Identifying Domains of Applicability of Machine Learning Models for Materials Science

Christopher Sutton,[1,*] Mario Boley,[2,*,**] Luca M. Ghiringhelli,[1] Matthias Rupp,[1,3]
Jilles Vreeken,[4] Matthias Scheffler[1,5]

[1]*Theory Department*
Fritz Haber Institute of the Max Planck Society
Berlin, Germany

[2]*Faculty of Information Technology*
Monash University
Clayton, Australia

[3] *Citrine Informatics*

Redwood City, California, USA

[4]*CISPA Helmholtz Center for Information Security*
Saarbrücken, Germany

[5]*Physics Department and IRIS Adlershof,*
Humboldt-Universität, Berlin, Germany

[*]Corresponding authors: sutton@fhi-berlin.mpg.de; mario.boley@monash.edu

[*]These authors contributed equally to the work

[**]Part of this work has been conducted while the author was at the Max Planck Institute for Informatics

Given sufficient predictive accuracy, machine learning (ML) can accelerate the discovery of novel materials by allowing to rapidly screen compounds at orders of magnitude lower computational cost than first-principles electronic-structure approaches.[1-7] In practice, however, the accuracy of ML models is often insufficient to draw reliable conclusions about materials for specific applications.[7] Therefore, different ML representations for materials are actively developed to provide accurate predictions over diverse materials classes and properties.[8-20] A critical obstacle for this effort is that the complex choices involved in designing an ML model are currently made based on the overly simplistic metric of the average model test error with respect to the entire materials class. We show that this treatment of models as a black box that produces a single error statistic can render models as generally insufficient for certain screening tasks while they actually predict the target property accurately in specific sub-domains of the considered materials. For that, we present an informed diagnostic tool based on subgroup discovery (SGD)[21-23] that detects *domains of applicability* (DA) of ML models within a materials class. These domains are given as a combination of simple conditions on the unit cell structure (e.g., on the lattice vectors, lattice angles, and bond distances) under which the model error is substantially lower than its global average in the complete materials class. We demonstrate this procedure by discriminating the performance of several state-of-the-art ML models for predicting the formation energy of transparent conducting oxides – an important open problem in materials design for which a large data-analytics competition was recently hosted by Kaggle.[24] We analyze three state-of-the-art models that all combine kernel ridge regression with various representations including the winning model of the competition, adapted from natural language processing (*n*-gram method),[24] smooth overlap of atomic positions (SOAP),[13-14] and the many-body tensor representation (MBTR).[12] The accuracies of these models are practically indistinguishable when considering the average test error alone. Importantly, they all appear unsatisfactory for screening applications as they fail to reliably identify the ground state polymorph structure for many of the examined systems. However, when applying the proposed DA method, the models show notably distinct performances and different domains of applicability. That is, they all require different characteristics of the unit cell to perform well. Each of the models performs substantially better within their domain of

applicability than what is indicated by their undifferentiated average error over the whole domain. However, the MBTR-based model stands out with an almost 2-fold reduction in the average error and a 5-fold reduction in the fraction of errors above the required accuracy to identify the ground state polymorph (i.e., from 12.8 to 2.6 percent). Thus, we demonstrate that the MBTR-based model is in fact feasible for screening materials that lie within its domain of applicability. This illustrates how the proposed method can be used to guide the development of ML representations through the identification of their systematic strengths and weaknesses. We expect this form of analysis to advance ML methods for materials as well as ML methods for science more broadly.

To formally introduce the method for DA identification, we recall some notions of ML for materials. In order to apply smooth function approximation techniques like Ridge Regression, the materials of interest are represented as vectors in a vector space X according to some chosen *representation*. The more complex state-of-the-art representations evaluated in this work are defined further below. A first simple example is to use features $\{\varphi_1, \dots, \varphi_n\}$ of the isolated atoms that constitute the material (e.g., $\varphi_i(Z)$ may be the "electronegativity of the species with atomic number Z", see Table 1) and then to lift these to representation coordinates x_i for compounds $\{(Z_j, \mu_j)\}_{j=1}^{k}$ defined as

$$x_i = \sum_{j=1}^{k} \mu_j \varphi_i(Z_j) \qquad \text{Eq. 1}$$

where μ_j corresponds to the mixture coefficient for atomic number Z_j. Moreover, let y be a numeric material property according to which screening should be performed (in this work, we focus on formation energy, which is relevant for performing a ground state search). A predictive ML model is then a function $f: X \to \mathbb{R}$ aiming to minimize the *expected error* (also called *prediction risk*)

$$e(f) = \int_{X \times \mathbb{R}} l(f(\boldsymbol{x}), y) dP(\boldsymbol{x}, y) \qquad \text{Eq. 2}$$

measured by some non-negative *loss function* l that quantifies the cost incurred by predicting the actual property value y with $f(\boldsymbol{x})$. Examples for loss functions are the

squared error ($l(y', y) = (y' - y)^2$), the *absolute error* ($l(y', y) = |y' - y|$), and, for non-zero properties, the *relative error* ($l(y', y) = |y' - y|/|y|$). Here *P* denotes some fixed probability distribution that captures how candidate materials are *assumed to be* sampled from the material class (this concept, while commonly assumed in ML, is an unnecessary restriction for high-throughput screening as we discuss in more detail below). Since the true prediction risk is impossible to compute directly without perfect knowledge of the investigated materials class, models are evaluated by the *test error* (or *empirical risk*)

$$\hat{e}(f) = \sum_{i=1}^{m} e_i(f)/m \qquad \text{Eq. 3}$$

defined as the average of the *individual errors (losses)* $e_i(f) = l(f(\boldsymbol{x}_i), y_i)$ on some *test set* of *m* reference data points $(\boldsymbol{x}_i, y_i)_{i=1}^{m}$. The samples in this test set are drawn independently and identically distributed according to *P* and are also independent of the model – which means in practice that it is a random subset of all available reference data that has been withheld from the ML algorithm. In order to reduce the variance of this estimate, a common strategy is *cross-validation,* where this process is repeated multiple times based on partitioning the data into a number of non-overlapping "folds" and then to use each of these folds as test sets and the remaining data as a *training set* to fit the model.

This test error properly estimates the model performance globally over the whole representation space *X* (weighted by the distribution *P* used to generate the test points). This is an appropriate evaluation metric for selecting a model that is required to work well on average for arbitrary new input materials that are sampled according to the same distribution P. This is, however, not the condition of high-throughput screening. Here, rather than being presented with random inputs, we can decide which candidate materials to screen next. This observation leads to the central idea enabled by the domain of applicability analysis proposed in this work: if the employed model is particularly applicable in a specific sub-domain of the materials class, and if that sub-domain has a

simple and interpretable shape that permits to generate new materials from it, then we can directly focus the screening there.

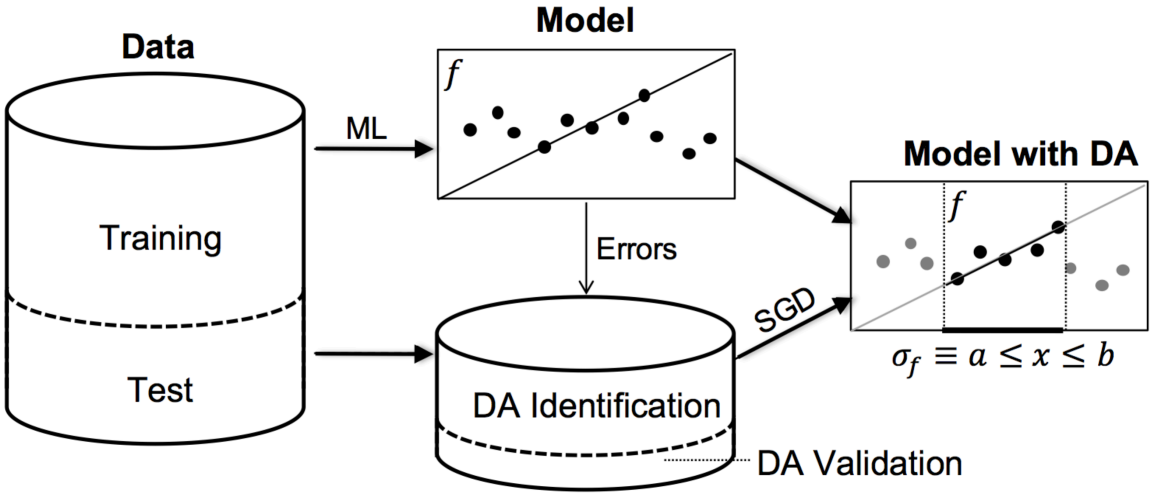

Figure 1. Workflow for the domain of applicability (DA) identification and validation for an ML model. The DA is described by a selector (σ_f) that is comprised of logical conjunctions of a representation space (here symbolized by a single-dimension x for simplicity but may be multidimensional). The selector is identified by applying subgroup discovery (SGD) to the individual ML-model errors for subset of test set (DA identification set). An unbiased estimate of the model performance within the DA is obtained on the remaining samples of the test set that were left out of the DA identification (DA validation set).

Such simply described domains of applicability (DA) can be identified by the descriptive data mining technique of SGD.[21-23, 25] This technique finds *selectors* in the form of logical conjunctions, i.e., Boolean functions ($\sigma: X \rightarrow \{\text{true,false}\}$) of the form:

$$\sigma(\boldsymbol{x}) \equiv \pi_1(\boldsymbol{x}) \wedge \pi_2(\boldsymbol{x}) \wedge \ldots \wedge \pi_p(\boldsymbol{x})$$

where "∧" denotes the "and"-operation and each proposition π_i is a simple inequality constraint on one of the coordinates, i.e., $\pi_i(\boldsymbol{x}) \equiv x_j \leq v$ for some constant v. Thus, these selectors describe intersections of axis-parallel half-spaces resulting in simple convex regions ($\{\boldsymbol{x} \in X: \sigma(\boldsymbol{x}) = \text{true}\}$) in X. This allows to systematically reason about the described sub-domains (e.g., it is easy to determine their differences and overlap) and also to sample novel points from them. To specifically obtain regions where a given model has a decreased error, SGD algorithms[26] can be configured to yield a selector with maximum *impact* on the model error. The impact is defined as the product of selector

coverage, i.e., the probability of the event $\sigma(\boldsymbol{x}) = \text{true}$, and the selector *effect* on the model error, i.e., the model error minus the model error given that the features satisfy the selector. Just as for the model fitting itself, we can only estimate these quantities based on empirical data. For that purpose, it is sensible to also split the test data into two parts: a *DA identification set* for optimizing the empirical impact and a *DA validation set* for obtaining an unbiased performance estimate of the identified DA (see Figure 1 for an illustration of the overall workflow). For ease of notation we assume the DA identification set consists of the first k points of the test set. We end up with the following objective function for the SGD algorithm:

$$\text{impact}(\sigma) = \underbrace{\left(\frac{s}{k}\right)}_{\text{coverage}} \underbrace{\left(\frac{1}{k}\sum_{i=1}^{k} l_i(f) - \frac{1}{s}\sum_{i \in I(\sigma)} l_i(f)\right)}_{\text{effect on test error}}$$

where s denotes the number of points in the DA identification set selected by σ and $I(\sigma) = \{i : 1 \leq i \leq k, \sigma(x_i) = \text{true}\}$ denotes the set of selected indices itself. In this work, we use the relative error as SGD target variable, which causes the applicability domain identification to be more sensitive to errors for small property values whereas it is more lenient for errors of large property values. This is a sensible behavior whenever we use the model to identify ground state structures.

The effect term of the objective function ensures that the model is estimated to be more accurate in the described region than in the global representation space. Thus, selectors with a large effect value describe domains of (increased[1]) applicability as desired. In addition, promoting large, i.e., *general*, DAs through the coverage term is important as those have a higher chance to a) contain data points of interest and b) to have an accurate effect estimate, i.e., the empirical error reduction measured by the effect term is likely to *generalize* to other points in the DA that are not contained in the DA identification set. Thus, the coverage term has a similar role as a regularization term in common objective

[1] The effect term captures a reduction in error relative to the global error. This calibrates the objective function, but it only guarantees that a positive objective value corresponds to an "increased" applicability as opposed to categorical applicability in terms of any absolute error constraint. Hence, the method detects the best DA possible out of all candidates, but in extreme cases no (notable) improvement over the global domain might be possible.

functions for model fitting. Technically, the data points withheld in the DA validation set mimic novel independent sample points that can be used to evaluate both: the coverage of the DA as well as the reduction in model error. As an extension of this, one can also repeat the domain of applicability optimization/validation on several splits (cross-validation) in order to reduce the variance of the coverage and model error estimates and, moreover, to assess the stability of the DA selector elements.

To illustrate the concept of applicability domains, let us consider a simple synthetic example (Figure 2) with a two-dimensional representation consisting of independent features x_1 and x_2 that are each distributed according to a normal distribution with mean 0 and variance 2 ($\mathrm{N}(0,2)$) and a target property y that is a 3rd degree polynomial in x_1 with an additive noise component that scales exponentially in x_2:

$$y \sim {x_1}^3 - x_1 + \mathrm{N}(0, \exp(x_2/2)).$$

That is, the y values are almost determined by the 3rd degree polynomial for low x_2 values but are almost completely random for high x_2 values. Discovering applicable domains reveals how different models cope differently with this setting even if they have a comparable average error. To show this, let us examine the error distributions obtained from three different kernelized regression models of the form

$$f(\cdot) = \sum_{i=1}^{n} \nu_i k(x_i^F, \cdot)$$

with parameter vector $\boldsymbol{\nu}$ that are fitted around a training (or fitting [F]) set $(\boldsymbol{x}_i^F, y_i^F)_{i=1}^n$ with three different choices for the kernel function k. We observe:

- When using the linear (lin) kernel ($k(x, x') = \langle x, x' \rangle$), the resulting linear model is globally incapable to trace the variation of the 3rd order polynomial except for a small stripe around the x_1-axis where it can be approximated well by a linear function. Consequently, there is a very high error globally that is substantially reduced in the applicability domain described by $\sigma_{\mathrm{lin}}(x_1, x_2) \equiv -0.3 \leq x_1 \leq 0.3$.
- When using the Gaussian kernel $k(x, x') = \exp(-\|x - x'\|^2/(2\varepsilon^2))$, the resulting radial basis function model is able to represent the target property well locally unless (a) the noise component is too large and (b) the variation of the target property is too high relative to the number of training points. The second

restriction is because the radial basis functions (rbf) have non-negligible values only within a small region around the training examples. Consequently, the discovered DA is not only restricted in x_2-direction but also excludes high absolute x_1-values: $\sigma_{\mathrm{rbf}} \equiv -3.3 \leq x_1 \leq 3.1 \ \wedge \ x_2 \leq 0.1$.

- In contrast, when using the non-local 3rd degree polynomial (ply) kernel $k(x, x') = (\langle x, x' \rangle + 1)^3$, data sparsity does not prevent an accurate modelling of the target property along the x_1 -axis. However, this non-locality is counterproductive along the x_2-axis where overfitting of the noise component has a global influence that results in higher prediction errors for the almost deterministic data points with low x_2-values. This is reflected in the identified applicability domain $\sigma_{\mathrm{ply}}(x_1, x_2) \equiv -3.5 \leq x_2 \leq 0.1$, which contains no restriction in x_1 -direction, but excludes both high and low x_2 -values. This highlights an important structural difference between the rbf and the polynomial model that is not reflected in their similar average errors.

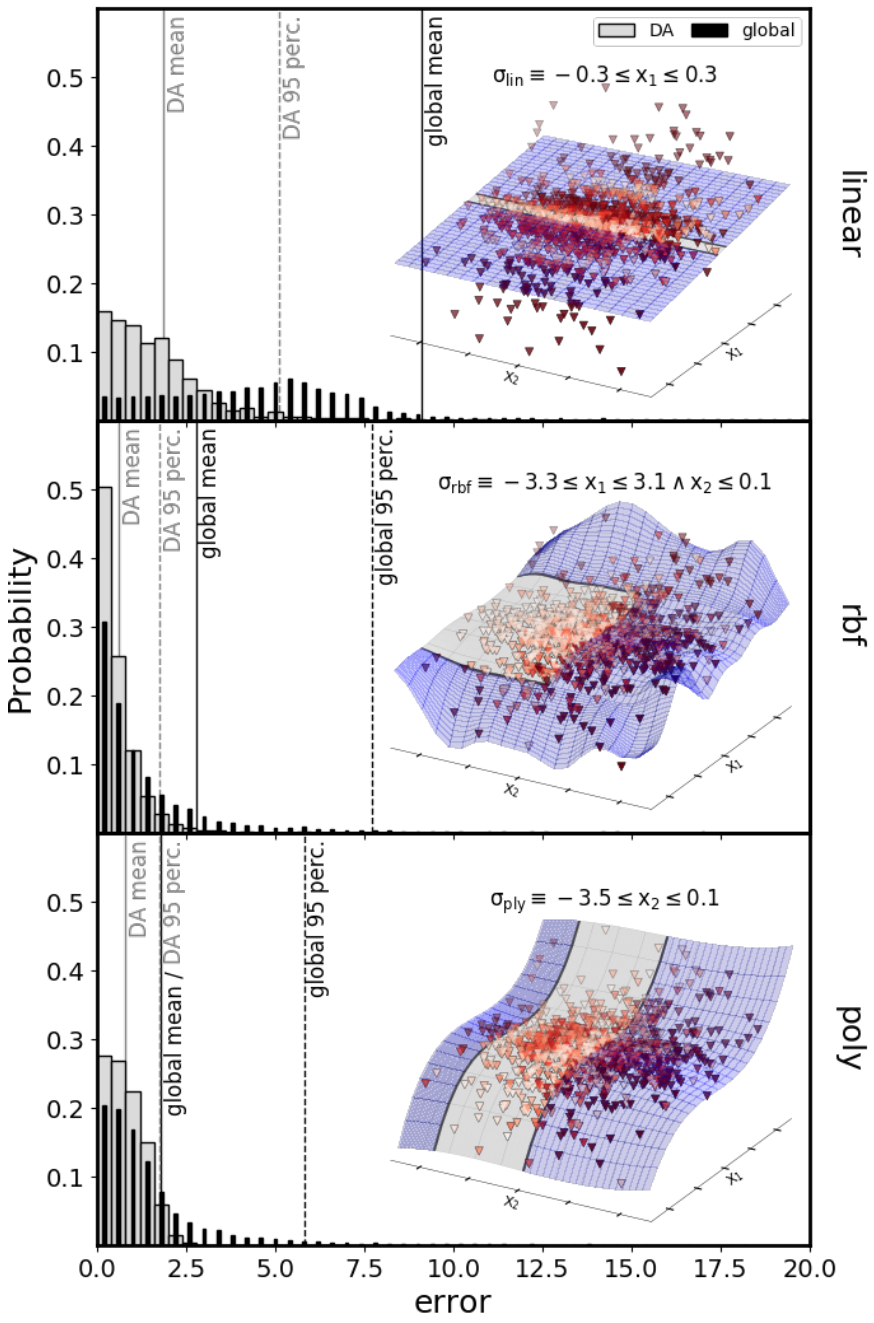

Figure 2. Domains of applicability (DA) and distributions of individual absolute errors for three different models approximating the same distribution of two independent features $x_1 \sim \mathrm{N}(0,2)$ and $x_2 \sim \mathrm{N}(0,2)$, and the target property $y \sim {x_1}^3 - x_1 + \mathrm{N}(0, \exp(x_2/2))$, where $\mathrm{N}(\mu, \varepsilon^2)$ denotes a normal distribution with mean μ and standard deviation s. Test points are plotted in 3d plots against the prediction surface of the models (color corresponds to absolute error) where the DA is highlighted in gray. The distributions of individual errors for the DA (gray) and globally (black) are shown. Note that the global error distribution of the linear model has a considerable long tail, which is capped in the image.

In the illustrative example above, all evaluated models share the same simple representation. However, in practice different models are typically fitted with different and more complicated representations. For instance, for the study on formation energies of transparent semiconductors below, we compare models based on the *n*-gram,[27] SOAP,[13-14] and MBTR[12] representations. These representations use different descriptions of the local atomic geometry, leading to high-dimensional non-linear transforms of the material configurations (e.g., 1400, 681, and 472 dimensions for MBTR, SOAP, and *n*-gram representations). A domain of applicability described directly in terms of these complex representations cannot easily be mapped back to intuitive conditions on the unit cell of a given material. This not only hinders interpreting the DA but also to construct novel materials from it. Finally, using different representations to describe applicability domains of different models makes it impossible to assess their overlap and differences. Therefore, we define a single neutral representation comprised of features that are specifically intended for the description of insightful sub-domains. A first natural group of features pertains directly to the shape of the unit cell such as the sorted lattice vectors and angles, the number of atoms in unit cell, and the unit-cell volume. (see Figure S2 for an illustration of the structural features of the unit cell). Additionally, when we are interested in a fixed compositional space, we can add features describing the composition (e.g., "percentage of Al cations") as well as structural features describing the bonding environments (e.g., "average nearest neighbor distance between Al and O"). See Table 1 for a summary of all features used.

Table 1: Features used for discovery of DA selectors.

Feature type	Feature label	Feature definition (units)
Unit cell	a, b, c	Lattice-vector lengths sorted from largest (a) to smallest (c) (Å)
	α	angle between b and c (°)
	β	angle between a and c (°)
	γ	angle between a and b (°)
	$\frac{V}{V_{atom}}$	volume of unit cell divided by atomic volumes derived from covalent radii
	N	number of atoms
Composition	%Al, %Ga, %In	number of cations divided by total number of cations
Computed bulk properties	E_g	PBE band gap energy
Structural	$R_{\{Al,Ga,In,O\}-\{Al,Ga,In,O\}}$	average nearest-neighbor distance between Al, Ga, In, and oxygen (Å)

Equipped with the concept of applicability domains, we can now examine the ML models for the prediction of stable alloys with potential application as transparent conducting oxides (TCOs). Materials that are both transparent to visible light and electrically conductive are important for a variety of technological devices such as photovoltaic cells, light-emitting diodes for flat-panel displays, transistors, sensors, touch screens, and lasers.[28-38] However, only a small number of TCOs have been realized because typically the properties that maximize transparency are detrimental to conductivity and vice versa. Because of their promise for technologically relevant applications, a public data-analytics competition was organized by the Novel Materials Discovery Centre of Excellence (NOMAD[39]) and hosted by the on-line platform Kaggle using a dataset of 3,000 $(Al_xGa_yIn_z)_2O_3$ sesquioxides, spanning six different spacegroups.[24] We emphasize that the

target property in the examination below is the formation energy, which is a measure of the energetic stability of the specific elements in a local environment that is defined by the specific lattice structure. Our aim is to demonstrate the ability of SGD to be used for differentiated model assessment to understand how well this physical picture is described by various representations that encode the local atomic information of each structure.

Before discussing the performance of the three ML models, we first briefly describe how the local atomic information is incorporated in each of the three representations. The MBTR representation space *X* can vary depending on the many-body order (e.g., interatomic distances for a two-body model, and/or angles for a two- and/or three-body model, and/or torsions for up to four-body models).[12] The results reported herein are calculated using a representation consisting of histograms of broadened pairwise inter-atomic distances, one for each unique pair of elements in the structure (i.e., for this dataset: Al-Al, Al-Ga, Al-In, Al-O, Ga-Ga, Ga-In, Ga-O, In-In, In-O, and O-O). These are generated according to:

$$g_i(r) = \sum_j \frac{1}{\sqrt{2\pi\varepsilon_{\text{atom}}^2}} \exp\left(-\left(r - r_{ij}\right)^2 / 2\varepsilon_{\text{atom}}^2\right) w_{\text{MBTR}}(i,j)$$

where a normal distribution function is centered at each distance between pairs of atoms (e.g., r_{ij}) to ensure smoothness of the representation. The function $w_{\text{MBTR}}(i,j)$ dampens contributions from atoms separated by large distances and is defined as: $w_{\text{MBTR}}(i,j) = \exp\left(-r_{ij}^2\eta\right)$, where both $\varepsilon_{\text{atom}}^2$ and η are hyperparameters.

The SOAP representation space is constructed by transforming pairwise atomic distances as overlapping densities of neighboring atoms and expanding the resulting density in terms of radial and spherical harmonics basis functions. The local density is modeled through a sum of normal distributions through each of the atomic neighbors *j* of atom *i*:

$$\rho_i(r) = \sum_j \frac{1}{\sqrt{2\pi\varepsilon_{\text{b}}^2}} \exp\left(-\left(r - r_{ij}\right)^2 / 2\varepsilon_{\text{atom}}^2\right) w_{\text{SOAP}}(r)$$

where *j* ranges over neighbors within a specific cutoff radius (r_{cut}) relative to i, where the cutoff function w_{SOAP} defined as:

$$w_{\mathrm{SOAP}}(r) = \begin{cases} \left[\cos\left(\frac{\pi r}{r_{cut}}\right) + 1\right]/2, & r \leq r_{\mathrm{cut}} \\ 0, & r > r_{\mathrm{cut}} \end{cases}$$

The density $\rho_i(r)$ is then expanded in terms of spherical harmonics $Y_{km}\left(\frac{r}{|r|}\right)$ and orthogonal radial functions $g_n(|r|)$:

$$\rho_i(r) = \sum_{nkm} \mathrm{c}_{nkm}\, g_n(|r|) Y_{km}\left(\frac{r}{|r|}\right).$$

The number of coefficients c_{nkm} is given by the choice of basis set expansion values. Rotationally invariant features are then computed from the coefficients of the expansion and averaged to create a single per-structure representation, forming the input space *X*. A real-space radial cutoff of r_{cut} = 10 Å and ε_{b} = 0.5 Å are used in this work.

The *n*-gram features are generated using a histogram of contiguous sequences of nodes (i.e., atoms) that are connected by edges (i.e., bonds) in a crystalline graph representation. An edge between nodes in the crystalline graph occurs if the interatomic distance in the 3D crystal is less than a pre-specified cut-off distance (r_{cut}) that is proportional to the sum of the ionic radii of the two species. The number of edges of a given node *i* corresponds to its coordination environment (CN_i):

$$CN_i = \begin{cases} 1 \text{ if } r_{ij} < r_{\mathrm{cut}} \\ 0 \end{cases}$$

r_{cut} was taken to be lattice dependent (details are provided in the supporting information). Here, only the cation coordination environment is considered, which is defined entirely by the number of oxygen atoms in the first coordination shell. The *n*-gram representation utilizes contiguous sequences of up to four nodes (see Ref. [24] for a detailed description of this approach).

As an additional benchmark, we also perform DA identification for a simple representation containing just atomic properties averaged by the compositions (this corresponds to the simplistic choice of a representation given in Eq. 1; see Table S1 for a list of atomic properties used in this representation). Since this representation is oblivious to configurational disorder (i.e., many distinct structures that are possible at a given

composition), it is expected to perform poorly across all space groups and concentrations. Formally, there is no unique y-value associated with each $\boldsymbol{x}$ but rather a distribution $P(y|\boldsymbol{x})$. Thus, even the optimal prediction at each composition of the test set (the median energy) to predict the test set energies results in a mean absolute error of 32.6 meV/cation, which is the highest accuracy that can be obtained using just composition-based properties. Therefore, it is a candidate for a representation that does not have any useful DA when compared to its full domain.

All representations are combined with kernel ridge regression using the rbf kernel. That is, ML models $f_{\boldsymbol{\nu}}(\boldsymbol{x}) = \sum_{i=1}^{n} \nu_i \exp\left(-\left\|\boldsymbol{x} - \boldsymbol{x}_i^F\right\|^2/(2\varepsilon^2)\right)$ with parameter vector $\boldsymbol{\nu}$ are found by minimizing the objective

$$\sum_{i=1}^{n} (f_{\boldsymbol{\nu}}(\boldsymbol{x}_i^F) - y_i^F)^2 + \lambda\|\boldsymbol{\alpha}\|$$

using a training set $(\boldsymbol{x}_i^F, y_i^F)_{i=1}^n$ of $n = 2400$ points. The values for the two hyperparameters ε and λ are determined through a grid search with 5-fold cross-validation. In addition to the training set, we have a hold-out test set $(\boldsymbol{x}_i, y_i)_{i=1}^m$ of $m = 600$ points. As described above, we partition the test set again into 6 folds of 100 points each such that we can evaluate the average DA performance over 6 different DA validation sets (in each case with the remaining 500 points of the test set as DA identification set). On top of that, we compare the identified DA selectors across the six individual experiments to assess their stability. SGD is performed with non-redundant branch-and-bound search with tight optimistic estimators and pre-discretization of cut-off values by 5-means clustering as described in Ref. [26].

MBTR, SOAP, and *n*-gram all display a similar test error (using the absolute error as the loss function l [see Eq. 3]; the resulting quantity we refer to as the *mean absolute error*, MAE) of 14.2 meV/cation, 13.6 meV/cation, and 15.0 meV/cation, respectively. This confirms previously reported virtually indistinguishable accuracies for MBTR and SOAP in the prediction of formation energies of alloys.[40] However, using the proposed method, key differences can be observed in the MAEs of their respective applicability domains. More specifically, the ML models built from MBTR, SOAP, and *n*-gram have an $\hat{e}$

averaged (standard deviation) over the six splits of the 100-sample DA validation set of 7.61 ($\pm$0.93) meV/cation, 11.24 ($\pm$2.87) meV/cation, 10.38 ($\pm$2.09) meV/cation, respectively. All identified DAs for the models utilizing MBTR, SOAP, and *n*-gram have a large coverage (i.e., percent of samples within in the DA) with an average (standard deviation) subpopulation contained within the DA validation set 0.44 ($\pm$0.03), 0.76 ($\pm$0.03), and 0.54 ($\pm$0.04), respectively.

In contrast, the atomic model is not only the worst model globally with a test error of 31.2 meV/cation, but, as anticipated, the DA error is virtually indistinguishable from the global model error (MAE = 29.9 meV/cation). This model performs slightly better than the MAE = 32.6 meV/cation that can be obtained by using the median energy at each composition of the test set to predict the test set energies. Therefore, this result illustrates the case of a weak representation for which no domain of applicability with substantial error reduction can be identified.

Although the reduction of the mean error for the three state-of-the-art representations is notable, the difference between the whole materials space and the DAs is even more pronounced when comparing the tails of the global error distributions using the 95 percentile. For the global models the average 95 percentile across all splits is reduced by a factor of 2.8, 1.3, and 1.5 for the DA compared with the global error for MBTR, SOAP, and *n*-gram (see Table S2 and Figure 3 for a summary of all model performances).

To put these error values into context, we consider the reference value of 24.9 meV/cation corresponding to half of the mean energy difference between the minimum energy and the second-to-minimum energy polymorph for all concentrations. The fraction of data points with these errors from the MBTR model above this reference value is reduced by a factor of 5 from 12.8% in the entire test set to 2.7% (averaged over each split) within the domain of applicability. A smaller reduction in the fraction of errors is observed for the SOAP model (13.8% in the entire test set to 9.6%) and *n*-gram model (16.7% vs. 11.5% in the global vs. test set). For the MBTR model, the 95-percentile of the DA errors (20.7 meV/cation) lies below the reference value.

The error and coverage estimates are not only consistent across the identification/validation splits but the *same selector* is identified across all of the splits. This is true in terms of both the referenced variables and the threshold values in the inequality constraints. The repeated selection of the same selector elements suggests that the identified variables describe some inherent structural strength/weaknesses of the investigated representations (note, however, that the exact numeric threshold value is also stabilized due to the clustering-based pre-discretization step performed by the SGD algorithm).

Interestingly, the variables that comprise the selectors of the domain of applicability are qualitatively different for each of these models. Selectors for MBTR include the number of atoms (N), the angle between the two longest lattice vectors in the unit cell (γ), and the average bond distance between Aluminum and Oxygen within the first coordination shell (that is defined by the effective coordination number), R_{Al-O}:

$$\sigma_{\text{MBTR}} \equiv\ N \geq 50 \text{ atoms} \wedge \gamma \leq 90.35° \wedge R_{Al-O}\ \leq 2.06 \text{ Å}\,.$$

For SOAP, selectors include features exclusively based on the unit cell shape such as the ratio of the longest (a) and shortest (c) lattice vectors, and lattice vector angles (β and γ):

$$\sigma_{\text{SOAP}} \equiv\ \frac{a}{c} \leq 3.89 \wedge \gamma < 90.35° \wedge\ \beta \geq 88.68°.$$

The selector of the *n*-gram model includes both features describing the unit cell shape [medium lattice vector (b) and angle (γ)] and structural motifs [interatomic bond distances between Al-O (R_{Al-O}) and Ga-O (R_{Ga-O}) within the first coordination shell]:

$$\sigma_{n\text{gram}} \equiv\ b \geq 5.59 \text{ Å} \wedge \gamma < 90.35° \wedge R_{Al-O}\ \leq 2.06 \text{ Å} \wedge R_{Ga-O}\ \leq 2.07 \text{ Å}.$$

It is worth noting that applying these DA selectors to the training set results in a similar reduction in error between the global and local populations and sample coverages (i.e., local population size) to what was observed for the test set: The training MAEs are reduced by factors of 1.67, 1.33 and 1.37 and the training DA coverages are 0.44, 0.76 and 0.54 for MBTR, SOAP, and *n*-gram models, respectively.

The qualitative differences observed in the selectors of the applicability domain for these three models can be quantified by examining the overlapping samples in the applicability domains using the *Jaccard similarity*, which is the ratio of the number of overlapping samples over the total number of samples in both DAs. We find Jaccard similarities of 0.60 for *n*-gram vs. SOAP, 0.67 for *n*-gram vs. MBTR, 0.58 for SOAP vs. MBTR (Figure S1). In other words, the discovered DA selectors are not only syntactically different, but, despite some overlap, they do indeed describe substantially different sub-populations of the investigated materials class.

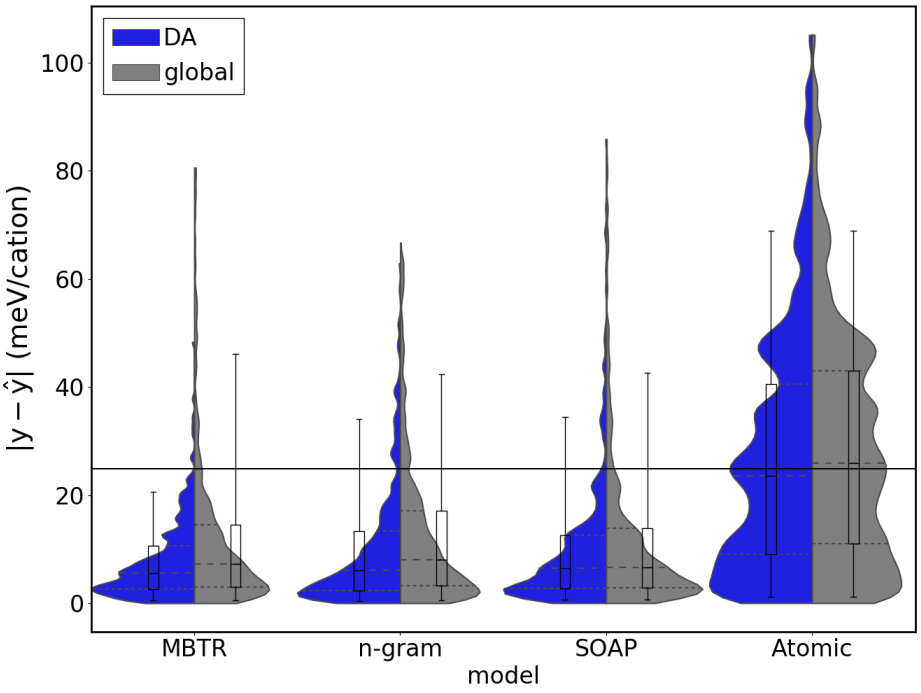

Figure 3. A comparison of the distribution of the absolute errors for the domain of applicability (DA) and the entire test set (global). Boxplots are included for each training and test set distribution to indicate the 25%, 50%, and 75% percentiles of the absolute errors. The violin plots only extend to the 98% percentile of the absolute errors, the box plots extend to the 95% percentile for the distribution contained in the violin plots. Horizontal line indicates reference error level of half of the mean energy difference between the minimum energy and the second-to-minimum energy polymorph (mean over all considered concentrations).

To further understand why the DAs of the three models are so different, we examine the distribution of each term of the selector for the SOAP representation because it has been used previously in several high-throughput screening applications. The inclusion of the attributes $\gamma < 90.35°$ and $\beta \geq 88.68°$ excludes 8.6% and 1.8% samples that have irregular unit cells based on the relatively large β and small γ values compared with the rest of the data points (see Figure 4 for the distribution of the three selectors). In contrast to these two selectors, $\frac{a}{c} \leq 3.89$ corresponds to a subgroup of 86% the test set samples. The inclusion of the $\frac{a}{c}$ is attributed to the fact that SOAP employs a real-space radial cutoff value $r_{cut} = 10$ Å in constructing the local atomic density for all samples (see above for a description of this representation). The algorithm threshold choice of $\frac{a}{c} \leq 3.89$ separates two modes of a relatively dense region of points (see Figure 4 top panel); However, for structures with asymmetric unit cell, the spherical radius could lead to inaccurate depiction of the local atomic environment, therefore, we repeat the procedure for two additional r_{cut} values of 20 Å and 30 Å. Compared with the selector identified for r_{cut} = 10 Å, a largely consistent selector is observed when the cut-off value is changed to a value of $r_{\text{cut}} = 20$ Å:

$$\sigma_{\text{SOAP},\, r_{\text{cut}} = 20\ \text{Å}} \equiv \frac{a}{c} \leq 3.89 \wedge \gamma \leq 90.35°.$$

However, increasing r_{cut} to a value of 30 Å – which exceeds the largest unit cell vector length (a) of 24 Å in the structures contained within this dataset – results in the selector:

$$\sigma_{\text{SOAP},\, r_{\text{cut}} = 30\ \text{Å}} \equiv c \geq 4.05\ \text{Å} \wedge \gamma \leq 90.35°.$$

The absence of the $\frac{a}{c}$ term for the SOAP representation utilizing a r_{cut} = 30 Å indicates that the choice of a cut off value less than the length of the unit cell directly impacts the model performance for the larger unit cells within this dataset, and thus, directly affects the selector chosen by SGD.

Finally, we note that it is an intuitive expectation that an improved model can be obtained by re-running the ML algorithm using only training data from within the discovered domain of applicability. However, this is not true in general: points outside of the DA, while having a higher error on average, can still contribute positively to the prediction inside the DA. For instance, refitting to a training set trimmed according to the DA selectors of the three model types investigated here leads to a change in test MAE of -1.5 (MBTR), -1.0 (SOAP), and +0.1 (*n*-gram) meV/cation. That is, we see an improvement for the MBTR and SOAP models when fitting to the reduced domain (with reduced training data), but a slight decline in model performance for the *n*-gram model. Note that, technically, only the DA validation set can be used to obtain an unbiased error estimate of a refitted model because this contains the only data that is independent of the refitted model. All other data, including the part of the test set that served as DA identification set, were involved in the overall process that yielded the refitted model. The statistical considerations related to model refitting are an interesting subject for future investigations and a better understanding could lead to an iterative refitting scheme where DAs are refined until convergence. Such a scheme could also contain an active learning component where additional data points are sampled from within the identified subdomains.

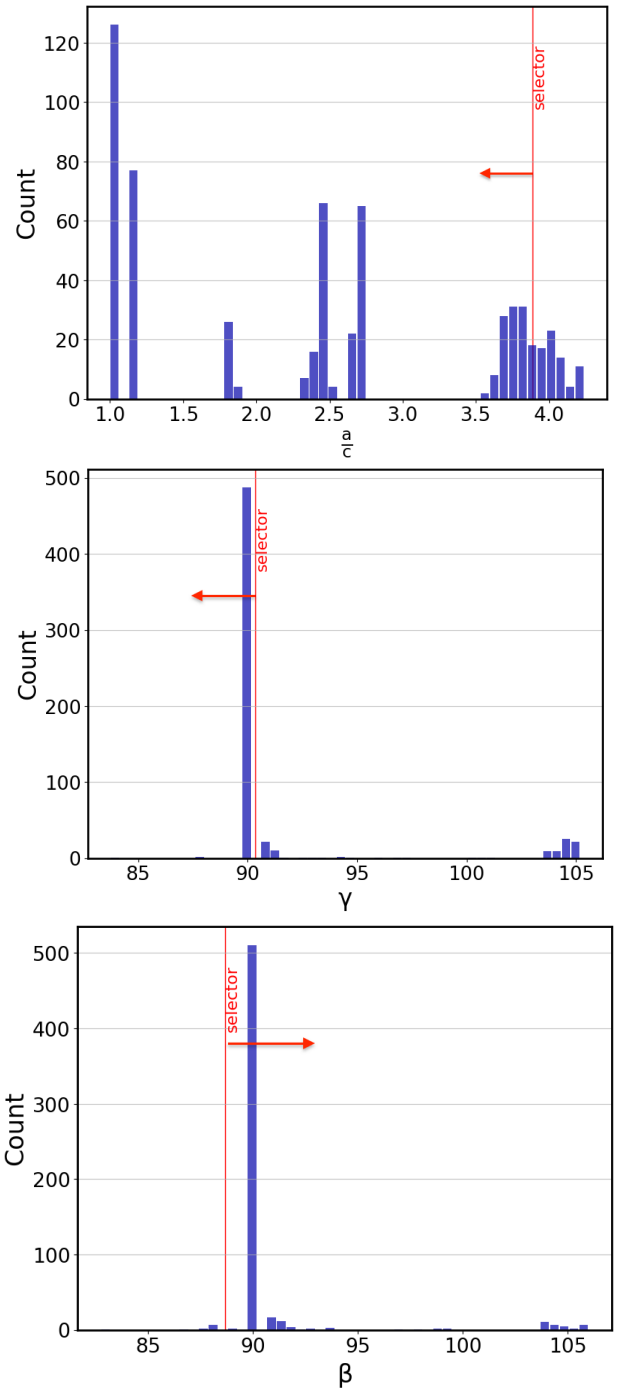

Figure 4. Distribution of features for the three selectors defining the domain of applicability of the SOAP-based model ($\sigma_{\mathrm{SOAP}} \equiv \frac{a}{c} \leq 3.89 \wedge \gamma < 90.35° \wedge \beta \geq 88.68°$).

CONCLUSION

We demonstrate a new approach to identify domains of applicability of machine learning models for materials properties, in which models achieve a substantially lower error than on the whole materials class. This approach is based on applying subgroup discovery to the individual errors of the model predictions in a model test set. Applying this idea to state-of-the art models of TCO formation energies (using kernel ridge regression combined with predictions from SOAP, *n*-gram, and MBTR) identifies distinct DAs for each model with notably improved accuracies and a large coverage of the underlying materials class (44% - 76%). In particular, the MBTR model displays a subdomain with a 95 percentile error that is about a factor of two smaller than its global 95 percentile error. Besides these quantitative assessments, the discovered DAs enable a qualitative comparison of the three investigated material representations by investigating their defining logical formulas. These DA selectors show notable differences that can be attributed to significant variation in the physics being captured by the models. For example, the appearance of a number of atoms in the selector for MBTR indicates heterogeneity in the error distribution based on the unit cell size because of the implementation of an unnormalized histogram in the representation. For SOAP, the selectors include features exclusively based on the unit cell shape, which is attributed to the choice of a cutoff radius in the construction of the local atomic environment. In order to be applicable on a wider domain, improved versions of these representations need to address those systematic shortcomings – a conclusion which is illustrative of how the method of DA identification can guide the improvement of materials representations and ML methods in general.

AWKNOWLEDGEMENTS

This work received funding from the European Union's Horizon 2020 Research and Innovation Programe (grant agreement No. 676580), the NOMAD laboratory CoE, and ERC:TEC1P (No. 740233). C.S. gratefully acknowledges funding by the Alexander von Humboldt Foundation.

REFERENCES

1. Meredig, B.; Agrawal, A.; Kirklin, S.; Saal, J. E.; Doak, J. W.; Thompson, A.; Zhang, K.; Choudhary, A.; Wolverton, C., Combinatorial screening for new materials in unconstrained composition space with machine learning. *Physical Review B* **2014,** *89* (9), 094104.
2. Isayev, O.; Fourches, D.; Muratov, E. N.; Oses, C.; Rasch, K.; Tropsha, A.; Curtarolo, S., Materials Cartography: Representing and Mining Materials Space Using Structural and Electronic Fingerprints. *Chemistry of Materials* **2015,** *27* (3), 735-743.
3. Oliynyk, A. O.; Antono, E.; Sparks, T. D.; Ghadbeigi, L.; Gaultois, M. W.; Meredig, B.; Mar, A., High-Throughput Machine-Learning-Driven Synthesis of Full-Heusler Compounds. *Chemistry of Materials* **2016,** *28* (20), 7324-7331.
4. Schmidt, J.; Shi, J.; Borlido, P.; Chen, L.; Botti, S.; Marques, M. A. L., Predicting the Thermodynamic Stability of Solids Combining Density Functional Theory and Machine Learning. *Chemistry of Materials* **2017,** *29* (12), 5090-5103.
5. Pilania, G.; Mannodi-Kanakkithodi, A.; Uberuaga, B. P.; Ramprasad, R.; Gubernatis, J. E.; Lookman, T., Machine learning bandgaps of double perovskites. **2016,** *6*, 19375.
6. Lee, J.; Seko, A.; Shitara, K.; Nakayama, K.; Tanaka, I., Prediction model of band gap for inorganic compounds by combination of density functional theory calculations and machine learning techniques. *Physical Review B* **2016,** *93* (11), 115104.
7. Draxl, C.; Scheffler, M., Big-Data-Driven Materials Science and its FAIR Data Infrastructure. . In *Handbook of Materials Modeling*, Andreoni, W.; Yip, S., Eds. Springer: 2019.
8. Rupp, M.; Tkatchenko, A.; Müller, K.-R.; von Lilienfeld, O. A., Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters* **2012,** *108* (5), 058301.
9. Grégoire, M.; Matthias, R.; Vivekanand, G.; Alvaro, V.-M.; Katja, H.; Alexandre, T.; Klaus-Robert, M.; Lilienfeld, O. A. v., Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics* **2013,** *15* (9), 095003.
10. Hirn, M.; Poilvert, N.; Mallat, S., Quantum energy regression using scattering transforms. https://arxiv.org/abs/1502.02077 **2015**.
11. Hansen, K.; Biegler, F.; Ramakrishnan, R.; Pronobis, W.; von Lilienfeld, O. A.; Müller, K.-R.; Tkatchenko, A., Machine Learning Predictions of Molecular Properties: Accurate Many-Body Potentials and Nonlocality in Chemical Space. *The Journal of Physical Chemistry Letters* **2015,** *6* (12), 2326-2331.
12. Huo, H.; Rupp, M., Unified Representation for Machine Learning of Molecules and Crystals. *arxiv.org* **2017**.
13. Bartók, A. P.; Payne, M. C.; Kondor, R.; Csányi, G., Gaussian Approximation Potentials: The Accuracy of Quantum Mechanics, without the Electrons. *Physical Review Letters* **2010,** *104* (13), 136403.
14. Bartók, A. P.; Kondor, R.; Csányi, G., On representing chemical environments. *Physical Review B* **2013,** *87* (18), 184115.
15. Seko, A.; Hayashi, H.; Nakayama, K.; Takahashi, A.; Tanaka, I., Representation of compounds for machine-learning prediction of physical properties. *Physical Review B* **2017,** *95* (14), 144110.

16. Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U., How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Physical Review B* **2014,** *89* (20), 205118.
17. Faber, F.; Lindmaa, A.; von Lilienfeld, O. A.; Armiento, R., Crystal structure representations for machine learning models of formation energies. *International Journal of Quantum Chemistry* **2015,** *115* (16), 1094-1101.
18. Behler, J., Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics* **2011,** *134* (7), 074106.
19. Behler, J., Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys* **2011,** *13*, 17930–17955.
20. Shapeev, A. V., Moment Tensor Potentials: A Class of Systematically Improvable Interatomic Potentials. *Multiscale Modeling & Simulation* **2016,** *14* (3), 1153-1173.
21. Atzmueller, M., Subgroup discovery. . *WIREs Data Mining Knowl Discov* **2015,** *5*, 35-49.
22. Wrobel, S., An algorithm for multi-relational discovery of subgroups. In *European Symposium on Principles of Data Mining and Knowledge Discovery*, Springe: Berlin, 1997.
23. Friedman, J. H.; Fisher, N. I., Bump hunting in high-dimensional data. *Statistics and Computing* **1999,** *9.2*, 123-143.
24. Breiman, L., Stacked regressions. *Machine Learning,* **1996,** *24* (1), 49-64.
25. Herrera, F.; Carmona, C. J.; González, P.; del Jesus, M. J., An overview on subgroup discovery: foundations and applications. *Knowledge and Information Systems* **2011,** *29* (3), 495-525.
26. Boley, M.; Goldsmith, B. R.; Ghiringhelli, L. M.; Vreeken, J., Identifying consistent statements about numerical data with dispersion-corrected subgroup discovery. *Data Mining and Knowledge Discovery* **2017,** *31* (5), 1391-1418.
27. Sutton, C., NOMAD 2018 Kaggle Competition: Solving Materials Science Challenges Through Crowd Sourcing. *Submitted* **2018**.
28. Kinoshita, A.; Hirayama, H.; Ainoya, M.; Aoyagi, Y.; Hirata, A., Room-temperature operation at 333 nm of Al0.03Ga0.97N/Al0.25Ga0.75N quantum-well light-emitting diodes with Mg-doped superlattice layers. *Applied Physics Letters* **2000,** *77* (2), 175-177.
29. Ohta, H.; Kawamura, K.-i.; Orita, M.; Hirano, M.; Sarukura, N.; Hosono, H., Current injection emission from a transparent p–n junction composed of p-SrCu2O2/n-ZnO. *Applied Physics Letters* **2000,** *77* (4), 475-477.
30. Tsukazaki, A.; Ohtomo, A.; Onuma, T.; Ohtani, M.; Makino, T.; Sumiya, M.; Ohtani, K.; Chichibu, S. F.; Fuke, S.; Segawa, Y.; Ohno, H.; Koinuma, H.; Kawasaki, M., Repeated temperature modulation epitaxy for p-type doping and light-emitting diode based on ZnO. *Nat Mater* **2005,** *4* (1), 42-46.
31. Nakamura, S.; Mukai, T.; Senoh, M., Candela‐class high‐brightness InGaN/AlGaN double‐heterostructure blue‐light‐emitting diodes. *Applied Physics Letters* **1994,** *64* (13), 1687-1689.
32. Arulkumaran, S.; Sakai, M.; Egawa, T.; Ishikawa, H.; Jimbo, T.; Shibata, T.; Asai, K.; Sumiya, S.; Kuraoka, Y.; Tanaka, M.; Oda, O., Improved dc characteristics of AlGaN/GaN high-electron-mobility transistors on AlN/sapphire templates. *Applied Physics Letters* **2002,** *81* (6), 1131-1133.

33. Kubovic, M.; Kasu, M.; Kallfass, I.; Neuburger, M.; Aleksov, A.; Koley, G.; Spencer, M.; Kohn, E., Microwave performance evaluation of diamond surface channel FETs. *Diamond and related materials* **2004,** *13* (4), 802-807.
34. Hoffman, R.; Norris, B. J.; Wager, J., ZnO-based transparent thin-film transistors. *Applied Physics Letters* **2003,** *82* (5), 733-735.
35. Nishii, J.; Hossain, F. M.; Takagi, S.; Aita, T.; Saikusa, K.; Ohmaki, Y.; Ohkubo, I.; Kishimoto, S.; Ohtomo, A.; Fukumura, T., High mobility thin film transistors with transparent ZnO channels. *Japanese journal of applied physics* **2003,** *42* (4A), L347.
36. Nomura, K.; Ohta, H.; Ueda, K.; Kamiya, T.; Hirano, M.; Hosono, H., Thin-film transistor fabricated in single-crystalline transparent oxide semiconductor. *Science* **2003,** *300* (5623), 1269-1272.
37. Nomura, K.; Ohta, H.; Takagi, A.; Kamiya, T.; Hirano, M.; Hosono, H., Room-temperature fabrication of transparent flexible thin-film transistors using amorphous oxide semiconductors. *Nature* **2004,** *432* (7016), 488-492.
38. Dehuff, N.; Kettenring, E.; Hong, D.; Chiang, H.; Wager, J.; Hoffman, R.; Park, C.-H.; Keszler, D., Transparent thin-film transistors with zinc indium oxide channel layer. *Journal of Applied Physics* **2005,** *97* (6), 064505.
39. Draxl, C.; Scheffler, M., NOMAD: the FAIR concept for big-data-driven materials science. *MRSBull* **2018,** *43*, 9.
40. Nyshadham, C.; Rupp, M.; Bekker, B.; Shapeev, A. V.; Mueller, T.; Rosenbrock, C. W.; Csányi, G.; Wingate, D. W.; Hart, G. L. W., Machine-learned multi-system surrogate models for materials prediction. *npj Computational Materials* **2019,** *5* (1), 51.