, Mo, Ta\}]$ is dominated by three feature contributions: δ_{cov}, δ_{E}, and ΔH^{el}. With the addition of Al, both δ_{E} and ΔH^{el} increase [Fig. 4(d)] and consequently drive the formation of IM phase to a large extent [Fig. 4(c)], but the contribution of δ_{E} to IM formation is considerably more due to a much steeper increase in δ_{E} feature value. On the other hand, with addition of Mo and Ta, δ_{cov} and ΔH^{el} contribute significantly toward the IM phase formation, whereas the contribution of δ_{E} is almost negligible, as seen in Figs. 5(c) and 5(g). This is because δ_{cov} increases sharply with Mo and Ta addition whereas δ_{E} saturates at low values, as seen in Figs. 5(d) and 4(h). To read further into these observations, we need to discuss how these features (δ_{cov}, δ_{E}, and ΔH^{el}) could be correlated to the formation of intermetallics. The bond formation in metals and alloys always has some covalent character [55] that is expected to be affected by δ_{cov}, which is a measure of asymmetry in bond lengths when constituent elements participate in covalent bond formation. The covalent nature is especially dominant in the intermetallics wherein a large δ_{cov} can dictate preferential formation of certain atomic pairs and coordination symmetries corresponding to favorable atomic-size ratios. Young's modulus (E) is correlated to the strength of interatomic bonds as it can be estimated from the potential energy vs separation curve. Since multiple elemental interactions are possible in MPEAs, δ_{E} is an indirect measure of the asymmetry in bond strengths of different atomic-pairs. A high δ_{E} would indicate the presence of certain atomic pairs with considerably higher or lower bond strengths that may promote ordering or clustering tendencies, and thus, dictate formation of intermetallics. The elastic enthalpy of mixing ΔH^{el}, calculated using the classical elasticity method by Eshelby and Friedel [56–58], is a measure of the elastic energy generated due to the internal strains induced by size asymmetry between constituent elements. Since the calculation of ΔH^{el} considers both local distortion (through volume corrections) and bonding characteristics (indirectly through shear and bulk modulus), a large ΔH^{el} represents instability induced in the solid-solution phase due to excessive strain energy and is expected to result in the formation of ordered structures that can relieve some of this strain energy. Thus, all three features, viz., δ_{cov}, δ_{E}, and ΔH^{el}, are closely correlated to the IM phase formation in MPEAs, and the fact that ML model gives significant and selective weightage to these features strongly indicates that it has successfully learned the underlying physics.

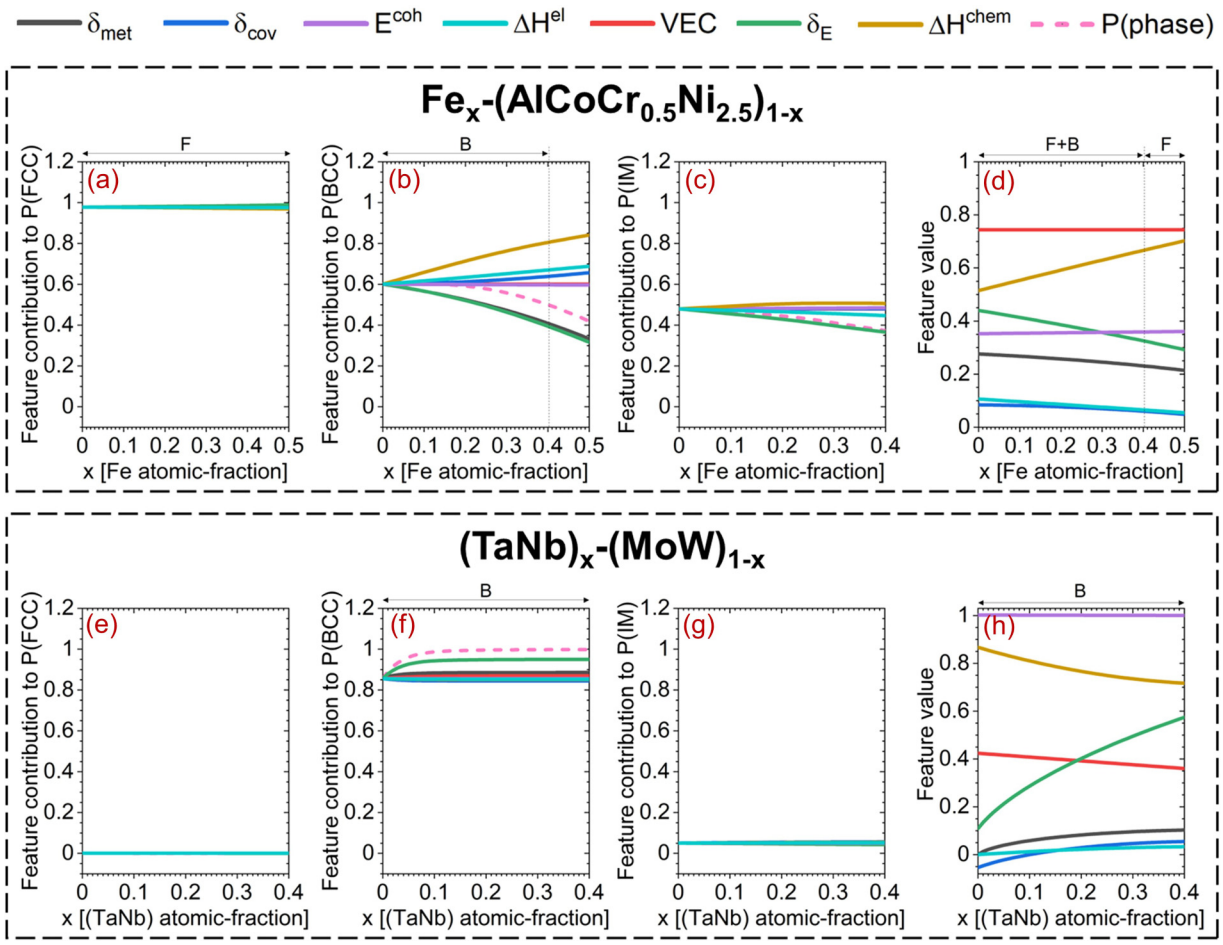

FIG. 6. Interpreting the decision-making process of phase-selection ML models using CoSMoR. Cumulative contribution of each feature towards the overall phase probability P(phase) in (a)–(c) Fe_x-$(AlCoCr_{0.5}Ni_{2.5})_{1-x}$ and (e)–(g) $(TaNb)_x$-$(MoW)_{1-x}$ alloy systems. Normalized feature values as a function of (d) Fe and (g) (TaNb) concentration. feature contributions here are cumulative contributions along the composition pathway with respect to baseline composition of $x = 0$.

D. Exploring Fe_x-$(AlCoCr_{0.5}Ni_{2.5})_{1-x}$ and $(TaNb)_x$-$(MoW)_{1-x}$ alloy systems

The implementation of CoSMoR in M_x-$(CoCrFeNi)_{1-x}$ alloy systems shows a strong dependence of fcc and bcc phase occurrence on VEC. This poses an interesting question as to whether the ML model can capture phase stabilities in systems where VEC remains constant. To address this, we looked at Fe_x-$(AlCoCr_{0.5}Ni_{2.5})_{1-x}$ alloy system that maintains a constant VEC of 8 and has been studied experimentally by Liu *et al.* [59]. One would expect this system to have fcc phase throughout the compositional range, but Liu et al. observed a transition from (fcc+bcc) dual-phase structure at $x = 0$ to an almost single-phase fcc structure at $x = 0.47$. The ML model captures this transition almost perfectly, as seen in Figs. 6(a) and 6(b), and the interpretation framework reveals that the extinction of bcc phase accompanying Fe addition is driven primarily by the decrease in δ_{met} and δ_E. δ_{met} is a measure of metallic radius asymmetry between the component elements, and increase in δ_{met} contributes to an increase in bcc and decrease in fcc phase probability, as seen in Figs. 4(a) and 4(b), 5(a), 5(b), 5(e), and 5(f), and Fig. 6(b). This is expected since bcc phase, due to its more open structure, can accommodate the larger size asymmetry more easily as compared to the close-packed fcc phase.

CoSMoR was further used to explore another interesting alloy system that contradicts some of the earlier observations related to IM formation. In M_x-$(CoCrFeNi)_{1-x}$ alloy systems, δ_{cov}, δ_E, and ΔH^{el} features contribute strongly towards the occurrence of IM phases. Especially with Al addition, the increase in δ_E contributed significantly to the formation of IM phases at high Al concentration. But, on the contrary, $(TaNb)_x$-$(MoW)_{1-x}$ system shows a very steep increase in δ_E value as (TaNb) concentration increases [Fig. 6(h)], but experimentally no IM formation is observed in this system [60]. This raises an interesting question as to whether the ML model can predict this behavior and how this decision would be made based on all feature values. The ML model does not predict any IM formation in this system, as seen in Fig. 6(g), which aligns with the experimental observations [60] wherein TaNbMoW forms a simple single-phase bcc structure. This can be attributed to the fact that all the other drivers for IM formation are virtually nonexistent in this system since (a) there are no significant lattice strains as δ_{met} and ΔH^{el} are almost zero, (b) no strong chemical interactions

as $-6.38 < \Delta H^{\text{chem}} \leftarrow -0.22\,\text{kJ/mol}$, and (c) δ_{cov} is almost zero. This indicates that a high δ_{E} alone is not sufficient for IM formation, and that the model understands these nuances in the feature variations that are associated with occurrence of IM phases.

V. USE-CASE SCENARIOS FOR CoSMoR

CoSMoR is a generalized framework that can be applied to any ML model (irrespective of the type or complexity of the models and features used) provided that the model has been built using features that are a direct function of material composition. Here we discuss the scenarios wherein CoSMoR can be used to evaluate, interpret, and improve the compositional ML models that have been built for predicting material properties.

A. Evaluating the nature of fit

The development and optimization of ML models is driven by the maximization of statistical performance that is quantified using a variety of metrics such as mean absolute error, R^2 value, percentage error, precision, recall, etc. But, while the statistical nature of fit is routinely evaluated, the physical nature of the fit (which implies the consistency between the ML model decision-making process and the known physical rules) is often ignored, resulting in a big question mark as to how well the model is expected to perform when extrapolated to novel compositions. It is an even bigger problem when using ML models (such as neural networks) that have multiple minima corresponding to similar statistical performance but different learning states, since now one can end up with different models with similar statistical performance. CoSMoR addresses these concerns by enabling the evaluation of physical nature of fit for compositional ML models through the comparison of feature contributions with well-established physical rules. For example, VEC is known to have a strong impact on fcc and bcc phase stability in MPEAs and thus when we probe the model decision-making process in M_x-$(\text{CoCrFeNi})_{1-x}$ alloys with CoSMoR, the fact that VEC contributions closely follow the P(fcc) and P(bcc) variations strongly indicates that the model has learned this rule. On the other hand, if it had so happened that VEC contributions did not align with the P(fcc) and P(bcc) variations, we could have safely rejected the model even if it showed good statistical accuracy. Thus, users can validate the physical consistency of their compositional ML models by using CoSMoR.

B. Revealing the drivers of change

The black-box treatment of ML models obscures their decision-making process, leaving open questions as to how exactly the model outcomes are computed. Implementation of CoSMoR provides the exact feature contributions along specific composition pathways and thus reveals material-specific insights into what drives the changes in target property as predicted by the ML model. This understanding can be both at feature level and physical level, as elaborated here.

First, CoSMoR identifies features that drive the changes as a function of composition which can lead to new fundamental insights and design principles. For example, since the advent of data-driven models for MPEAs, the intermetallic formation has been frequently assumed to be driven by large metallic radius asymmetry (δ_{met}). But, the implementation of CoSMoR in this work shows that it is actually the covalent radius asymmetry that is considerably more critical for intermetallic formation, and that δ_{met} is more relevant for fcc and bcc phase stability rather than intermetallic formation. Similarly, even though δ_{met} and elastic enthalpy of mixing (ΔH^{el}) features are highly correlated, the intermetallic formation is predominantly driven by ΔH^{el}. Thus, CoSMoR can provide insights into the relevance and significance of features used in the ML model.

Secondly, when combined with experimental observations and *ab initio* calculations, CoSMoR can lead to insights into the physical origins of the target property predicted by the ML model. Suppose we have a compositional ML model that uses a set of input material descriptors (such as VEC, δ_{met}, δ_{cov}, ΔH^{el}, etc.) to predict the strength of an alloy. Using CoSMoR, we can identify the origin of strength at feature level while the density-functional theory calculations and experimental results can assist in identification of baselines as to what these origins at feature level may represent at a physical level within the material. For example, one of the expected observations in this case would be that whenever fcc↔bcc transitions occur, the contribution of VEC towards the resultant changes in strength becomes significant. Such baselines, once established, will allow exploring the novel compositional spaces using CoSMoR to predict not only the variations in predicted property, but also the underlying physical phenomena that may be responsible for these variations. In our previous work [12], we have shown how this methodology can be used to extract physical origins of hardness in MPEAs.

The model-agnostic implementation offered by CoSMoR allows its use for both continuous as well as discrete decision-based models such as decision trees and random forest (RF). To illustrate this, we developed RF models for the same problem presented in this work, i.e., presence/absence of fcc, bcc, and IM phases in MPEAs and implemented CoSMoR to probe the decision-making process of these RF models along some of the compositional spaces. The performance of these models is inferior to the neural network models and they were developed only to illustrate the applicability of CoSMoR to such models. The results in Supplemental Material [61] show that CoSMoR successfully brings out the contribution of different features towards the phase occurrence for these RF models also. The instability of PLDs (which may occur in such discrete models) at phase-transition boundaries does not appear to be a problem for fcc and bcc models. But, for the IM model, we do see instability in both the model output as well as the feature contributions. In fact, the instability in feature contributions could largely be due to the instability in the model output itself because the IM RF model was the least accurate model with $F1$ score of 0.74 as compared to fcc and bcc models with $F1$ score of 0.92 and 0.94, respectively. Thus, while the breakdown of continuous models using CoSMoR will always be more accurate as compared to that for decision-based models, we believe that certain best practices can improve the performance for RF models: (a) Reducing the complexity of the RF by controlling the depth and splitting criteria, (b) preventing overfit of the model,

and (c) improving generality of the model by using bootstrapping or other techniques.

VI. CONCLUSION

We have presented here the model-agnostic **Co**mpositional **S**timulus and **Mo**del **R**esponse (**CoSMoR**) framework, which is an interpretation framework capable of decoding the decision-making process of machine learning models trained for material properties along continuous compositional pathways. A compositional pathway is defined as a pseudobinary A_xB_{1-x} where components A and B can be either elements (e.g., Al or Ti) or any stoichiometric combination of elements (e.g., CrFe, Cr_2Fe, or Cr_2FeNi_2). As composition changes, the ML model predictions also change, but the variations in model predictions cannot be explained if the ML model is treated as a black box. CoSMoR addresses this by calculating the exact contribution of each feature along any compositional pathway based on the partial-local dependence of ML model (with respect to each feature) and the sensitivity of that feature (with respect to the composition). CoSMoR offers three unique advantages. First, it adopts a model agnostic approach and thus can be applied to any compositional ML model, irrespective of the algorithm used. Second, the causality of model understanding is shifted to variations in alloy composition rather than some arbitrary feature variations. This enables integration of the understanding derived from CoSMoR with *ab initio* calculations and experimental results to correlate the feature contributions with physical phenomena. Third, the feature contributions extracted by CoSMoR are quantitatively exact with respect to the model decision-making process. This enables direct and reliable comparison between multiple composition pathways to assess similarities and differences between how ML model treats these different alloy systems. We have showcased the importance of CoSMoR through implementation on phase-selection problem in multiprincipal element alloys. Individual feature contributions towards predicted phase probabilities in a variety of alloy systems, viz., M_x-$(CoCrFeNi)_{1-x}[M = \{Al, Cu, Mn, Mo, Ta\}]$, Fe_x-$(AlCoCr_{0.5}Ni_{2.5})_{1-x}$, and $(TaNb)_x$-$(MoW)_{1-x}$ were obtained using CoSMoR. The interpretation results show that the ML model has learned the underlying physics associated with phase stability and point toward the features that are essential for predicting relative stability of fcc/bcc phases and formation of intermetallic phase. Thus, CoSMoR enables a systematic and insightful exploration of compositional spaces using ML models to not only validate the physical nature of fit but to also extract material-specific insights as to how the target property manifests in a material.

The CoSMoR code, along with the documentation, examples and video tutorials, is available at IDEAsLab-Materials-Informatics organization page on GitHub [62].

ACKNOWLEDGMENTS

D.B. acknowledges the support from Prime Minister's Research Fellows scheme run by Ministry of Education, Government of India. P.K.R. acknowledges the support from Science and Engineering Research Board, Department of Science and Technology, Ministry of Science and Technology, Government of India (Grant No. CRG/2021/006974).

[1] K. T. Butler, D. W. Davies, H. Cartwright, O. Isayev, and A. Walsh, Machine learning for molecular and materials science, Nature (London) **559**, 547 (2018).

[2] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques, Recent advances and applications of machine learning in solid-state materials science, npj Comput. Mater. **5**, 83 (2019).

[3] R. Batra, L. Song, and R. Ramprasad, Emerging materials intelligence ecosystems propelled by machine learning, Nat. Rev. Mater. **6**, 655 (2021).

[4] D. Beniwal, Jhalak, and P. K. Ray, Data-Driven Phase Selection, Property prediction and force-field development in multi-principal element alloys, in *Forcefields for Atomistic-Scale Simulations: Materials and Applications*, edited by A. Verma, S. Mavinkere Rangappa, S. Ogata, and S. Siengchin (Springer Nature, Singapore, 2022), pp. 315–347.

[5] K. Choudhary, B. DeCost, C. Chen, A. Jain, F. Tavazza, R. Cohn, C. W. Park, A. Choudhary, A. Agrawal, S. J. L. Billinge *et al.*, Recent advances and applications of deep learning methods in materials science, npj Comput. Mater. **8**, 59 (2022).

[6] G. L. W. Hart, T. Mueller, C. Toher, and S. Curtarolo, Machine learning for alloys, Nat. Rev. Mater. **6**, 730 (2021).

[7] D. Beniwal and P. K. Ray, Learning phase selection and assemblages in high-entropy alloys through a stochastic ensemble-averaging model, Comput. Mater. Sci. **197**, 110647 (2021).

[8] K. Lee, M. V. Ayyasamy, P. Delsa, T. Q. Hartnett, and P. V. Balachandran, Phase classification of multi-principal element alloys via interpretable machine learning, npj Comput. Mater. **8**, 25 (2022).

[9] A. Roy, T. Babuska, B. Krick, and G. Balasubramanian, Machine learned feature identification for predicting phase and Young's modulus of low-, medium- and high-entropy alloys, Scr. Mater. **185**, 152 (2020).

[10] W. Huang, P. Martin, and H. L. Zhuang, Machine-learning phase prediction of high-entropy alloys, Acta Mater. **169**, 225 (2019).

[11] V. Shivam, D. Beniwal, Y. Shadangi, P. Singh, V. S. Hariharan, G. Phanikumar, D. D. Johnson, P. K. Ray, and N. K. Mukhopadhyay, Effect of Zn addition on phase selection in AlCrFeCoNiZn high-entropy alloy, SSRN Electron. J. 24 (2022).

[12] D. Beniwal, P. Singh, S. Gupta, M. J. Kramer, D. D. Johnson, and P. K. Ray, Distilling physical origins of hardness in multi-principal element alloys directly from ensemble neural network models, npj Comput. Mater. **8**, 153 (2022).

[13] J. M. Rickman, H. M. Chan, M. P. Harmer, J. A. Smeltzer, C. J. Marvel, A. Roy, and G. Balasubramanian, Materials informatics for the screening of multi-principal elements and high-entropy alloys, Nat. Commun. **10**, 2618 (2019).

[14] C. Yang, C. Ren, Y. Jia, G. Wang, M. Li, and W. Lu, A machine learning-based alloy design system to facilitate the rational design of high entropy alloys with enhanced hardness, Acta Mater. **222**, 117431 (2022).

[15] I. Roy, P. K. Ray, and G. Balasubramanian, Modeling oxidation of AlCoCrFeNi high-entropy alloy using stochastic cellular automata, Entropy **24**, 1263 (2022).

[16] C. D. Taylor and B. M. Tossey, High temperature oxidation of corrosion resistant alloys from machine learning, npj Mater. Degrad. **5**, 38 (2021).

[17] N. Birbilis, S. Choudhary, J. R. Scully, and M. L. Taheri, A perspective on corrosion of multi-principal element alloys, npj Mater. Degrad. **5**, 14 (2021).

[18] A. Roy, M. F. N. Taufique, H. Khakurel, R. Devanathan, D. D. Johnson, and G. Balasubramanian, Machine-learning-guided descriptor selection for predicting corrosion resistance in multi-principal element alloys, npj Mater. Degrad. **6**, 9 (2022).

[19] S. Kr. Bhattacharya, R. Sahara, and T. Narushima, Predicting the parabolic rate constants of high-temperature oxidation of Ti alloys using machine learning, Oxid. Met. **94**, 205 (2020).

[20] C. J. Bartel, S. L. Millican, A. M. Deml, J. R. Rumptz, W. Tumas, A. W. Weimer, S. Lany, V. Stevanović, C. B. Musgrave, and A. M. Holder, Physical descriptor for the gibbs energy of inorganic crystalline solids and temperature-dependent materials chemistry, Nat. Commun. **9**, 4168 (2018).

[21] C. J. Bartel, A. Trewartha, Q. Wang, A. Dunn, A. Jain, and G. Ceder, A critical examination of compound stability predictions from machine-learned formation energies, npj Comput. Mater. **6**, 97 (2020).

[22] Y. Liu, T. Zhao, W. Ju, and S. Shi, Materials discovery and design using machine learning, J. Materiomics **3**, 159 (2017).

[23] Y. Yan, D. Lu, and K. Wang, Accelerated discovery of single-phase refractory high entropy alloys assisted by machine learning, Comput. Mater. Sci. **199**, 110723 (2021).

[24] Y. Iwasaki, I. Takeuchi, V. Stanev, A. G. Kusne, M. Ishida, A. Kirihara, K. Ihara, R. Sawada, K. Terashima, H. Someya *et al.*, Machine-learning guided discovery of a new thermoelectric material, Sci. Rep. **9**, 2751 (2019).

[25] K. Kaufmann, D. Maryanovsky, W. M. Mellor, C. Zhu, A. S. Rosengarten, T. J. Harrington, C. Oses, C. Toher, S. Curtarolo, and K. S. Vecchio, Discovery of high-entropy ceramics via machine learning, npj Comput. Mater. **6**, 42 (2020).

[26] S. Lu, Q. Zhou, L. Ma, Y. Guo, and J. Wang, Rapid discovery of ferroelectric photovoltaic perovskites and material descriptors via machine learning, Small Methods **3**, 1900360 (2019).

[27] D. Khatamsaz, B. Vela, P. Singh, D. D. Johnson, D. Allaire, and R. Arróyave, Bayesian optimization with active learning of design constraints using an entropy-based approach, npj Comput. Mater. **9**, 49 (2023).

[28] B. Vela, C. Acemi, P. Singh, T. Kirk, W. Trehern, E. Norris, D. D. Johnson, I. Karaman, and R. Arróyave, High-throughput exploration of the WMoVTaNbAl refractory multi-principal-element alloys under multiple-property constraints, Acta Mater. **248**, 118784 (2023).

[29] B. Kailkhura, B. Gallagher, S. Kim, A. Hiszpanski, and T. Y.-J. Han, Reliable and explainable machine-learning methods for accelerated material discovery, npj Comput. Mater. **5**, 108 (2019).

[30] K. Lee, M. V. Ayyasamy, Y. Ji, and P. V. Balachandran, A comparison of explainable artificial intelligence methods in the phase classification of multi-principal element alloys, Sci. Rep. **12**, 11591 (2022).

[31] F. Oviedo, J. L. Ferres, T. Buonassisi, and K. T. Butler, Interpretable and explainable machine learning for materials science and chemistry, Acc. Mater. Res. **3**, 597 (2022).

[32] G. Pilania, Machine learning in materials science: From explainable predictions to autonomous design, Comput. Mater. Sci. **193**, 110360 (2021).

[33] X. Zhong, B. Gallagher, S. Liu, B. Kailkhura, A. Hiszpanski, and T. Y.-J. Han, Explainable machine learning in materials science, npj Comput. Mater. **8**, 204 (2022).

[34] Z. C. Lipton, The Mythos of model interpretability: In machine learning, the concept of interpretability is both important and slippery,, Queue **16**, 31 (2018).

[35] D. Beniwal and P. K. Ray, FCC vs. BCC phase selection in high-entropy alloys via simplified and interpretable reduction of machine learning models, Materialia **26**, 101632 (2022).

[36] C. K. H. Borg, C. Frey, J. Moh, T. M. Pollock, S. Gorsse, D. B. Miracle, O. N. Senkov, B. Meredig, and J. E. Saal, Expanded dataset of mechanical properties and observed phases of multi-principal element alloys, Sci. Data **7**, 430 (2020).

[37] B. Cantor, Multicomponent high-entropy cantor alloys, Prog. Mater. Sci. **120**, 100754 (2021).

[38] S. Guo, C. Ng, Z. Wang, and C. T. Liu, Solid solutioning in equiatomic alloys: Limit set by topological instability, J. Alloys Compd. **583**, 410 (2014).

[39] B. Cantor, I. T. H. Chang, P. Knight, and A. J. B. Vincent, Microstructural development in equiatomic multicomponent alloys, Mater. Sci. Eng. A **375–377**, 213 (2004).

[40] L. Tang, K. Yan, B. Cai, Y. Wang, B. Liu, S. Kabra, M. M. Attallah, and Y. Liu, Deformation mechanisms of $FeCoCrNiMo_{0.2}$ high entropy alloy at 77 and 15 K, Scr. Mater. **178**, 166 (2020).

[41] A.-C. Fan, J.-H. Li, and M.-H. Tsai, On the phase constituents of three CoCrFeNiX (X = V, Nb, Ta) high-entropy alloys after prolonged annealing, J. Alloys Compd. **823**, 153524 (2020).

[42] C. Dai, T. Zhao, C. Du, Z. Liu, and D. Zhang, Effect of molybdenum content on the microstructure and corrosion behavior of FeCoCrNiMox high-entropy alloys, J. Mater. Sci. Technol. **46**, 64 (2020).

[43] C. Ai, F. He, M. Guo, J. Zhou, Z. Wang, Z. Yuan, Y. Guo, Y. Liu, and L. Liu, Alloy design, micromechanical and macromechanical properties of CoCrFeNiTax eutectic high entropy alloys, J. Alloys Compd. **735**, 2653 (2018).

[44] Y. Cai, S. M. Manladan, and Z. Luo, Tribological behaviour of the double FeCoNiCrCux middle-entropy alloy coatings, Surf. Eng. **35**, 14 (2019).

[45] J. Cieslak, J. Tobola, K. Berent, and M. Marciszko, Phase composition of AlxFeNiCrCo high entropy alloys prepared by sintering and Arc-melting methods, J. Alloys Compd. **740**, 264 (2018).

[46] K. A. Christofidou, E. J. Pickering, P. Orsatti, P. M. Mignanelli, T. J. A. Slater, H. J. Stone, and N. G. Jones, On the influence of Mn on the phase stability of the CrMnxFeCoNi high entropy alloys, Intermetallics **92**, 84 (2018).

[47] U. Dahlborg, J. Cornide, M. Calvo-Dahlborg, T. C. Hansen, Z. Leong, L. Asensio Dominguez, S. Chambreland, A. Cunliffe, R. Goodall, and I. Todd, Crystalline structures of some high entropy alloys obtained by neutron and x-ray diffraction, Acta Phys. Pol. A **128**, 552 (2015).

[48] F. Tian, L. Delczeg, N. Chen, L. K. Varga, J. Shen, and L. Vitos, Structural stability of $NiCoFeCrAl_x$ high-entropy alloy from ab initio theory, Phys. Rev. B **88**, 085128 (2013).

[49] C. Li, J. C. Li, M. Zhao, and Q. Jiang, Effect of alloying elements on microstructure and properties of multiprincipal elements high-entropy alloys, J. Alloys Compd. **475**, 752 (2009).

[50] Y.-J. Hsu, W.-C. Chiang, and J.-K. Wu, Corrosion behavior of FeCoNiCrCux high-entropy alloys in 3.5% sodium chloride solution, Mater. Chem. Phys. **92**, 112 (2005).

[51] H.-P. Chou, Y.-S. Chang, S.-K. Chen, and J.-W. Yeh, Microstructure, thermophysical and electrical properties in AlxCoCrFeNi ($0 \leqslant x \leqslant 2$) high-entropy alloys, Mater. Sci. Eng. B **163**, 184 (2009).

[52] P. Singh, A. Sharma, A. V. Smirnov, M. S. Diallo, P. K. Ray, G. Balasubramanian, and D. D. Johnson, Design of high-strength refractory complex solid-solution alloys, npj Comput. Mater. **4**, 16 (2018).

[53] S. Guo, C. Ng, J. Lu, and C. T. Liu, Effect of valence electron concentration on stability of fcc or bcc phase in high entropy alloys, J. Appl. Phys. **109**, 103505 (2011).

[54] Y. Zhang, Y. J. Zhou, J. P. Lin, G. L. Chen, and P. K. Liaw, Solid-solution phase formation rules for multi-component alloys, Adv. Eng. Mater. **10**, 534 (2008).

[55] F. E. Wang, *Bonding Theory for Metals and Alloys* (Elsevier, Amsterdam, 2005), Vol. 1.

[56] J. D. Eshelby, The continuum theory of lattice defects, in *Solid State Physics*, edited by F. Seitz and D. Turnbull (Academic Press, London, 1956), Vol. 3, pp. 79–144.

[57] J. Friedel, Electronic structure of primary solid solutions in metals, Adv. Phys. **3**, 446 (1954).

[58] R. F. Zhang, S. H. Zhang, Z. J. He, J. Jing, and S. H. Sheng, Miedema calculator: A thermodynamic platform for predicting formation enthalpies of alloys within framework of miedema's theory, Comput. Phys. Commun. **209**, 58 (2016).

[59] M. Liu, W. Xu, S. Zhang, Z. Wang, Z. Wang, B. Wang, D. Wang, and F. Li, Microstructures and hardnesses of AlCoCr0.5FexNi2.5 high entropy alloys with equal valence electron concentration, J. Alloys Compd. **824**, 153881 (2020).

[60] O. N. Senkov, G. B. Wilks, J. M. Scott, and D. B. Miracle, Mechanical properties of $Nb_{25}Mo_{25}Ta_{25}W_{25}$ and $V_{20}Nb_{20}Mo_{20}Ta_{20}W_{20}$ refractory high entropy alloys, Intermetallics **19**, 698 (2011).

[61] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevMaterials.7.043802 for results from implementation of CoSMoR on discrete decision-based Random Forest (RF) models. Figure S1 shows feature contributions extracted using CoSMoR for RF models that were trained to predict presence/absence of fcc, bcc, and intermetallic phases in multiprincipal element alloys.

[62] https://github.com/IDEAsLab-Materials-Informatics/CoSMoR.