Predicting phase preferences of two-dimensional transition metal dichalcogenides using machine learning

Pankaj Kumar,[1] Vinit Sharma [⦿],[2] Sharmila N. Shirodkar,[1] and Pratibha Dev [⦿][1,*]

[1]*Department of Physics and Astronomy, Howard University, Washington, D.C. 20059, USA*
[2]*National Institute for Computational Sciences, Oak Ridge National Laboratory, Oak Ridge, Tennessee 37831, USA*

Two-dimensional transition metal dichalcogenides (TMDs) can adopt one of several possible structures, with the most common being the trigonal prismatic and octahedral symmetry phases. Since the structure determines the electronic properties, being able to predict phase preferences of TMDs from just the knowledge of the constituent atoms is highly desired, but has remained a long-standing problem. In this study, we applied high-throughput quantum mechanical computations with machine learning algorithms to solve this old problem. Our analysis provides insights into determining physiochemical factors that dictate the phase preference of a TMD, identifying and going beyond the attributes considered by earlier researchers in predicting crystal structures. A knowledge of these underlying physiochemical factors not only helps us to rationalize, but also to accurately predict structural preferences. We show that machine learning algorithms are powerful tools that can be used not only to find new materials with targeted properties, but also to find connections between elemental attributes and the target property/properties that were not previously obvious.

I. INTRODUCTION

Layered transition metal dichalcogenides (TMDs), given by the formula MX_2 (M = transition metal; X = chalcogen), represent a chemically well-defined family of materials. As bulk crystals, TMDs are formed from weakly-bonded van der Waals stacks of MX_2 layers, and were studied extensively in the 1960s–1980s [1–4]. Interest in TMDs was revived nearly two decades ago when it was realized that bulk TMD crystals can be separated into two-dimensional (2D) monolayers and few-layer thick crystals [5]. Hence, a considerable experimental effort has been dedicated to their synthesis [6–12]. The diverse electronic, magnetic and topological properties of TMDs [13–17] makes them attractive for a number of technologies [18–28]. With the recent creation of novel Janus TMD crystals, MXY, in 2017 (X and Y are two different chalcogens on two faces) [29], it has become clear that the TMD family is growing larger, along with the list of their properties, which now includes ferroelasticity and ferroelectricity [29–31].

Unlike 2D layered materials such as graphene and hexagonal boron nitride, TMDs can exist in multiple crystal phases. Both traditional (MX_2) and Janus (MXY) TMDs most commonly adopt either a trigonal prismatic phase (*1H* phase) or an octahedral symmetry phase (*1T* phase). As the structure determines their properties, knowledge of factors that determine phase-preferences of TMDs is very important and has long been debated in the literature [3,4]. These earlier works, using chemical and physical considerations, showed how the structural preferences are related to different factors, such as:

(i) the Pauling's cationic to anionic radius ratio [32], with the ratio obtained by combining the structural constraints with the fractional ionicity of bonds within a TMD layer [3], and (ii) the d electron counts of the transition metals [4]. All of these earlier studies involved traditional TMDs that were known at the time of the respective studies. Due to the rapid materials discoveries in recent years, which include discoveries of novel TMDs, it is important to develop phase-prediction methods that can predict preferred structures for a much larger chemical phase-space and are generalizable to even different classes of materials, while still having roots in physical and chemical intuition.

In this paper, we created a generalizable framework for predicting phase preferences of TMDs by combining high-throughput quantum mechanical computations with machine learning (ML) algorithms (see detailed discussion within the Supplemental Material [33]). This powerful combination can help to explore the materials phase space more economically, and has been successfully deployed in several studies [34–40]. We also went beyond the known TMDs, creating novel combinations of transition elements and chalcogen atoms, thereby expanding the chemical phase space over which our analysis is valid. Density functional theory (DFT) calculations were used to generate a database of formation energies (ΔE_f) for different TMDs. Machine learning models were then trained and tested for their ability to predict chemical stability and phase preference of the TMDs. We quantitatively show that the most important factors determining the formation energy, as well as the crystal structure of a TMD, are the relative elemental dipole polarizabilities, electronegativities, ionization energies, and commonly exhibited valencies of the elements comprising the TMD structure. Two additional attributes that play a role (albeit to a lesser extent) are the electron-affinities
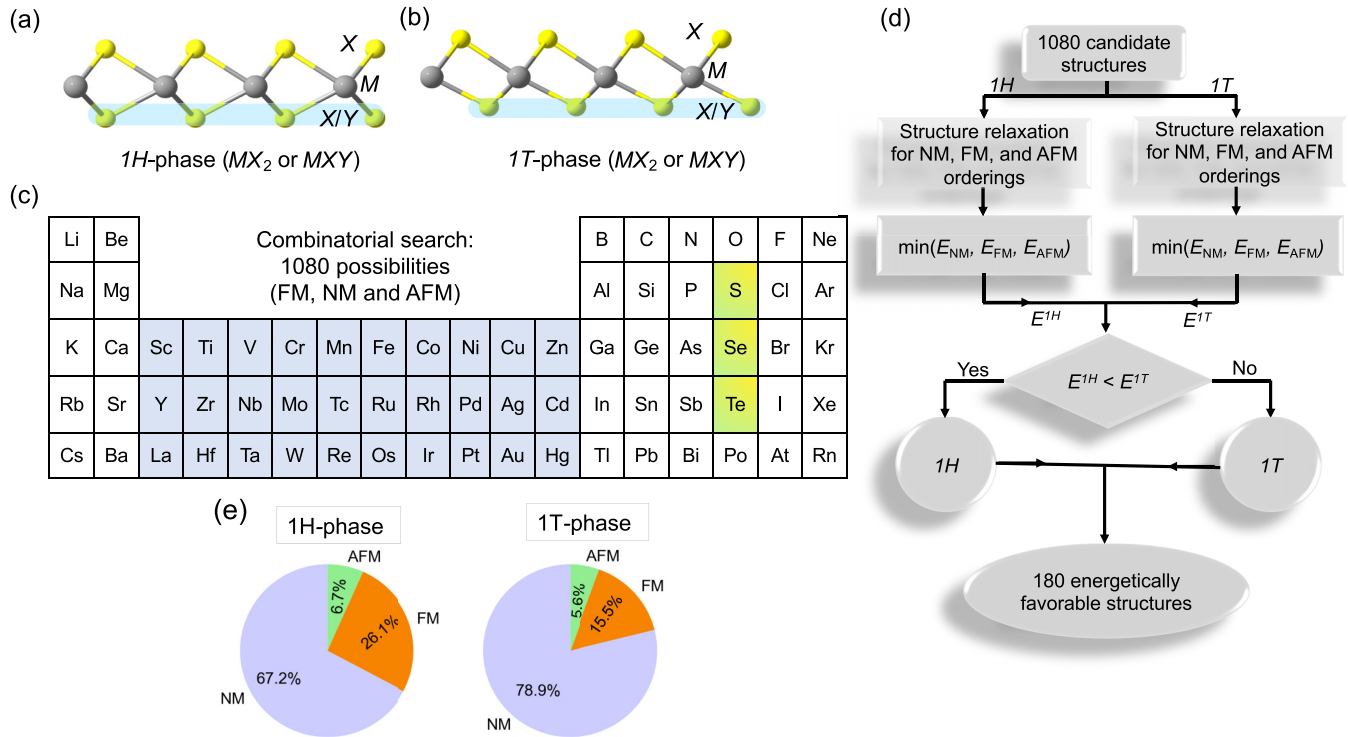
* pratibha.dev@howard.edu

FIG. 1. Schematic representation of (a) *1H* and (b) *1T* phase, traditional and Janus TMD monolayers. The Janus structure consists of two different chalcogens (*X* and *Y*) on two faces of a TMD. (c) Periodic table highlighting the transition elements and the chalcogen atoms considered for the combinatorial study. (d) Flow chart of systematic combinatorial search of traditional and Janus monolayers. (e) Pie charts showing the % distributions of the most stable magnetic orderings—ferromagnetic (FM), antiferromagnetic (AFM), and nonmagnetic (NM)—for *1H* phase and *1T* phase TMDs.

and heat of gas-phase formation. We also found that swapping the dipole polarizabilities with covalent radii, and vice versa, in our analysis yields nearly equivalent ML models. This is not surprising because an atom's dipole polarizability is a size-dependent attribute and is a measure of the covalency of the bonds that it forms. The relative radii, degree of ionicity (as given by the difference of electronegativities) and relative valencies (related to d-electron counts), are expected to contribute to the relative stability of the *1H* and *1T* phases. These dependencies are in accordance with the papers by Gamble [3] and by Kertesz *et al.* [4] who used chemical and physical arguments to rationalize phase preferences, while the other features represent previously unknown contributors. Hence, by employing ML algorithms to find new connections between relative elemental attributes of the constituents and the target property/properties of the resulting compounds, we are able to both rationalize and predict chemical stabilities and structural preferences of the TMDs.

II. TECHNICAL DETAILS

A. High-throughput first principles calculations

In this paper, we used ML algorithms to identify relationships between phase-stability of TMDs and their respective chemistries in a high-dimensional space of independent variables. Since the accuracy of datasets as well as their sizes dictate the accuracy of ML models trained on these datasets, it was important to go beyond known TMDs. Hence, we

performed high-throughput first principles calculations to generate both conventional and Janus TMDs [Figs. 1(a) and 1(b)] from different combinatorial possibilities of chalcogens (*X, Y* = S, Se, Te) and transition metal (TM) atoms belonging to the 3d-, 4d- and 5d series [highlighted in Fig. 1(c)]. There were a total of 360 chemically- and structurally-distinct crystals (i.e., *1H* phase and *1T* phase structures) formed from 30 transition elements (Sc-Zn, Y-Cd, and La-Hg), and three chalcogen atoms. Since these structures can be magnetic, each TMD was further allowed to adopt ferromagnetic (FM), antiferromagnetic (AFM), and nonmagnetic (NM) configurations, yielding a total of 1080 candidate structures [see Fig. 1(d)]. Our spin-resolved DFT calculations were performed using the Vienna *ab initio* simulations package (VASP) [41,42]. The Perdew, Burke, and Ernzerhof (PBE) generalized gradient approximation [43] was used to account for the exchange-correlation effects. In order to minimize interactions between the periodic images of the monolayers, a large vacuum of 20 Å was used in the direction normal to the 2D layers. In order to determine if the TMDs will be magnetic, we calculated the energy differences (per formula unit) between the nonmagnetic and magnetic structures, where the latter can have either FM or AFM alignments of the moments. For the FM alignment, we performed a spin resolved calculation, wherein the moments on the atoms within a unitcell were initialized to non-zero values. This generated a ferromagnetic structure since the periodic boundary conditions ensured that the moments (if nonzero) in the cell and its images were all

aligned. To create magnetic structures with AFM alignment of moments, we doubled the size of the unit cell in one of the lateral directions, and initialized the moments on the magnetic species in the neighboring cells so as to align them antiferromagnetically [44,45]. All structures were optimized using an energy cutoff of 550 eV and a Γ-centered k-point mesh [46], which was equivalent to a $21 \times 21 \times 1$ grid for a unit cell, until the force on each of the atoms was less than 1 meV/Å and the total energy values converged to better than 10^{-7} eV.

Formation energies for different TMD monolayers were calculated using

$$\Delta E_f = E_{MXY} - (E_M + E_X + E_Y). \qquad (1)$$

Here, E_{MXY} is the total energy of the monolayer. E_M, E_X, and E_Y are the total energies for the the transition metals (M) and the chalcogens (X, Y), respectively, in their most stable bulk form. Formation energy is an indicator of a material's chemical stability, although it does not account for possible mechanical instabilities. For the latter, one needs to determine dynamical stability by calculating the phonon spectra, and thermal stability using *ab initio* molecular dynamics at different temperatures. It is also worth mentioning that some of the materials that may display mechanical instabilities as freestanding monolayers can possibly be stabilized on appropriate substrates. However, determining mechanical stability and/or conditions under which a monolayer can be stabilized is beyond the scope of present paper.

B. ML-based materials property prediction

1. Primary descriptor/feature dataset

Choosing an optimal set of independent descriptors/features is of utmost importance in building effective machine learning models. As a first step, our goal was to find easily accessible descriptors that can uniquely describe the system and provide a meaningful mapping to the target property. To achieve this goal, we initially chose thirteen elemental properties of the atoms that form the TMD monolayers. More specifically, the primary atomic and bulk elemental descriptors considered here are: heat of gas-phase formation (HF), metallic radius (MR), specific heat (SP), covalent radius (CR), electron affinity (EA), atomic weight (AW), electronegativity (EN), heat of fusion (HS), atomic radius (AR), dipole polarizability (DP), ionization energy (IE), valency (VS), and Shannon ionic radius (IR) of the elemental species forming the TMDs [47,48]. Since we studied compounds in this paper, we actually needed compound descriptors/features that could describe a TMD with a particular combination of elements. Hence, instead of the aforementioned elemental attributes in their unchanged forms, we considered different combinations of the elemental descriptors to form compound descriptors. We also refrained from using properties of TMDs themselves (e.g., lattice constants) that one obtains only after performing a DFT-calculation or an experiment on a TMD, which would have defeated the purpose of using machine learning. The compound descriptors formed from the elemental descriptors were divided into three sets: (i) Set A, formed by finding the ratio between the elemental attributes of the metal and the averaged attributes of the chalcogens forming each of the TMDs, such that a compound attribute in Set A is given by $\mathcal{F}_{M/\bar{X}} = 2\,\mathcal{F}_M/(\mathcal{F}_X + \mathcal{F}_Y) = \mathcal{F}_M/\mathcal{F}_{\bar{X}}$, where $\mathcal{F}_M$, $\mathcal{F}_X$, $\mathcal{F}_Y$ and $\mathcal{F}_{\bar{X}}$ are the elemental attributes of atoms M, X, Y, and the averaged attributes of the chalcogen atoms [$\mathcal{F}_{\bar{X}} = (\mathcal{F}_X + \mathcal{F}_Y)/2$], respectively. (ii) Set B, obtained by finding the difference in the averaged attributes of the chalcogen atoms and the elemental attributes of M atoms, such that a compound attribute in this set is represented by $\mathcal{F}_{\bar{X}M} = \mathcal{F}_{\bar{X}} - \mathcal{F}_M$. (iii) Set C, obtained by finding the difference of the elemental properties of the two chalcogens involved in the formation of different TMDs, such that a compound attribute is given by $\mathcal{F}_{XY} = \mathcal{F}_X - \mathcal{F}_Y$. Sets A–C thus contained a total of 39 compound descriptors. In order to identify independent compound descriptors, we calculated the Pearson correlation coefficients between the features. Formally, the Pearson correlation coefficient (or bivariate correlation) between two variables is obtained by finding the covariance of the two variables and dividing it by the product of their standard deviations. The Pearson correlation coefficient $\mathcal{P}$, when applied to the features is effectively obtained by taking the inner product,

$$\mathcal{P}(\vec{u}, \vec{v}) = \frac{\sum_{j=1}^{N}[u_j - \bar{u}][v_j - \bar{v}]}{\sqrt{\sum_{j=1}^{N}[u_j - \bar{u}]^2}\sqrt{\sum_{j=1}^{N}[v_j - \bar{v}]^2}}. \qquad (2)$$

Here, $\vec{u}$ and $\vec{v}$ are any two of the M descriptors, and are $N \times 1$-dimensional vectors, N being the total number of materials. The mean value of the descriptor $\vec{u}$ is given by $\bar{u} = (1/N)\sum_{j=1}^{N} u_j$, where u_j is the jth component of $\vec{u}$. As the inner product in Eq. (2) is normalized, the lower and upper bounds of $\mathcal{P}$ are -1 and $+1$, corresponding to strong negative and positive linear relationships between the features, respectively. A value of $\mathcal{P} = 0$ indicates that the features are completely independent.

2. Machine learning algorithms

The down-selected features were used as inputs in supervised learning of formation energies. For this, we employed two ML algorithms—random forest (RF) and Kernel ridge regression (KRR).

Random forest algorithm. The RF algorithm was used for two tasks: (i) regression to predict formation energies of the TMDs and (ii) classification to predict the phase of the TMDs [49]. RF consists of constructing a large number of individual uncorrelated decision trees (models) that work as an ensemble [50]. In order to ensure that the different models (decision trees) are uncorrelated, each independently-constructed decision tree randomly samples the dataset with replacement, thereby avoiding overfitting. The final prediction is made by aggregating, either by averaging or with a simple majority vote. This ML method has only two parameters: (i) the number of variables in the random subset at each node and (ii) the number of trees in the forest. Although RF is not very sensitive to the specific values chosen for these two parameters, it is generally desirable to optimize these parameters during the model training stages.

Kernel ridge regression. In addition to RF regression, we also applied KRR to verify the robustness of our predictions.
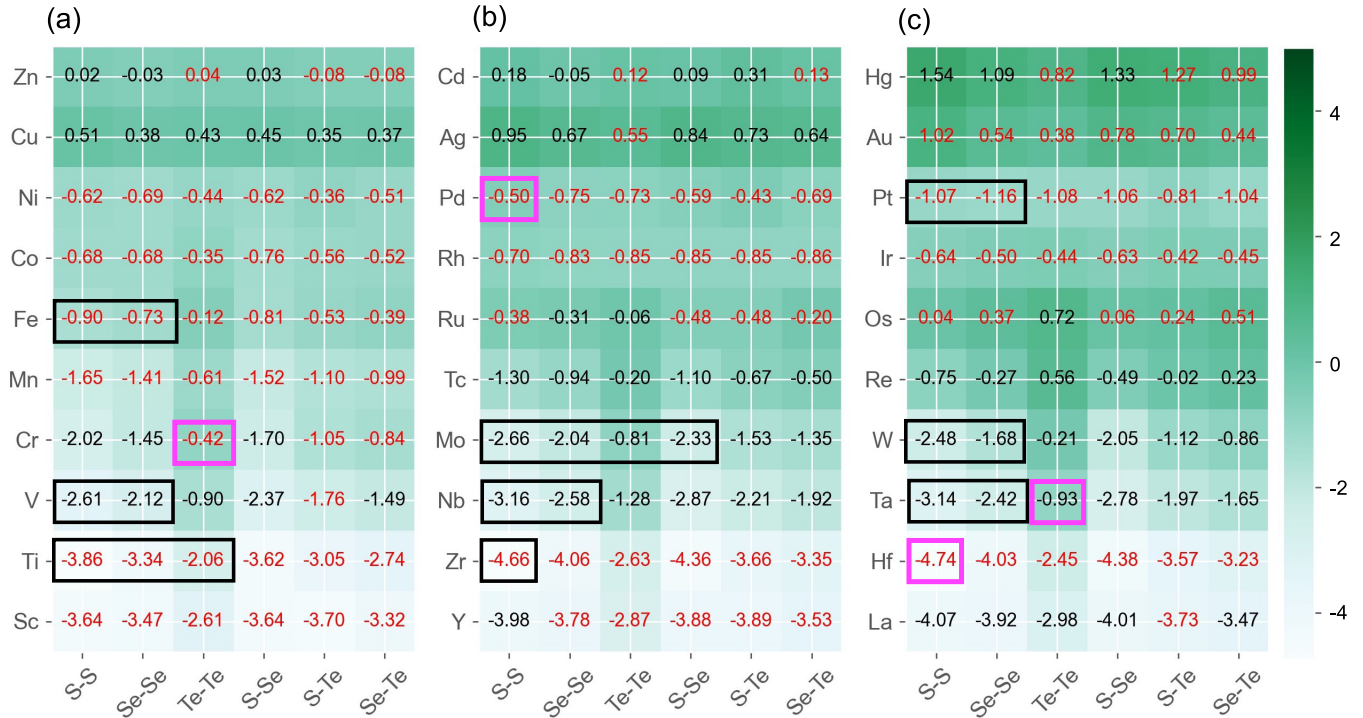
(a)

	S-S	Se-Se	Te-Te	S-Se	S-Te	Se-Te
Zn	0.02	-0.03	0.04	0.03	-0.08	-0.08
Cu	0.51	0.38	0.43	0.45	0.35	0.37
Ni	-0.62	-0.69	-0.44	-0.62	-0.36	-0.51
Co	-0.68	-0.68	-0.35	-0.76	-0.56	-0.52
Fe	-0.90	-0.73	-0.12	-0.81	-0.53	-0.39
Mn	-1.65	-1.41	-0.61	-1.52	-1.10	-0.99
Cr	-2.02	-1.45	-0.42	-1.70	-1.05	-0.84
V	-2.61	-2.12	-0.90	-2.37	-1.76	-1.49
Ti	-3.86	-3.34	-2.06	-3.62	-3.05	-2.74
Sc	-3.64	-3.47	-2.61	-3.64	-3.70	-3.32

(b)

	S-S	Se-Se	Te-Te	S-Se	S-Te	Se-Te
Cd	0.18	-0.05	0.12	0.09	0.31	0.13
Ag	0.95	0.67	0.55	0.84	0.73	0.64
Pd	-0.50	-0.75	-0.73	-0.59	-0.43	-0.69
Rh	-0.70	-0.83	-0.85	-0.85	-0.85	-0.86
Ru	-0.38	-0.31	-0.06	-0.48	-0.48	-0.20
Tc	-1.30	-0.94	-0.20	-1.10	-0.67	-0.50
Mo	-2.66	-2.04	-0.81	-2.33	-1.53	-1.35
Nb	-3.16	-2.58	-1.28	-2.87	-2.21	-1.92
Zr	-4.66	-4.06	-2.63	-4.36	-3.66	-3.35
Y	-3.98	-3.78	-2.87	-3.88	-3.89	-3.53

(c)

	S-S	Se-Se	Te-Te	S-Se	S-Te	Se-Te
Hg	1.54	1.09	0.82	1.33	1.27	0.99
Au	1.02	0.54	0.38	0.78	0.70	0.44
Pt	-1.07	-1.16	-1.08	-1.06	-0.81	-1.04
Ir	-0.64	-0.50	-0.44	-0.63	-0.42	-0.45
Os	0.04	0.37	0.72	0.06	0.24	0.51
Re	-0.75	-0.27	0.56	-0.49	-0.02	0.23
W	-2.48	-1.68	-0.21	-2.05	-1.12	-0.86
Ta	-3.14	-2.42	-0.93	-2.78	-1.97	-1.65
Hf	-4.74	-4.03	-2.45	-4.38	-3.57	-3.23
La	-4.07	-3.92	-2.98	-4.01	-3.73	-3.47

FIG. 2. Heatmaps showing calculated formation energies (in eV per formula unit) of traditional and Janus TMD monolayers of (a) 3*d*-, (b) 4*d*-, and (c) 5*d*-series elements in their most stable phase. The values printed in black (red) color correspond to the TMDs that prefer the *1H* phase (*1T* phase). Black boxes highlight examples of TMDs that have already been synthesized in either the *1H* phase or *1T* phase, whereas those in pink already exist in multilayers or in bulk, with a possibility of creating monolayers from those structures.

KRR implements the L2 regularization to prevent overfitting of the data, and takes into account different nonlinear effects by invoking various kinds of kernel functions. Within KRR, the predicted target property, $f^{ML}(j)$, of the *j*th material is expressed as $f^{ML}(j) = \sum_i^n \alpha_i \mathcal{K}(\vec{x}_i, \vec{x}_j)$ Here, the summation is carried out over the entire training set with *n* materials, α_i's are the weights and $\mathcal{K}(\vec{x}_i, \vec{x}_j)$ is the kernel function, representing the Euclidean distance between the feature vectors of the *i*th material from the training set and the *j*th material. To predict the formation energies of TMDs, we have used the Gaussian kernel, defined as: $\mathcal{K}(\vec{x}_i, \vec{x}_j) = exp(-\gamma |\vec{x}_i - \vec{x}_j|^2)$. Training the KRR model involves finding the optimized values of the kernel coefficients α_i, and the Gaussian width γ, through iterative minimization of the errors in the prediction.

III. RESULTS AND DISCUSSION

A. Monolayer TMD dataset

The flowchart depicted in Fig. 1(d) outlines our systematic combinatorial search of different 2D TMD monolayers. Starting with 1080 total structures that consisted of FM, AFM, and NM magnetic orderings for both *1H* and *1T* phases of the TMDs, we obtained the lowest energy structures from amongst the three possible magnetic orderings. This provided us with 180 structures within each of the two phases. These 360 structures were subsequently used to create machine learning models (detailed discussion in the next section). The pie chart in Fig. 1(e) shows the relative percentages of the structures that are FM, AFM, or NM for the TMDs in the *1H* and *1T* phases. In both phases, most of the TMDs are

nonmagnetic, consistent with previous results for already known TMDs [51,52]. Nevertheless, there is still a significant fraction of TMDs that adopt FM or AFM structures and can potentially have tremendous technological impact as 2D magnets. The pie charts in Fig. 1(e), also reveal that a larger percentage of structures in the *1H* phase (as compared to those in *1T* phase) adopted either FM or AFM ordering. The preferred magnetic orderings of each TMD in the *1H* and *1T* phases, along with their spin polarization energies and exchange energies per formula unit (f.u.) are provided within the Supplemental Material (see Tables S1– S3) [33]. Moreover, these results are in agreement with previous experimental and theoretical reports [52–58].

In order to study the chemical stability of the screened traditional and Janus TMDs, we computed the formation energies of the relaxed structures using Eq. (1). Further comparison between the two phases for each of the TMDs reduces the number of lowest energy structures to 180 (either *1H* or *1T* phase). Out of these 180 structures, 74 TMDs preferred to be in the *1H* phase and 106 preferred to be in the *1T* phase. Figure 2 shows the computed formation energies/f.u. of monolayers of the 180 TMDs in their most favored phase. The formation energy values given in black and red correspond to the TMD monolayers that preferentially form in the *1H* phase and *1T* phase, respectively. Some of these 180 TMDs have already been synthesized (examples indicated by black boxes in Fig. 2). TMDs encased in pink boxes in Fig. 2 are examples of those materials for which bulk counterparts exist, with a possibility of creating the monolayers from those structures. We find that several TMDs exhibit small energy

differences (less than 30 meV/f.u.) between the *1H* and *1T* phases, and are therefore promising phase-change materials. These are: VTe_2, $VSTe$, $VSeTe$, $ZnSe_2$, YSe_2, $YSSe$, RuS_2, $RuSe_2$, $TaTe_2$, and $OsTe_2$ (see Tables S1– S3 within the Supplemental Material [33]). The small energy differences suggest that these TMDs, can either exist in both phases (one of them being metastable), or can readily undergo a phase transition between the two phases upon application of external stimuli, such as gating, doping and applying strain. Owing to a similar energy difference in two of its lowest energy phases, one of the well-known TMDs, $MoTe_2$ undergoes a phase transition from the *1H* phase to the distorted octahedral symmetry, T' phase [21,27]. The latter phase was not included in this study. Also note that a few of the TMDs, such as WTe_2, ReS_2, and $ReSe_2$, are known to form in low-symmetry distorted phases—T' and T'' [54,59,60]. However, as the *1H* and *1T* phases are the most common phases among TMDs, we have not included the other possible structures at this stage. In addition, it is also possible for these materials to exist in higher symmetry phases (which may be metastable states for the crystals). Nevertheless, the dynamical and thermal stabilities as well as distances from the convex hull of the most promising novel TMDs, which are predicted to be chemically stable in this work (listed in Tables S1– S3 within the Supplemental Material [33]), need to be explored in future studies.

Figures 2(a)–2(c) show which of the TMDs are thermodynamically stable (i.e., have negative values for the formation energy). These formation energies become less negative across the TM series (e.g., across the 3*d* series) and finally become positive, suggesting that the TMDs, which involve TM elements at the end of the series, are unstable. The inclusion of a large number of conventional and Janus TMDs in this paper allowed us to effectively capture and recognize physicochemical trends in thermodynamically stable TMDs. The data trends in the calculated formation energies (Fig. 2) hint at two different connections. First, the number of d electrons increases, while the atomic radii of the TM atoms generally decrease as we go across a period (e.g., from left to right in the 3*d* series of TM). As mentioned above, the stability of the TMDs also decreases as the TM radii decrease across a period, implying that the TM radii and the *d* electron counts play a major role in the stability of the structures. Second, for a fixed transition element (e.g., Mo), the formation energy decreases as one goes from S → Se → Te (for example, see the ΔE_fs listed in Fig. 2 for MoS_2, $MoSe_2$, and $MoTe_2$). This trend is generally seen for most of the stable traditional and Janus TMDs, though not for all. For a fixed TM, what has varied are the properties of the chalcogens, such as their electronegativities and their radii, which generally increase as one goes down a group.

The aforementioned observations are consistent with Pauling's rules [32] for ionic solids, and the work by F. R. Gamble [3] on layered dichalcogenides, according to which the ratio of the atomic radii of the cations and anions, and the difference in their electronegativities play an important role in deciding the geometrical phase preferences of these materials. They are also consistent with the findings of Kertesz *et al.* [4], who explained the dependence of phase stability on *d* electron counts. Hence, even before we apply ML algorithms, the observed trends hint at relationships between the formation energies and different attributes of the elements within the TMDs. Nevertheless, a deeper analysis is required to reveal and quantify the connections between different attributes of elements involved in a TMD, and its formation energy and phase stability.

B. Down-selection of features

A large number of descriptors increases the dimensionality and complexity of the machine learning model, not only making it difficult to interpret the results, but also leading to over-fitting. As there were a large number of compound descriptors (see Sec. II Technical Details), we first needed to find a subset of highly relevant feature vectors in this large phase space of variables. To do so, we calculated the pairwise Pearson correlation coefficients for all feature vectors, as well as the correlation coefficients between the feature vectors and the target property vector (i.e., formation energy), which was calculated using DFT. If any two compound features showed $|\mathcal{P}| > 0.85$, only the one with the larger correlation with the target property was kept. Simultaneously, we also required the feature vectors to have $|\mathcal{P}| > 0.15$ with the property vector. We found that the feature vectors from Sets A and B not only have large pairwise correlations with each other, but also have comparable pairwise correlations with the target property (see Table S4 within the Supplemental Material [33]). In addition, we found that compound feature vectors from Set C, created by subtracting the elemental attributes of the two chalcogen atoms (involved in a TMD), showed poor correlation with the target property and hence, were discarded. Therefore, in order to reduce the complexity of the model, we considered only the feature vectors of Set B.

The pairwise correlations of the feature vectors from Set B amongst themselves and with the target property are shown in a correlogram chart in Fig. 3(a). The absolute values of feature-feature and feature-target property correlations are pictorially represented through pie charts in the upper triangle and colored boxes in the lower triangle, both presenting the same information. The blue and red colors corresponds to positive and negative correlations, and the relative intensity of their colors corresponds to the relative strengths. Figure 3(a) shows that not only are the covalent radii and dipole polarizabilities highly correlated, even their feature-target property correlations are comparable, being 0.81 for $CR_{\bar{X}M}$ and 0.83 for $DP_{\bar{X}M}$. The large linear dependence between the covalent radii and dipole polarizabilities is understandable because the atomic dipole polarizability is highly correlated with an atom's size. The larger the size, the more polarizable an atom is, and in turn, the more readily it will form covalent bonds. It should also be pointed out that it is the differences in the covalent radii of the atoms involved that are more important as a feature as compared to the differences in ionic radii [see Fig. 3(a)]. In contrast, in the earlier paper by Gamble [3], wherein the author studied deciding factors that determine phase preferences (though not the formation energies), it was found that the ionic radii are more important attributes. However, TMDs are mostly covalent, with a very small fractional ionic character [3,61]. Our correlation analysis highlights this fact by assigning a much greater correlation fraction
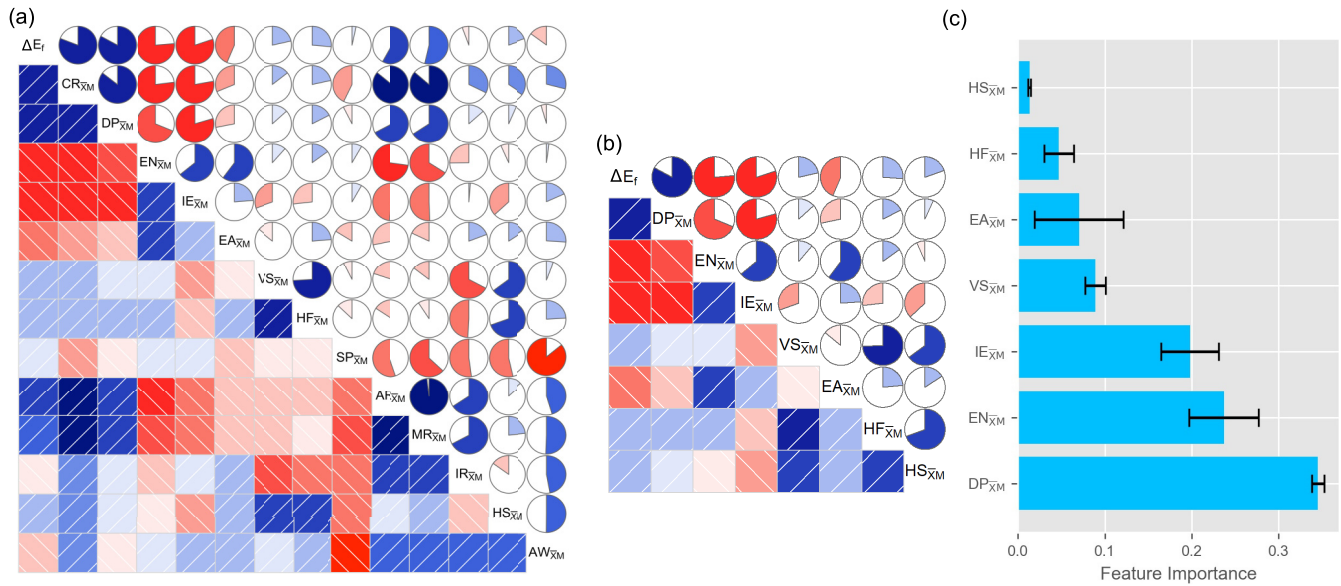
FIG. 3. Down-selection of features: Pearson correlation for (a) all thirteen features from Set B and (b) seven down-selected features from Set B, summarizing the pairwise correlation of different features among themselves and with the target property vector (formation energy ΔE_f). The target property and the features are listed along the diagonal; the blue and red color correspond to positive and negative correlation coefficients, respectively. The boxes in the lower triangle and the pie charts in the upper triangle are two different pictorial representations of the same information, with the absolute value of the associated Pearson correlation coefficient given by either the lighter and darker shades of the colors in the boxes, or the filled fraction of the pie chart. (c) Relative importance of Set B feature vectors in predicting the formation energies of TMDs computed using the RF algorithm. The subscript $\bar{X}M$ indicates that the down-selected compound features are obtained as differences of the average of attributes of the chalcogens with the that of the metal atom's attributes.

to covalent radii differences as compared to the ionic radii differences. As the covalent radii and dipole polarizabilities show comparable feature-target property correlations, we decided to create two different subsets of down-selected features based on our aforementioned criteria—one subset included $DP_{\bar{X}M}$ and another included $CR_{\bar{X}M}$. From this point forward, we present our analysis for the short-listed feature subset (from set B) that includes $DP_{\bar{X}M}$, while the analysis for the $CR_{\bar{X}M}$ -containing subset is presented within the Supplemental Material [33].

The correlogram chart in Fig. 3(b) summarizes the pairwise correlations between the $DP_{\bar{X}M}$-containing subset of seven down-selected features. Thus far, we have used Pearson correlation for the first round of feature down-selection, which helped to reduce the dimensionality of the problem. The next round of down-selection of features from Set B came from the ML algorithms themselves. In this round of down-selection, we calculated the relative feature importances using an RF model, aiming to predict the thermodynamic stability of the TMDs. As mentioned in the previous section, the RF regression algorithm has two main hyperparameters: number of trees and maximum allowable depth. Tuning these parameters is essential to avoid suboptimal performance and overfitting of the model. Therefore, to find optimal values of these parameters, we trained the RF regression algorithm over 90% of the dataset and tested the model with 10% unseen data for each set of hyperparameters. The ML model was trained over 20 randomly distributed train/test splits of data for different k-fold cross-validations (with k = 5, 8, 10). We then computed the average root mean squared error (RMSE) and R^2-score (goodness-of-fit score) using the dataset from the *1H* phase of

the TMDs. The average R^2-scores and RMSEs for different cross-validations are given in the Table S5 within the Supplemental Material [33]. A 10-fold cross-validation was used in the rest of the paper as it shows very good training and testing R^2-scores and RMSE values. The optimized number of trees and maximum allowable depth for 10-fold cross-validation are 48 and 16, respectively. Using the optimized values of hyperparameters, we computed the importance of different features in predicting the formation energies of the TMDs. The bar graph in Fig. 3(c) shows the mean values of the relative feature importance, with the error bars representing their standard deviations, obtained from 20 randomly distributed train/test splits of data. Figure 3(c) shows that $DP_{\bar{X}M}$, $EN_{\bar{X}M}$, and $IE_{\bar{X}M}$ are the most important features, with a relatively small contribution from $VS_{\bar{X}M}$, $EA_{\bar{X}M}$, and $HF_{\bar{X}M}$, corroborating the results from our pair-wise correlation study. Since $HS_{\bar{X}M}$ has a negligible effect on the prediction accuracy of the formation energy, following the general *principle of parsimony*, we excluded $HS_{\bar{X}M}$ from the down-selected feature set. The six chosen features were used to design machine learning models.

Before implementing the ML algorithms, we would like to further elaborate upon the physically-intuitive results that emerge from our down-selection analysis of the feature set. In Fig. 4, we plot the down-selected features as a function of the formation energy for the *1H* [Figs. 4(a)–4(f)] and *1T* [Figs. 4(g)–4(l)] phases. Overall, one can see a strong dependence of formation energy on $DP_{\bar{X}M}$ (equivalently, $CR_{\bar{X}M}$ in Fig. S1 within the Supplemental Material [33]), $EN_{\bar{X}M}$, $IE_{\bar{X}M}$, and $VS_{\bar{X}M}$, making them the key features. Figures 4(a) and 4(g) show that the data are more scattered as the difference in dipole polarizabilities ($DP_{\bar{X}M}$) decreases (same trend for
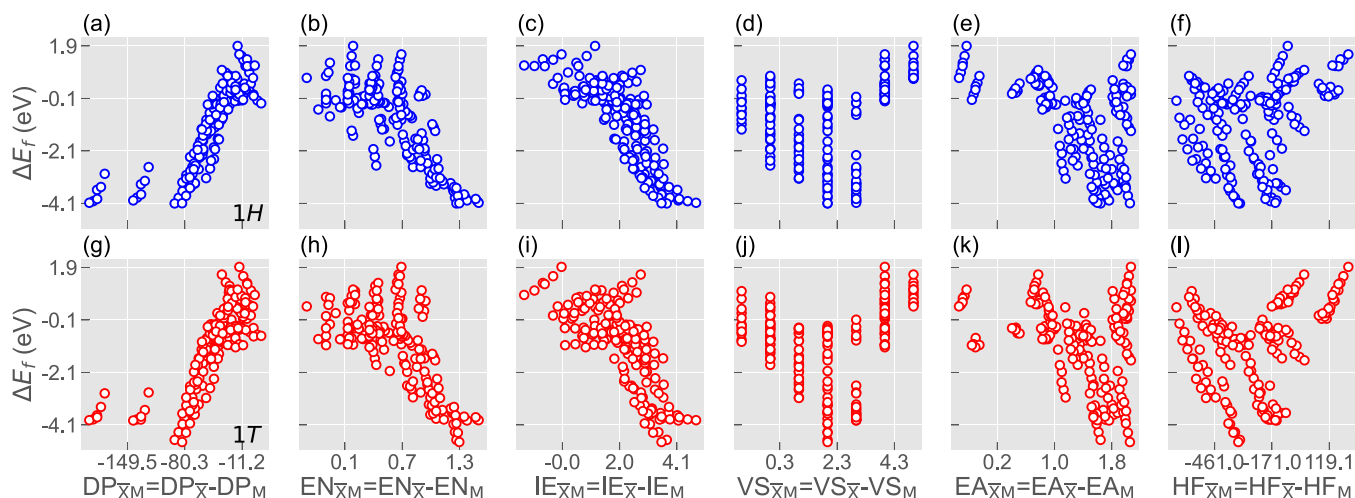
FIG. 4. Formation energy vs the down-selected feature vectors for different TMDs in: [(a)–(f)] the *1H* phase and [(g)–(l)] *1T* phase. Along the *x* axis, the compound feature vectors are computed from elemental features according to $\mathcal{F}_{\bar{X}M} = \frac{1}{2}(\mathcal{F}_X + \mathcal{F}_Y) - \mathcal{F}_M$.

covalent radii), indicating that dipole polarizabilities (covalent radii) alone are not a deciding factor in the thermodynamical stability of TMDs. Another important key feature vector is the difference in electronegativities of the TM atom and chalcogens ($EN_{\bar{X}M}$), which accounts for the fractional ionic character [Figs. 4(b) and 4(h)]. The fractional ionic character f_i, as defined by Pauling is given by $f_i = 1 - exp[-(EN_A - EN_B)^2/4]$, where EN_A and EN_B are the respective elemental electronegativities of a binary compound of elements A and B. Although there is a spread of the formation energy values, the overall trend shows a decrease in the formation energy with increasing values of $EN_{\bar{X}M}$. The relative ionization energies of atoms forming the materials ($IE_{\bar{X}M}$) is generally related to how readily the respective atoms enter into chemical reactions and form bonds with each other. It is, therefore, not surprising that this is one of the important features selected by the correlation analysis [Figs. 4(c) and 4(i)]. Lastly, the relative difference in valencies of the cations and anions $VS_{\bar{X}M}$ shows a correlation with the formation energies, displaying a U-shaped trend, with the bottom of the U shape at $V_{\bar{X}M} = 2$ [Figs. 4(d) and 4(j)]. In the compound attribute $VS_{\bar{X}M}$, the average valency of the chalcogens is fixed as we go from one TMD to the next, while that of the transition element changes according to its *d*-electron count. This trend for $VS_{\bar{X}M}$ is understandable, based on work by Kertesz *et al.* [4], who explained how the stability of TMDs changes as a function of *d*-electron counts. Although no such obvious trends can be seen for the $EA_{\bar{X}M}$ [Figs. 4(e) and 4(k)] and $HF_{\bar{X}M}$ [Figs. 4(f) and 4(l)] feature vectors, they were included when creating an ML regression model because the calculated feature importance [Fig. 3(b)] for these two attributes is comparable to that for the valence state.

C. Machine learning algorithms for regression and classification

1. Regression algorithms for formation energies

Having established the physical rationale behind the down-selected feature vectors, we applied regression algorithms to predict the formation energy of TMDs. The robustness of the statistical models was tested in several steps, which are described below. Unless stated otherwise, all ML regression

models were trained on *1H*, *1T* or the complete datasets using 20 different randomly selected train/test splits of 90%/10%. The resulting average R^2-scores and RMSE values were then calculated to determine each model's performance.

We first trained the RF regression model to predict formation energies of TMDs in the *1H* phase. With the six down-selected features, the average training and test R^2-scores are 98.7% and 96.3%, respectively. The corresponding average RMSEs for the training and test data points are 0.17 eV and 0.28 eV, respectively. The prediction performance of the model for the *1H* dataset is shown in Fig. 5(a). Most of the data points lie on or are near the straight line (so-called line of accuracy), suggesting that the model is highly accurate. The deviation from the line of accuracy is quantified by calculating the error in predicting the DFT computed formation energy at each point and is indicated by an error bar. The error bars for the data points with negative formation energies are relatively smaller than those for data points with positive formation energies, indicating that our trained RF regression model is able to accurately predict systems that are structurally stable. We also went one step further and tested the prediction performance of our ML model trained on the *1H* data and testing it with 10% of the *1T* phase dataset, with accuracy scores of 91.5%.

Figure 5(b) plots the prediction performance of the RF model trained on the *1T* dataset. This model achieves an average accuracy of 98.5% for training data and 96.1% for test data, with the corresponding RMSEs of 0.19 eV and 0.31 eV respectively. The prediction performance of our ML model trained on the *1T* data was also tested on 10% of the *1H*-phase dataset, with an accuracy score of 94.0%. The average RMSEs numbers for the *1H* and *1T* testing datasets are comparable [see Figs. 5(a) and 5(b)], showing the robustness of the ML model in predicting the chemical stability of both crystal geometries. Hence, these results show that the six-feature ML model is efficient in predicting the formation energy over a vast chemical space and for different phases. Furthermore, we verified the robustness of the ML model by showing that our results are insensitive to the train/test split (see Tables S6 and S7 within the Supplemental Material [33]).
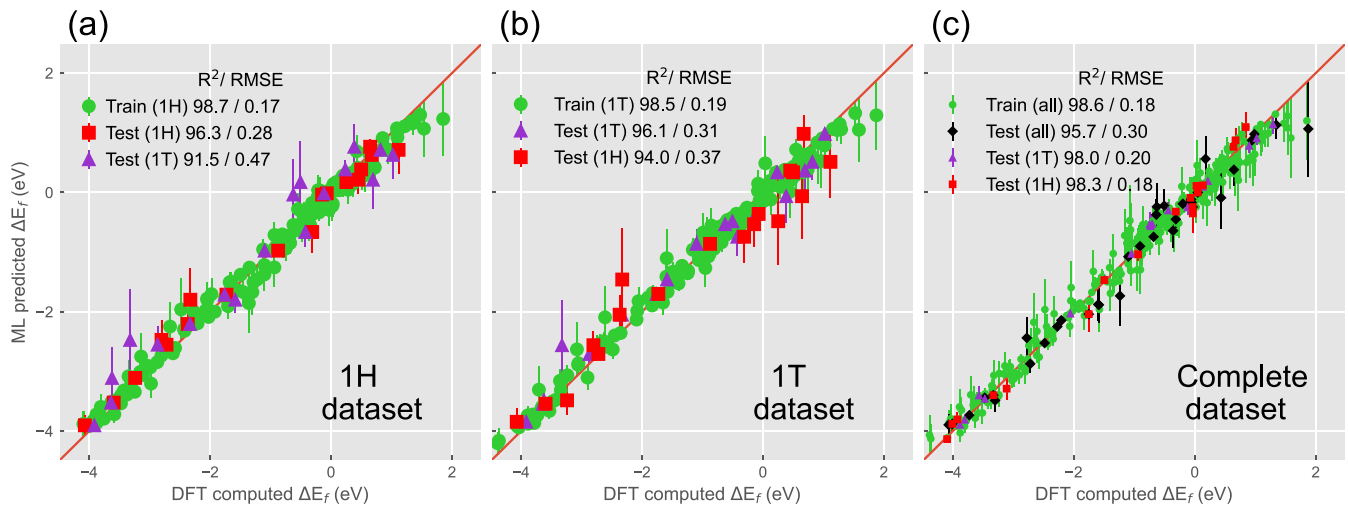
FIG. 5. Parity plots of formation energies of: (a) *1H*, (b) *1T* phases of TMD monolayers, and (c) the entire dataset. The DFT computed formation energies are along the abscissa and those predicted using RF along the ordinate. Each dataset is divided such that 90% of the data is used to train the RF regression model and 10% data is used for validation of the model. The error bars on each data point correspond to prediction errors.

So far, the ML model focused on predicting formation energies of TMDs in either the *1H* or *1T* phase. We further checked the prediction performance of our model by applying a more stringent test. We considered the entire dataset of 1080 candidate structures and augmented the feature set by four additional feature vectors (one feature vector each to indicate FM, AFM, or NM ground state, and one feature vector corresponding to the phase). We then trained the RF regression model with the set of hyperparameters that were optimized using the *1H* dataset alone. The calculated R^2-score and RM-SEs are shown in Fig. 5(c). The prediction performance of the enhanced ML model across a different dataset suggests that it can be further generalized to other TMDs that preferentially form configurations different from either the *1H* or *1T* phases.

In order to check the validity of our results over different algorithms, we also applied KRR to the *1H* dataset. The key hyperparameters of the KRR model, namely, α and γ were obtained through iterative minimization of the errors in the prediction. In our case, we calculated these hyperparameters by training the KRR model over 20 randomly distributed train/test splits of data in the ratio of 90%/10% and minimizing the root mean squared error (RMSE) of the model. In order to avoid overfitting in the model, we used a 10-fold cross-validation. Using this scheme, we calculated the optimized values of α and γ to be 0.0001 and 0.0001, respectively. We then separately trained KRR model on the *1H* and *1T* datasets using the optimized hyperparameters of KRR model. For the *1H* dataset, we calculated the average R^2-score and RMSE over 20 random train/test data splits to be 98.7%/95.3% and 0.17/0.30 eV, respectively. The average R^2-score and RMSE for the *1T* phase were 99.1%/96.3% and 0.14/0.28 eV, respectively. The parity plots of these trained datasets using KRR algorithm are shown in Fig. S5 within the Supplemental Material [33]. From Figs. 5 and S5, we can see that the corresponding accuracies of the RF regression and the KRR models are comparable, proving that our down-selected features are robust across different algorithms and describe

the material properties of TMDs with a fairly high degree of accuracy.

2. Random forest classification algorithm for predicting phase preference

So far, we have shown that we have built an efficient and accurate machine learning model to predict the formation energies of different TMDs using six descriptors. We next tested our model to see if it can predict the phase preferences of TMDs [see Fig. 2]. Starting with the down-selected features and the hyperparameters from the RF regression model discussed above, we augmented the feature set by the addition of one categorical feature to indicate the preferred phases of the TMDs. We then trained the RF classification model and used it to predict the phase preferences of the TMDs, classifying them as *1H* and *1T* phase crystals. In order to avoid overfitting with respect to the randomly selected data, and to ensure that our results were converged, we trained the model over different numbers of random train/test splits of data, predicting the unseen data each time. We found that the correctly predicted phases do not change if we use more than 55 different random train/test splits of data. This analysis yielded a high training and test score of 98.6% and 91.8%, respectively. The preferred phases predicted for the 180 down-selected traditional and Janus structures are shown in Fig. 6. We were able to accurately predict the phases of 172 monolayers. We note that for a small fraction of materials, the RF classification model did not predict the phase preference accurately; these are marked with an "X" in the figure. A comparison of Figs. 2 and 6 shows that some of the TMDs for which our model could not predict the phases accurately are the chemically unstable ones (positive formation energies). However, some of the chemically stable TMDs are also predicted inaccurately (e.g., RuSe$_2$, RuTe$_2$, and LaSeTe). There may be several reasons for these inaccuracies. Some of these systems either have a relatively smaller differences in energies between the *1H* and *1T* phases, or they
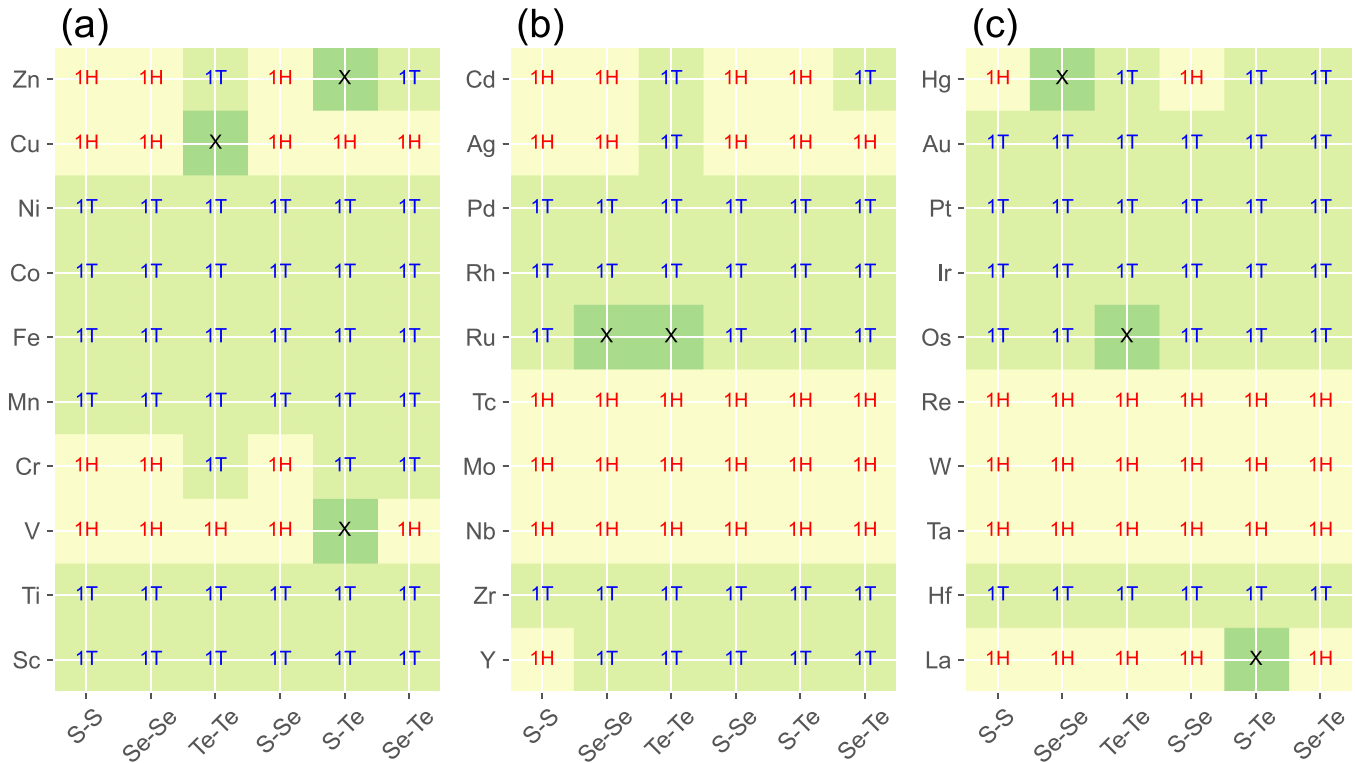
FIG. 6. Phase-preference prediction performance of RF classification algorithm for the 180 down selected TMDs (see Fig. 2). The boxes marked with an "X" correspond to systems that were not predicted accurately by our model. We consider our prediction to be accurate if we obtain the correct phase at least 90% of the time.

may prefer other distorted phases (*1T'* or ZT), which were not considered in this study.

It is interesting to note that the same set of down-selected compound features account for both the chemical stability and the relative phase stability of the TMDs. This implies that the RF regression algorithm itself can be used to predict the correct phase ordering by determining the differences between predicted formation energies for *1H*- and *1T*-phase TMDs. Figure S6 within the Supplemental Material [33] shows that the RF regression model indeed predicts the formation energies for the two phases at the level of accuracy needed to predict phase preferences for the TMDs, correctly predicting the crystal structures of 169 out of 180 TMDs.

IV. CONCLUSIONS

Predicting phase preferences of TMDs from just the knowledge of the constituent atoms is important, but has remained a long-standing problem. Earlier works had used packing fraction considerations (starting with Pauling [32] for ionic crystals), an empirical measure consisting of ionic radius ratios and degree of ionicity (Gamble [3]), and/or electronic factors (*d*-electron counts by Kertesz *et al.* [4]) to rationalize the observed structures of the traditional TMDs. Our machine learning-based quantitative analysis not only rediscovers each of these different physicochemical factors, but also uncovers previously unknown relative attributes of the constituent atoms that govern chemical stability and crystal structures of TMDs. Establishing different connections between compound

attributes of the TMDs and the target property/properties enabled us to not only rationalize, but also to accurately predict phases for 172 out of 180 TMDs. The framework developed here can be generalized for other classes of materials, and provides a path forward in materials research.

[1] J. Wilson and A. Yoffe, Adv. Phys. **18**, 193 (1969).

[2] A. D. Yoffe, Annu. Rev. Mater. Sci. **3**, 147 (1973).

[3] F. Gamble, J. Solid State Chem. **9**, 358 (1974).

[4] M. Kertesz and R. Hoffmann, J. Am. Chem. Soc. **106**, 3453 (1984).

[5] K. S. Novoselov, D. Jiang, F. Schedin, T. J. Booth, V. V. Khotkevich, S. V. Morozov, and A. K. Geim, Proc. Natl. Acad. Sci. USA **102**, 10451 (2005).

[6] A. Ayari, E. Cobas, O. Ogundadegbe, and M. S. Fuhrer, J. Appl. Phys. **101**, 014507 (2007).

[7] H. Ramakrishna Matte, A. Gomathi, A. Manna, D. Late, R. Datta, S. Pati, and C. Rao, Angew. Chem., Int. Ed. **49**, 4059 (2010).

[8] J. N. Coleman, M. Lotya, A. Ø'Neill, S. D. Bergin, P. J. King, U. Khan, K. Young, A. Gaucher, S. De, R. J. Smith *et al.*, Science **331**, 568 (2011).

[9] K.-K. Liu, W. Zhang, Y.-H. Lee, Y.-C. Lin, M.-T. Chang, C.-Y. Su, C.-S. Chang, H. Li, Y. Shi, H. Zhang *et al.*, Nano Lett. **12**, 1538 (2012).

[10] M. Nakano, Y. Wang, Y. Kashiwabara, H. Matsuoka, and Y. Iwasa, Nano Lett. **17**, 5595 (2017).

[11] H. Kaur, S. Yadav, A. K. Srivastava, N. Singh, S. Rath, J. J. Schneider, O. P. Sinha, and R. Srivastava, Nano Res. **11**, 343 (2018).

[12] A. B. Puthirath, A. P. Balan, E. F. Oliveira, V. Sreepal, F. C. Robles Hernandez, G. Gao, N. Chakingal, L. M. Sassi, P. Thibeorchews, G. Costin *et al.*, J. Phys. Chem. C **125**, 18927 (2021).

[13] M. Chhowalla, H. S. Shin, G. Eda, L. J. Li, K. P. Loh, and H. Zhang, Nat. Chem. **5**, 263 (2013).

[14] X. Qian, J. Liu, L. Fu, and J. Li, Science **346**, 1344 (2014).

[15] D. Xiao, G.-B. Liu, W. Feng, X. Xu, and W. Yao, Phys. Rev. Lett. **108**, 196802 (2012).

[16] K. F. Mak, K. He, J. Shan, and T. F. Heinz, Nat. Nanotechnol. **7**, 494 (2012).

[17] N. M. Ravindra, W. Tang, and S. Rassay, *Transition Metal Dichalcogenides Properties and Applications*, edited by M. Pech-Canul and N. Ravindra (Springer International Publishing, Cham, 2019), pp. 333–396.

[18] Q. H. Wang, K. Kalantar-Zadeh, A. Kis, J. N. Coleman, and M. S. Strano, Nat. Nanotechnol. **7**, 699 (2012).

[19] Y. F. Lin, Y. Xu, S. T. Wang, S. L. Li, M. Yamamoto, A. Aparecido-Ferreira, W. Li, H. Sun, S. Nakaharai, W. B. Jian *et al.*, Adv. Mater. **26**, 3263 (2014).

[20] H. Hayat, K. Kohary, and C. D. Wright, Nanotechnology **28**, 035202 (2017).

[21] K.-A. N. Duerloo, Y. Li, and E. J. Reed, Nat. Commun. **5**, 4214 (2014).

[22] Y. Li, K. A. N. Duerloo, K. Wauson, and E. J. Reed, Nat. Commun. **7**, 10671 (2016).

[23] H. Yang, S. W. Kim, M. Chhowalla, and Y. H. Lee, Nat. Phys. **13**, 931 (2017).

[24] D. Rhodes, D. A. Chenet, B. E. Janicek, C. Nyby, Y. Lin, W. Jin, D. Edelberg, E. Mannebach, N. Finney, A. Antony *et al.*, Nano Lett. **17**, 1616 (2017).

[25] Y. Wang, J. Xiao, H. Zhu, Y. Li, Y. Alsaid, K. Y. Fong, Y. Zhou, S. Wang, W. Shi, Y. Wang *et al.*, Nature (London) **550**, 487 (2017).

[26] C. Palacios-Berraquero, D. M. Kara, A. R.-P. Montblanch, M. Barbone, P. Latawiec, D. Yoon, A. K. Ott, M. Loncar,

A. C. Ferrari, and M. Atatüre, Nat. Commun. **8**, 15093 (2017).

[27] P. Manchanda, P. Kumar, and P. Dev, Phys. Rev. B **101**, 144104 (2020).

[28] P. Manchanda, P. Kumar, and P. Dev, Phys. Rev. B **103**, 144403 (2021).

[29] A. Y. Lu, H. Zhu, J. Xiao, C. P. Chuu, Y. Han, M. H. Chiu, C. C. Cheng, C. W. Yang, K. H. Wei, Y. Yang *et al.*, Nat. Nanotechnol. **12**, 744 (2017).

[30] J. Zhang, S. Jia, I. Kholmanov, L. Dong, D. Er, W. Chen, H. Guo, Z. Jin, V. B. Shenoy, L. Shi *et al.*, ACS Nano **11**, 8192 (2017).

[31] C. Zhang, Y. Nie, S. Sanvito, and A. Du, Nano Lett. **19**, 1366 (2019).

[32] L. Pauling, J. Am. Chem. Soc. **51**, 1010 (1929).

[33] See Supplemental Material at http://link.aps.org/supplemental/10.1103/PhysRevMaterials.6.094007 for (i) a note on efficacy of combining high-throughput quantum mechanical computations with machine learning algorithms, (ii) additional RF regression and classification results for different feature sets, as well as KRR results, and (iii) tables with details from high-throughput DFT calculations and ML analysis, including different robustness tests for the ML models.

[34] T. D. Huan, A. Mannodi-Kanakkithodi, C. Kim, V. Sharma, G. Pilania, and R. Ramprasad, Sci. Data **3**, 160012 (2016).

[35] S. A. Tawfik, O. Isayev, C. Stampfl, J. Shapter, D. A. Winkler, and M. J. Ford, Adv. Theory Simul. **2**, 1800128 (2019).

[36] V. Sharma, P. Kumar, P. Dev, and G. Pilania, J. Appl. Phys. **128**, 034902 (2020).

[37] S. Masubuchi, E. Watanabe, Y. Seo, S. Okazaki, T. Sasagawa, K. Watanabe, T. Taniguchi, and T. Machida, npj 2D Mater. Appl. **4**, 3 (2020).

[38] G. R. Schleder, C. M. Acosta, and A. Fazzio, ACS Appl. Mater. Interfaces **12**, 20149 (2020).

[39] K. Choudhary, K. F. Garrity, V. Sharma, A. J. Biacchi, A. R. Hight Walker, and F. Tavazza, npj Comput. Mater. **6**, 64 (2020).

[40] K. Choudhary, K. F. Garrity, A. C. E. Reid, B. DeCost, A. J. Biacchi, A. R. Hight Walker, Z. Trautt, J. Hattrick-Simpers, A. Gilad Kusne, A. Centrone *et al.*, npj Comput. Mater. **6**, 173 (2020).

[41] G. Kresse and J. Hafner, Phys. Rev. B **49**, 14251 (1994).

[42] G. Kresse and J. Furthmüller, Comput. Mater. Sci. **6**, 15 (1996).

[43] J. P. Perdew, K. Burke, and M. Ernzerhof, Phys. Rev. Lett. **77**, 3865 (1996).

[44] P. Dev, Y. Xue, and P. Zhang, Phys. Rev. Lett. **100**, 117204 (2008).

[45] P. Dev and P. Zhang, Phys. Rev. B **81**, 085207 (2010).

[46] H. J. Monkhorst and J. D. Pack, Phys. Rev. B **13**, 5188 (1976).

[47] R. D. Shannon, Acta Crystallogr. Sect. A **32**, 751 (1976).

[48] L. M. Mentel, mendeleev – A python resource for properties of chemical elements, ions and isotopes, https://github.com/lmmentel/mendeleev (2014).

[49] L. Breiman, Mach. Learn. **45**, 5 (2001).

[50] C. Zhang and Y. Ma, *Ensemble Machine Learning* (Springer, Boston, MA, 2012).

[51] C. Ataca, H. Şahin, and S. Ciraci, J. Phys. Chem. C **116**, 8983 (2012).

[52] F. A. Rasmussen and K. S. Thygesen, J. Phys. Chem. C **119**, 13169 (2015).

[53] D. Gao, Q. Xue, X. Mao, W. Wang, Q. Xu, and D. Xue, J. Mater. Chem. C **1**, 5909 (2013).

[54] J. Zhou, J. Lin, X. Huang, Y. Zhou, Y. Chen, J. Xia, H. Wang, Y. Xie, H. Yu, J. Lei *et al.*, Nature (London) **556**, 355 (2018).

[55] Y. Lu and S. B. Sinnott, ACS Appl. Nano Mater. **3**, 384 (2020).

[56] D. J. O'Hara, T. Zhu, A. H. Trout, A. S. Ahmed, Y. K. Luo, C. H. Lee, M. R. Brenner, S. Rajan, J. A. Gupta, D. W. McComb *et al.*, Nano Lett. **18**, 3125 (2018).

[57] T. Kanazawa, T. Amemiya, A. Ishikawa, V. Upadhyaya, K. Tsuruta, T. Tanaka, and Y. Miyamoto, Sci. Rep. **6**, 22277 (2016).

[58] N. Mounet, M. Gibertini, P. Schwaller, D. Campi, A. Merkys, A. Marrazzo, T. Sohier, I. E. Castelli, A. Cepellotti, G. Pizzi *et al.*, Nat. Nanotechnol. **13**, 246 (2018).

[59] S. K. C, C. Zhang, S. Hong, R. M. Wallace, and K. Cho, 2D Mater. **2**, 035019 (2015).

[60] H.-X. Zhong, S. Gao, J.-J. Shi, and L. Yang, Phys. Rev. B **92**, 115438 (2015).

[61] R. Huisman, R. de Jonge, C. Haas, and F. Jellinek, J. Solid State Chem. **3**, 56 (1971).