scientific reports

OPEN

A universal similarity based approach for predictive uncertainty quantification in materials science

Vadim Korolev[1]✉, Iurii Nevolin[2] & Pavel Protsenko[1]

Immense effort has been exerted in the materials informatics community towards enhancing the accuracy of machine learning (ML) models; however, the uncertainty quantification (UQ) of state-of-the-art algorithms also demands further development. Most prominent UQ methods are model-specific or are related to the ensembles of models; therefore, there is a need to develop a universal technique that can be readily applied to a single model from a diverse set of ML algorithms. In this study, we suggest a new UQ measure known as the Δ-metric to address this issue. The presented quantitative criterion was inspired by the *k*-nearest neighbor approach adopted for applicability domain estimation in chemoinformatics. It surpasses several UQ methods in accurately ranking the predictive errors and could be considered a low-cost option for a more advanced deep ensemble strategy. We also evaluated the performance of the presented UQ measure on various classes of materials, ML algorithms, and types of input features, thus demonstrating its universality.

Supervised machine learning (ML) has tremendously transformed the modeling of structure–property relationships[1–3]. Cutting-edge studies have addressed the implementation of new featurization schemes, such as materials representations[4–7] and the adaptation of neural network architectures to domain-specific input data (crystal structures[8–10] and chemical compositions[11–13]). In parallel, the materials informatics community has created a diverse suite of user-friendly packages covering different stages of the ML pipeline[14–21]. A sharp increase in the number of predictive algorithms, materials representations, and applications has driven the development of benchmarking protocols and data sets[22–24]. However, another aspect of using ML, namely, uncertainty quantification (UQ)[25,26], has received much less attention, although it has also been crucial to the exploration of subdomains of materials spaces that significantly differ from the training set. Thus, advanced materials informatics tasks, such as materials discovery, have been shown to be most sensitive to this issue[27–30].

Most studies have focused on the applicability of predictive models and reliability of individual outputs (prediction intervals) to perform UQ using universal, but cost-prohibitive, variational-inference-based methods, such as Monte Carlo dropout[31,32], bootstrapping/subsampling[29,31,33,34], and deep ensembles[31]. Gaussian process regression (GPR)[35–37] is an alternative approach that intrinsically provides the output and associated uncertainty. Beyond these well-known straightforward solutions, only a few techniques have been adopted for materials informatics. In particular, Janet et al.[38] introduced the distance to training data in the latent space of a neural network as a low-cost UQ metric, and predictive error decreased systematically by tightening the threshold of this parameter. In addition, Sutton et al.[39] presented a tool based on subgroup discovery. The conditioned combinations of structural and compositional features derived by the method defined the subdomains of a materials space with a model error that was lower than average for all considered materials.

A closely related field to materials informatics, chemoinformatics, has also been confronted with the problem of UQ. For example, three main groups of algorithms have been adapted for ML-assisted drug design[40]. First, frequentist methods drew the statistical inference only from the likelihood without a prior hypothesis, and the UQ for the molecular property prediction tasks was provided[41–44]. Second, Bayesian approaches have been successfully utilized for the same purpose[45–50]. The third group of methods, known as empirical techniques, have relied on the concept of applicability domain[51–55]. Generally, this can be expressed as a chemical structure (descriptor) subspace where the predictive model provides a reliable output. Thus, UQ methods involving applicability domain estimation do not consider information from the approximation algorithm. Consequently, the universality of model-agnostic methods interacts with the simple assumption that their performance is only determined by the distribution of training points in chemical space.

[1]Department of Chemistry, Lomonosov Moscow State University, Moscow 119991, Russia. [2]Frumkin Institute of Physical Chemistry and Electrochemistry, Russian Academy of Sciences, Moscow 119071, Russia. ✉email: korolewadim@gmail.com

In this study, we have presented a new UQ measure, Δ-metric, based on ideas of applicability domain estimation originated from chemoinformatics, and the remainder of this manuscript describes the principles of its construction and benchmarking results. We considered four use cases of bandgap prediction to assess the efficacy of the Δ-metric in ranking predictive errors and calculating predictive intervals. The performance of the suggested metric in UQ was compared to the performance of widespread methods, including deep ensembles, subsampling, and the infinitesimal jackknife variance. In addition, two out-of-domain use cases were also discussed. The provided results depicted a set of scenarios where the Δ-metric would be a potentially helpful UQ technique.

Methods

Data sets.

For this study, 10,434 inorganic crystal structures and their corresponding bandgap values were obtained from the computational database presented by Kim et al.[56] Initially, the band edges were identified within the generalized gradient approximation, as implemented by Perdew, Burke, and Ernzerhof (PBE)[57]. Then, one-shot calculations with the screened hybrid functional of Heyd, Scuseria, and Ernzerhof (HSE06)[58] were conducted to estimate the bandgap values at the k points of the band edges found with the PBE functional. Structures containing noble gases were excluded from consideration. Three hundred fifty-eight two-dimensional materials and their corresponding bandgap values were collected from the Computational 2D Materials Database[59–61]. We considered structures with bandgaps that were calculated within three available levels of theory: PBE, hybrid HSE06, and many-body GW approximation. A total of 14,204 MOF structures and their corresponding PBE bandgap values were obtained from the Quantum MOF database[62], and 12,500 molecular crystals and their corresponding PBE bandgap values were obtained from the Organic Materials Database[63,64].

Feature extraction and preprocessing.

For the MEGNet[9] model, we used elemental embeddings from the original study trained on the formation energy, which was kept fixed during our model training. The CFIDs[65], as implemented in the matminer[17] package, were used to featurize the two-dimensional materials. The thirty CFIDs with the highest F-values were selected for further consideration, and the PBE bandgap was incorporated as a crude estimator of the GW bandgap[66–68]. The StandardScaler was applied to normalize the above features. The attributes proposed by Meredig et al.[69] (atomic fractions of elements and statistics of elemental properties), as implemented in the matminer[17] package, were used to featurize the MOFs, and the MinMaxScaler was implemented to normalize these features. We used the PLMFs proposed by Isayev et al.[7] to featurize the molecular organic materials, and specifically, the linear chains up to four atoms and the first shell of the nearest-neighbor atoms were considered. Only atomic numbers were taken into account to label local fragments. Atomic connectivity was defined according to a distance criterion with a tolerance of 0.25 Å. In contrast to the original study, the Voronoi–Dirichlet partition was not applied for reasons of simplicity. The PLMFs normalized per Å^3 were processed by VarianceThreshold, where features with a training-set variance greater than 10^{-7} were chosen for the following consideration. The selected features were normalized using the MinMaxScaler. All the data were prepared within the scikit-learn[70] processing routines.

Machine learning and uncertainty quantification models.

The GPR, KRR, and RF models implemented in the scikit-learn[70] library were trained on 80% of the data and tested with the remaining 20%. A train-validation-test split of 80%–10%–10% was applied for the MEGNet[9] model. For the GPR model, we used a sum-kernel including the radial-basis function and rational quadratic kernels, where the noise level was set to 1.0 via the α parameter. The KRR model was trained using the Laplacian kernel, with a regularization strength α of 0.1, and parameter γ of 0.1. The RF model was trained using 1000 trees, and all other hyperparameter values of the above models were set to the default. The MEGNet model was trained using the hyperparameter values proposed in the original study.

The SOAP-like descriptor for the Δ-metric was constructed using the DScribe[18] library. The number of radial basis functions n_{max} and maximum degree of spherical harmonics l_{max} were set to eight and six, respectively, while the degree ζ in Eq. (3) was set to four. Quantile regression analysis was performed using the statsmodels[71] package.

Results and discussion

Definition of the UQ metric.

The presented UQ measure, Δ-metric, was deeply inspired by the k-nearest neighbor approach adopted for applicability domain evaluation. Specifically, the average distance to the k closest training set points was compared to the pre-defined threshold in the most widespread formulation[72]. According to the original definition of the weighted k-nearest neighbor algorithm, we proposed the following formula for the i-th structure in the test set:

$$\Delta_i = \frac{\sum_j K_{ij} |\varepsilon_j|}{\sum_j K_{ij}}, \tag{1}$$

where ε_j is the error between the true and predicted target values of the j-th neighbor structure in the training set, and K_{ij} is the corresponding weight coefficient. It was natural to represent K_{ij} as a similarity measure between the i-th and j-th structures. For this purpose, we implemented a kernel proposed by Bartok et al.[4], which used the form of a normalized dot product raised to the ζ-th power:

$$K_{ij} = \left(\frac{p_i p_j}{p_i p_j} \right)^{\zeta}, \tag{2}$$

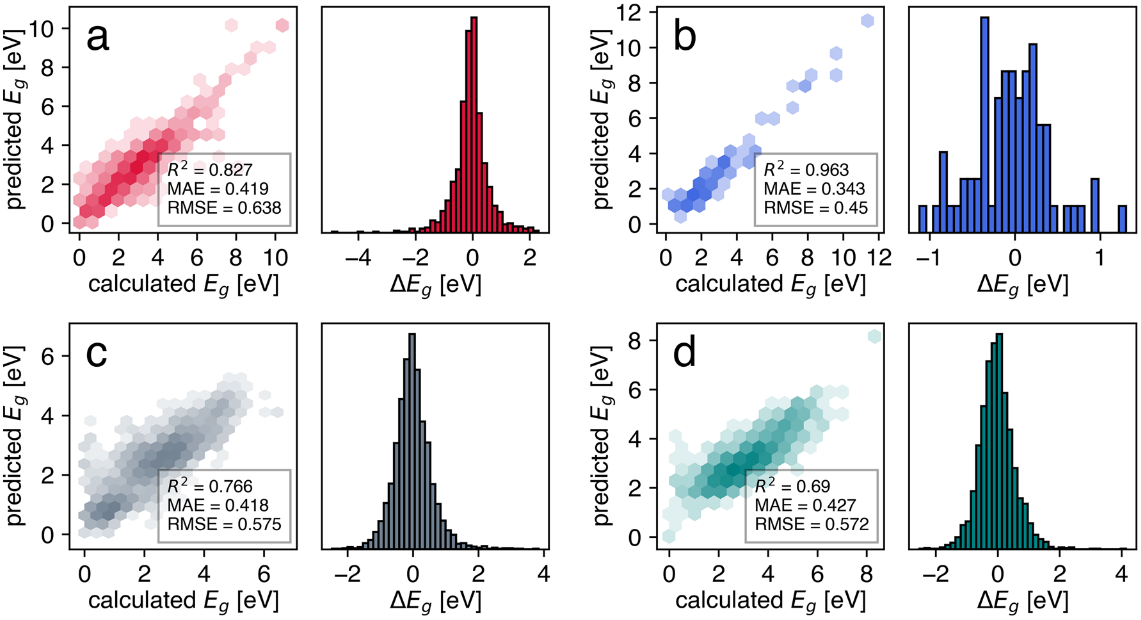

Figure 1. Parity plots and error histograms for the four considered use cases: (**a**) SNUMAT-MEGNet, (**b**) C2DB-GPR-CFID, (**c**) QMOF-KRR-ElemStat, and (**d**) OMDB-RF-PLMF.

where p is a global descriptor. To featurize the structures, we used a smooth overlap of the atomic positions (SOAP) descriptor[73], as given by

$$p_{n_1 n_2 l} = \frac{\pi}{N^2} \sqrt{\frac{8}{2l+1}} \sum_m \sum_{i,j} \left(c^i_{n_1 lm}\right)^\dagger c^j_{n_2 lm}, \tag{3}$$

where $c^{i,j}_{n_{1,2}lm}$ are the expansion coefficients in terms of the radial basis functions (labeled by $n_{1,2}$) and angular momentum channels (labeled by l) for the i, j-th atom, and N is the number of atoms. Therefore, the Δ-metric calculations required only that the following quantities be obtained from ML model training: the atomic structure of interest, the atomic structures from the training set, and the absolute errors of prediction at these points.

UQ in bandgap prediction. To show the efficacy of the proposed UQ metric, we selected bandgap prediction because of its fundamental role in determining material performance in many applications[74–76]. Concretely, four use cases that varied in the predictive model algorithm, type of materials representation, and materials class (the corresponding abbreviations are shown in brackets) were considered: the materials graph network for inorganic crystals (SNUMAT-MEGNet), Gaussian process regression based on classical force-field inspired descriptors for two-dimensional inorganic materials (C2DB-GPR-CFID), kernel ridge regression based on the atomic fractions and statistics of elemental properties for metal–organic frameworks (QMOF-KRR-ElemStat), and random forests based on the fragments of the simplified version of property-labeled materials fragments for organic materials (OMDB-RF-PLMF). A summary of trained model performance is shown in Fig. 1. Although we did not intend to achieve state-of-the-art (SOTA) accuracy in this study, it was still valuable to compare the presented algorithms with those known from the literature. As shown below, these estimates provided a first insight into the ability of the Δ-metric to improve the performance of the predictive model by selecting a specific subset of structures.

Wang et al.[77] implemented three models to predict the HSE bandgap of inorganic crystals from the SNUMAT database[56]. These models were trained using information from the constituent elements, PBE bandgap, and the combination of inputs from the first two models, respectively. Because we did not use the crude estimator (PBE bandgap) of the target property (HSE bandgap) in the SNUMAT-MEGNet use case, the first model with a root mean squared error (RMSE) of 0.75 eV was considered the SOTA model.

Liang and Zhu[67] used a set of physicochemical descriptors as inputs for the Lasso algorithm. The model was trained and tested using structures from the C2DB[59,60] database, reaching a mean absolute error (MAE) of 0.31 eV, which was nearly equivalent to that of our model (0.343 eV). Recently, Satsangi et al.[78] applied a novel feature selection approach to predict the GW bandgap of two-dimensional materials collected from the C2DB[59,60] and aNANt repository. The presented GPR model with blended features demonstrated an impressive RMSE of 0.15 eV. However, it should be emphasized that structures from C2DB that belonged only to the two space groups ($P\bar{6}m2$ and $P\bar{3}m1$) were considered. For this reason, we assumed that the results provided by Liang and Zhu[67] were more relevant for comparison.

Fung et al.[24] carried out a consistent benchmark study of graph neural networks on several material data sets, including the Quantum MOF database[62]. Several neural networks architectures demonstrated very similar MAE, and slightly outperformed the crystal graph convolutional neural network[8] model (MAE of 0.274 eV) trained in a study where the MOF data set was presented. Concretely, the SchNet[79] model had an MAE of 0.228 eV.

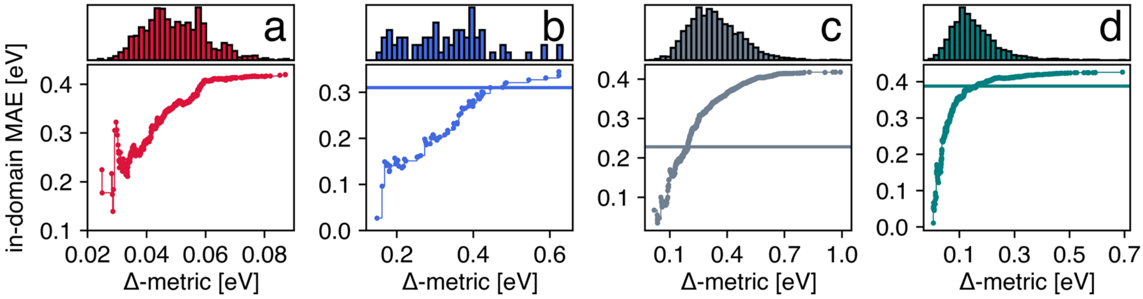

Figure 2. In-domain MAE as a function of the Δ-metric cutoff for the four considered use cases: (**a**) SNUMAT-MEGNet, (**b**) C2DB-GPR-CFID, (**c**) QMOF-KRR-ElemStat, and (**d**) OMDB-RF-PLMF, where the solid horizontal lines indicate the performance of the SOTA models available to date.

Geilhufe and Olsthoorn[80] applied the same graph neural network architecture to predict the PBE bandgap of the molecular crystals from the OMDB data set, and the model showed an MAE value of 0.406 eV. Olsthoorn et al.[64] achieved SOTA performance for this task (MAE of 0.388 eV) by averaging the predictions of KRR built on the SOAP kernel and SchNet model.

As a profitable UQ criterion, the Δ-metric should be able to demarcate the applicability domain of the predictive model. The structure was considered inside the applicability domain if the corresponding Δ-metric value did not exceed the pre-defined threshold, and an increase in model accuracy with a decrease in the threshold was desired. To validate the suggested UQ measure in this context, we examined the in-domain MAE as a function of the Δ-metric cutoff (Fig. 2). Indeed, the most general trend corresponded to the expectation that MAE could be reduced by gradually excluding it from consideration structures with a high Δ-metric value, i.e., high predictive uncertainty. As stated earlier, the performance of the SOTA models for considered tasks appeared to be a reasonable starting point to estimate the significance of decreasing the in-domain MAE. Specifically, the SOTA values of MAE in the C2DB-GPR-CFID, OMDB-RF-PLMF, and QMOF-KRR-ElemStat use cases were achieved at 87.3, 62.3, and 13.2% of the test points with the smallest Δ-metric values, respectively. In the SNUMAT-MEGNet use case, the model implemented in this study entirely covered the applicability domain of the available SOTA algorithm due to the impressive predictive performance of the graph neural network architecture. The results were mainly dependent on the combination of algorithm and material representations that we used, and those that demonstrated SOTA precision; therefore, they should not be considered a universal benchmark of suggested UQ measures. We presented an illustrative example of how Δ-metric helped to extract a (tiny) subspace of structures for which the model built on composition-only features competed on equal terms with the powerful graph neural network (QMOF-KRR-ElemStat use case).

To further explore the efficacy of the Δ-metric, we quantified the sequence monotonicity of the in-domain MAE values $\left\{\mathrm{MAE}_{(i)}\right\}_{i=1}^{N}$, as depicted in Fig. 2, using increasing coefficient IC values as defined by the following equation:

$$IC = \tfrac{1}{N-1}\left|\left\{\mathrm{MAE}_{(i)} | \mathrm{MAE}_{(i+1)} > \mathrm{MAE}_{(i)}\right\}\right|, \quad (4)$$

where a higher IC corresponds to a higher degree of monotonicity. Surprisingly, the IC values were relatively low, being 0.41, 0.63, 0.50, and 0.46 for the SNUMAT-MEGNet, C2DB-GPR-CFID, QMOF-KRR-ElemStat, and OMDB-RF-PLMF use cases, respectively. Despite the above trend in increasing in-domain MAE values with expanding applicability domain, nearly half of the entities in the $\left\{\mathrm{MAE}_{(i)}\right\}_{i=1}^{N}$ sequences were less than their predecessors. IC as a local feature helped to capture the data noise, but it was not valuable for defining the global ordering of structures according to the applied UQ measure. Next, we calculated the ranking-based metric for the area under the confidence-oracle error (AUCO):

$$AUCO = \sum_{i=1}^{N-1}\left(\mathrm{MAE}_{(i)}^{conf} - \mathrm{MAE}_{(i)}^{orac}\right), \quad (5)$$

where $\mathrm{MAE}_{(i)}^{conf}$ and $\mathrm{MAE}_{(i)}^{orac}$ were the MAEs calculated based on the subsets of structures where the i structures with the highest approximate (Δ-metric) and true (absolute error) UQ measure were removed, respectively. The corresponding $\left\{\mathrm{MAE}_{(i)}^{conf}\right\}_{i=1}^{N-1}$ and $\left\{\mathrm{MAE}_{(i)}^{orac}\right\}_{i=1}^{N-1}$ confidence curves were normalized to the $[0, 1]$ range (Fig. 3), where a lower AUCO corresponded to a higher-ranking capability. It was possible to compare Δ-metric with other UQ strategies with the confidence curves in the unified form. Specifically, we considered the following methods. In the SNUMAT-MEGNet use case, a deep ensemble[81] was implemented. The outputs of ten MEGNet models differing only in their initial weights were averaged, and the corresponding standard deviations served as the UQ measure. In the C2DB-GPR-CFID use case, the predictive variance naturally provided by GPR[82] was taken into consideration. In the QMOF-KRR-ElemStat use case, we used the subsampling[83] technique. Thirty KRR models were trained on 50% of the initial training set randomly sampled. The predictions on the test set were averaged, and the corresponding standard deviations served as the UQ measure. In the OMDB-RF-PLMF use case, the infinitesimal jackknife variance[84] was employed for UQ. As shown in Fig. 3, the deep ensemble had a significantly smaller AUCO than the Δ-metric. In all other use cases, the suggested UQ criterion outperformed

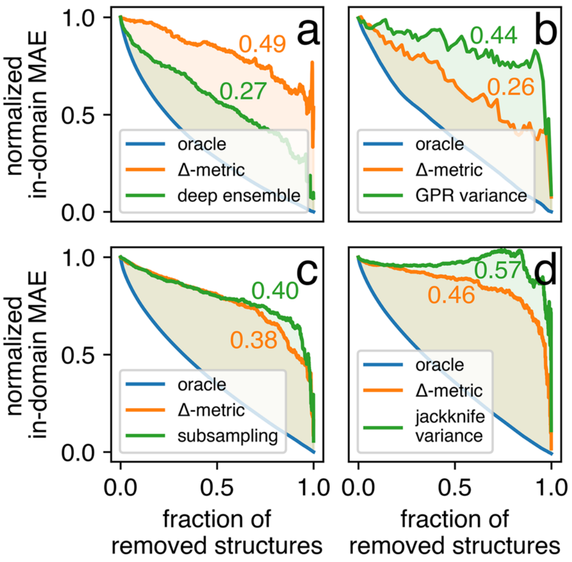

Figure 3. Confidence curves for the four considered use cases: (**a**) SNUMAT-MEGNet, (**b**) C2DB-GPR-CFID, (**c**) QMOF-KRR-ElemStat, and (**d**) OMDB-RF-PLMF. The subgraph includes three cases, the so-called oracle curve (ideal ranking according to observable error), Δ-metric curve (UQ measure presented in this study), and the curve corresponding to the competitive method.

the competitive methods. It should be emphasized that UQ via the deep ensemble strategy required resource-intensive calculations associated with training multiple models, especially with advanced neural network architectures. As a result, the Δ-metric could act as a low-cost alternative with lower accuracy, but it could be readily applied to the UQ of a single model.

Strictly speaking, the above four UQ methods served to derive epistemic uncertainty, whereas the absolute observable error incorporated into the Δ-metric also included aleatoric uncertainty[85,86]. The former was the uncertainty related to the approximate predictive model, while the latter captured the noise inherent in the data. Scalia et al.[50] demonstrated that epistemic uncertainty was the main contributor to the total uncertainty in the case of molecular data sets derived from the electronic structure theory. These data were typically self-consistent and characterized by low internal variability, i.e., low aleatoric uncertainty. Therefore, we concluded a provided comparison of the total uncertainty measure (Δ-metric) versus epistemic uncertainty techniques (deep ensemble, GPR variance, subsampling, and infinitesimal jackknife variance) was still valuable for the DFT and GW-derived bandgaps. However, this trick could probably not be used for the experimentally obtained[87,88] bandgaps.

Predictive intervals estimation. After providing a general picture of the efficacy of the Δ-metric as a ranking criterion in UQ, we presented a strategy to compute the corresponding predictive intervals. An approximate form was required to transfer from the UQ measure to the predictive error. For instance, Janet et al.[38] fitted the predictive variance to the conditional Gaussian distribution $\mathcal{N}\left(0, \sigma_1^2 + d\sigma_2^2\right)$, where d denotes the latent space distance (UQ measure), and σ_1 and σ_2 are the variable parameters. For the same purpose, we used quantile regression[89]. Given the errors in the test set and the corresponding Δ-metric values $\{\varepsilon_i, \Delta_i\}_{i=1}^{N}$, the following optimization problem was solved by:

$$\min_{\beta \in \mathbb{R}^p} \sum_{i=1}^{N} \rho_\tau(\varepsilon_i - \xi(\Delta_i, \beta)), \tag{6}$$

where ξ is a parametric function of β, and ρ_τ is a tilted absolute value function for the τ quantile. For the sake of simplicity, ξ was assumed a linear function of the parameters. We used quantiles that corresponded to one and two standard deviations, suggesting the normal distribution of errors (Fig. 4). As expected from previous analysis, the predictive intervals derived by quantile regression broadened significantly with increasing Δ-metric value, confirming its usefulness as a measure of predictive uncertainty. Guided by this illustrative representation of UQ measure-error dependence, one could define the cutoff Δ-metric value based on a desirable level of uncertainty in terms of the width of the predictive intervals.

The suggested strategy reproduced predictive intervals surprisingly well in terms of the quantiles. Thus, the model that was trained on 80% of the test data and examined the remaining 20% predicted the fraction of points within one standard deviation and within errors of 0.15, 2.38, and 1.47% in the SNUMAT-MEGNet, QMOF-KRR-ElemStat, and OMDB-RF-PLMF use cases, respectively. In the C2DB-GPR-CFID use case, where the test set contained just 72 points, the quantile regression model had a remarkably higher error of 18.39% for the same task. For instance, the 3 UQ methods considered by Tavazza et al.[37] reached an error of about 10% in most cases.

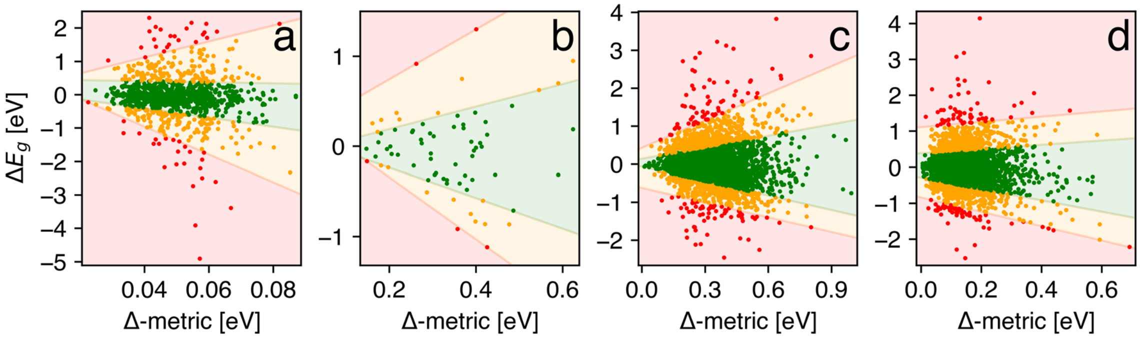

Figure 4. Predictive error as a function of the Δ-metric for the four considered use cases: (**a**) SNUMAT-MEGNet, (**b**) C2DB-GPR-CFID, (**c**) QMOF-KRR-ElemStat, and (**d**) OMDB-RF-PLMF. The colored areas correspond to the predictive intervals extracted by quantile regression: within one standard deviation (green), within two standard deviations (orange), and over two standard deviations (red).

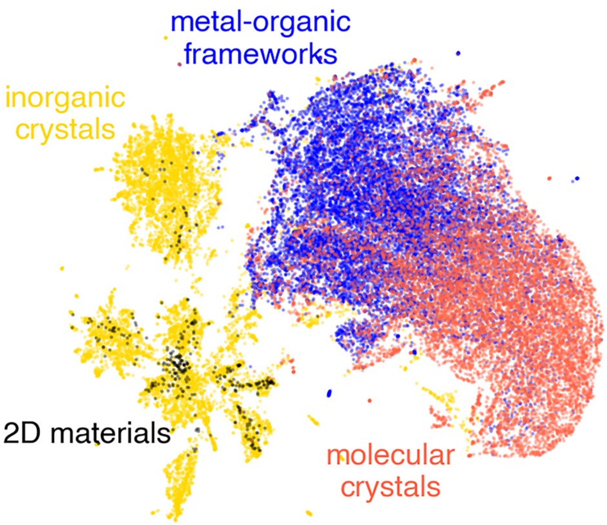

Figure 5. Two-dimensional representation of four considered materials subclasses extracted by the uniform manifold approximation and projection algorithm.

Out-of-domain applications. In the two previous sections, we used the subsets of the considered databases to verify the efficacy of the Δ-metric in UQ. Thus, a strong assumption that the training data and structures of interest sampled from the same distribution, i.e., independent and identically distributed, was made. In general, this was not the case[90]. Moreover, distinct classes of materials could form the former and latter distributions[91]. The performance of UQ measures in the out-of-domain regime has been of intense interest in advanced materials informatics applications, in particular, inverse design[92–95] has been associated with the exploration of materials space beyond its well-known subregions. To model a scenario where the assumption of independent and identically distributed data was not satisfied, we applied implemented predictive models to predict the bandgap of materials from other data sets. Specifically, the MEGNet model trained on inorganic crystals was used to estimate the bandgap of two-dimensional materials (SNUMAT-MEGNet → C2DB-GPR-CFID), and the KRR model trained on metal–organic frameworks was tested on molecular crystals (QMOF-KRR-ElemStat → OMDB-RF-PLMF). Two-dimensional representations of SOAP-like descriptors obtained by the uniform manifold approximation and projection[96,97] algorithm provided a first glimpse of the structural relationships between the donor and acceptor subsets (Fig. 5). The structures from the C2DB database appeared to be a nested subspace of the SNUMAT set of inorganic crystals. Indeed, approximately 9% of the structures from the SNUMAT database were identified as layered two-dimensional materials using the scoring parameter presented by Larsen et al.[98] These layered compounds were hypothetical precursors for the monolayers that formed the C2DB database, where 21% of the considered C2DB structures had counterparts with identical chemical formulas among the two-dimensional SNUMAT compounds. Clouds of points corresponding to the metal–organic frameworks (QMOF-KRR-ElemStat) and molecular crystals (OMDB-RF-PLMF) partially overlapped.

A summary of MEGNet and KRR model performances in the out-of-domain applications is shown in Fig. 6. The MEGNet model demonstrated an acceptable level of accuracy in terms of MAE, RMSE, and the coefficient of determination (R^2), whereas the KRR model trained on the MOFs did not give a practical estimation of the

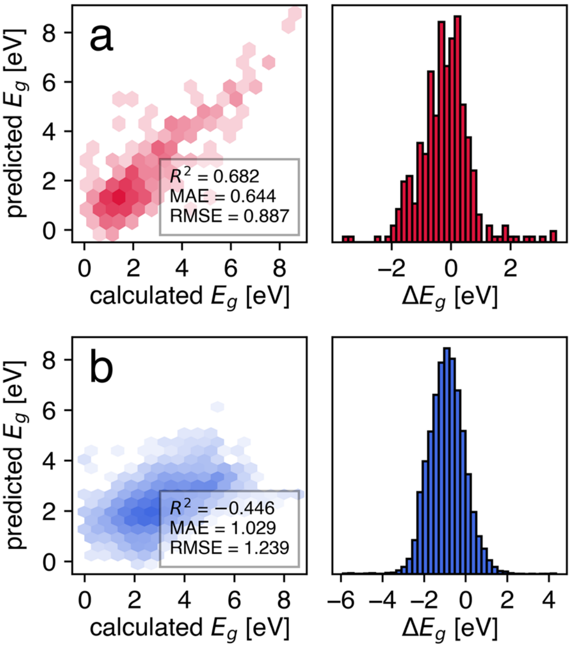

Figure 6. Parity plots and error histograms for the two considered out-of-domain use cases: (**a**) SNUMAT-MEGNet → C2DB-GPR-CFID, (**b**) QMOF-KRR-ElemStat → OMDB-RF-PLMF.

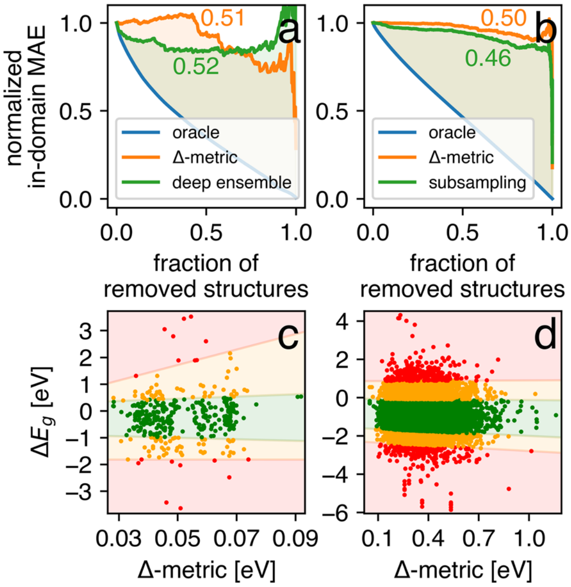

Figure 7. Confidence curves and predictive error as a function of the Δ-metric for the two considered out-of-domain use cases: (**a**, **c**) SNUMAT-MEGNet → C2DB-GPR-CFID, (**b**, **d**) QMOF-KRR-ElemStat → OMDB-RF-PLMF. Colored areas in subgraphs (**c**) and (**d**) correspond to the predictive intervals extracted by quantile regression: within one standard deviation (green), within two standard deviations (orange), and over two standard deviations (red).

bandgap for molecular crystals. Surprisingly, this model was worse than a dummy predictor, whose output was the average ensemble value for every structure (R^2 = 0.0, RMSE = 1.031, MAE = 0.798). The Δ-metric and the corresponding competitive UQ methods (deep ensemble and subsampling) showed comparable AUCO values of nearly 0.5 (Fig. 7), indicating a low ranking capability of all the above algorithms in the out-of-domain regime. Nevertheless, the Δ-metric helped to reduce in-domain MAE by 10 (20)% by considering 43 (25)% of the C2DB structures with the lowest UQ measure value. Similar behavior in the QMOF-KRR-ElemStat → OMDB-RF-PLMF use case was of little practical relevance because the reduced MAE value was still too high (greater than the error provided by a dummy predictor).

Conclusions

In summary, we considered the performance of a new UQ measure in detail, which directly provided predictive intervals for individual model output in conjunction with quantile regression. Moreover, the Δ-metric made it possible to decrease the ensemble predictive error by choosing a proper subset of structures. In contrast to the well-known variational-inference-based methods, the proposed measure was directly applicable to the UQ of a single model and agnostic to the specific predictive algorithms and featurization schemes. We believed that the Δ-metric would also help explore new subregions of materials space beyond the assumption of independent and identically distributed data.

Data availability

All data used to train ML models are from the publicly available datasets: SNUMAT band gap dataset (https://www.snumat.com), Computational 2D Materials Database (https://cmrdb.fysik.dtu.dk/c2db), Quantum MOF database (https://materialsproject.org/mofs), and Organic Materials Database (https://omdb.mathub.io).

Code availability

The code to calculate the presented UQ metric is available at https://github.com/korolewadim/dmetric.

Received: 31 May 2022; Accepted: 25 August 2022
Published online: 02 September 2022

References

1. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
2. Schmidt, J., Marques, M. R. G., Botti, S. & Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *NPJ Comput. Mater.* **5**, 1–36 (2019).
3. Jablonka, K. M., Ongari, D., Moosavi, S. M. & Smit, B. Big-data science in porous materials: Materials genomics and machine learning. *Chem. Rev.* **120**, 8066–8129 (2020).
4. Bartók, A. P., Kondor, R. & Csányi, G. On representing chemical environments. *Phys. Rev. B* **87**, 184115 (2013).
5. Faber, F., Lindmaa, A., von Lilienfeld, O. A. & Armiento, R. Crystal structure representations for machine learning models of formation energies. *Int. J. Quant. Chem.* **115**, 1094–1101 (2015).
6. Huo, H. & Rupp, M. Unified representation of molecules and crystals for machine learning. https://arXiv.org/1704.06439 (2017).
7. Isayev, O. *et al.* Universal fragment descriptors for predicting properties of inorganic crystals. *Nat. Commun.* **8**, 1–12 (2017).
8. Xie, T. & Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Phys. Rev. Lett.* **120**, 145301 (2018).
9. Chen, C., Ye, W., Zuo, Y., Zheng, C. & Ong, S. P. Graph networks as a universal machine learning framework for molecules and crystals. *Chem. Mater.* **31**, 3564–3572 (2019).
10. Choudhary, K. & DeCost, B. Atomistic line graph neural network for improved materials property predictions. *NPJ Comput. Mater.* **7**, 1–8 (2021).
11. Jha, D. *et al.* Elemnet: Deep learning the chemistry of materials from only elemental composition. *Sci. Rep.* **8**, 1–13 (2018).
12. Goodall, R. E. A. & Lee, A. A. Predicting materials properties without crystal structure: Deep representation learning from stoichiometry. *Nat. Commun.* **11**, 1–9 (2020).
13. Wang, A.Y.-T., Kauwe, S. K., Murdock, R. J. & Sparks, T. D. Compositionally restricted attention-based network for materials property predictions. *NPJ Comput. Mater.* **7**, 1–10 (2021).
14. Ong, S. P. *et al.* Python Materials Genomics (pymatgen): A robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **68**, 314–319 (2013).
15. Ward, L., Agrawal, A., Choudhary, A. & Wolverton, C. A general-purpose machine learning framework for predicting properties of inorganic materials. *NPJ Comput. Mater.* **2**, 1–7 (2016).
16. Gossett, E. *et al.* AFLOW-ML: A RESTful API for machine-learning predictions of materials properties. *Comput. Mater. Sci.* **152**, 134–145 (2018).
17. Ward, L. *et al.* Matminer: An open source toolkit for materials data mining. *Comput. Mater. Sci.* **152**, 60–69 (2018).
18. Himanen, L. *et al.* DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **247**, 106949 (2020).
19. Shao, Y., Hellström, M., Mitev, P. D., Knijff, L. & Zhang, C. PiNN: A python library for building atomic neural networks of molecules and materials. *J. Chem. Inf. Model.* **60**, 1184–1193 (2020).
20. Choudhary, K. *et al.* The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *NPJ Comput. Mater.* **6**, 1–13 (2020).
21. Jacobs, R. *et al.* The Materials Simulation Toolkit for Machine Learning (MAST-ML): An automated open source toolkit to accelerate data-driven materials research. *Comput. Mater. Sci.* **176**, 109544 (2020).
22. Dunn, A., Wang, Q., Ganose, A., Dopp, D. & Jain, A. Benchmarking materials property prediction methods: The matbench test set and automatminer reference algorithm. *NPJ Comput. Mater.* **6**, 1–10 (2020).
23. Bartel, C. J. *et al.* A critical examination of compound stability predictions from machine-learned formation energies. *NPJ Comput. Mater.* **6**, 1–11 (2020).
24. Fung, V., Zhang, J., Juarez, E. & Sumpter, B. G. Benchmarking graph neural networks for materials chemistry. *NPJ Comput. Mater.* **7**, 1–8 (2021).
25. Hu, R., Monebhurrun, V., Himeno, R., Yokota, H. & Costen, F. An adaptive least angle regression method for uncertainty quantification in FDTD computation. *IEEE Trans. Antennas Propag.* **66**, 7188–7197 (2018).
26. Hu, R., Monebhurrun, V., Himeno, R., Yokota, H. & Costen, F. A general framework for building surrogate models for uncertainty quantification in computational electromagnetics. *IEEE Trans. Antennas Propag.* **70**, 1402–1414 (2021).
27. Xue, D. *et al.* Accelerated search for materials with targeted properties by adaptive design. *Nat. Commun.* **7**, 1–9 (2016).
28. Talapatra, A. *et al.* Autonomous efficient experiment design for materials discovery with Bayesian model averaging. *Phys. Rev. Mater.* **2**, 113803 (2018).
29. Balachandran, P. V., Xue, D., Theiler, J., Hogden, J. & Lookman, T. Adaptive strategies for materials design using uncertainties. *Sci. Rep.* **6**, 1–9 (2016).

30. Lookman, T., Balachandran, P. V., Xue, D. & Yuan, R. Active learning in materials science with emphasis on adaptive sampling using uncertainties for targeted design. *NPJ Comput. Mater.* **5**, 1–17 (2019).
31. Tran, K. *et al.* Methods for comparing uncertainty quantifications for material property predictions. *Mach. Learn. Sci. Technol.* **1**, 25006 (2020).
32. Noh, J., Gu, G. H., Kim, S. & Jung, Y. Uncertainty-quantified hybrid machine learning/density functional theory high throughput screening method for crystals. *J. Chem. Inf. Model.* **60**, 1996–2003 (2020).
33. Musil, F., Willatt, M. J., Langovoy, M. A. & Ceriotti, M. Fast and accurate uncertainty estimation in chemical machine learning. *J. Chem. Theory Comput.* **15**, 906–915 (2019).
34. Tian, Y. *et al.* Role of uncertainty estimation in accelerating materials development via active learning. *J. Appl. Phys.* **128**, 14103 (2020).
35. Flores, R. A. *et al.* Active learning accelerated discovery of stable iridium oxide polymorphs for the oxygen evolution reaction. *Chem. Mater.* **32**, 5854–5863 (2020).
36. Li, Z., Achenie, L. E. K. & Xin, H. An adaptive machine learning strategy for accelerating discovery of perovskite electrocatalysts. *ACS Catal.* **10**, 4377–4384 (2020).
37. Tavazza, F., DeCost, B. & Choudhary, K. Uncertainty prediction for machine learning models of material properties. *ACS Omega* **6**, 32431–32440 (2021).
38. Janet, J. P., Duan, C., Yang, T., Nandy, A. & Kulik, H. J. A quantitative uncertainty metric controls error in neural network-driven chemical discovery. *Chem. Sci.* **10**, 7913–7922 (2019).
39. Sutton, C. *et al.* Identifying domains of applicability of machine learning models for materials science. *Nat. Commun.* **11**, 1–9 (2020).
40. Mervin, L. H., Johansson, S., Semenova, E., Giblin, K. A. & Engkvist, O. Uncertainty quantification in drug design. *Drug Discov. Today* **26**, 474–489 (2021).
41. Mayr, A., Klambauer, G., Unterthiner, T. & Hochreiter, S. DeepTox: Toxicity prediction using deep learning. *Front. Environ. Sci.* **3**, 80 (2016).
42. Sun, J. *et al.* Applying mondrian cross-conformal prediction to estimate prediction confidence on large imbalanced bioactivity data sets. *J. Chem. Inf. Model.* **57**, 1591–1598 (2017).
43. Cortés-Ciriano, I. & Bender, A. Deep confidence: A computationally efficient framework for calculating reliable prediction errors for deep neural networks. *J. Chem. Inf. Model.* **59**, 1269–1281 (2018).
44. Mervin, L. H., Afzal, A. M., Engkvist, O. & Bender, A. Comparison of scaling methods to obtain calibrated probabilities of activity for protein-ligand predictions. *J. Chem. Inf. Model.* **60**, 4546–4559 (2020).
45. Bruneau, P. & McElroy, N. R. logD 7.4 modeling using Bayesian regularized neural networks. Assessment and correction of the errors of prediction. *J. Chem. Inf. Model.* **46**, 1379–1387 (2006).
46. Zhang, Y. *et al.* Bayesian semi-supervised learning for uncertainty-calibrated prediction of molecular properties and active learning. *Chem. Sci.* **10**, 8154–8163 (2019).
47. Ryu, S., Kwon, Y. & Kim, W. Y. A Bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification. *Chem. Sci.* **10**, 8438–8446 (2019).
48. Williams, D. P., Lazic, S. E., Foster, A. J., Semenova, E. & Morgan, P. Predicting drug-induced liver injury with Bayesian machine learning. *Chem. Res. Toxicol.* **33**, 239–248 (2019).
49. Semenova, E., Williams, D. P., Afzal, A. M. & Lazic, S. E. A Bayesian neural network for toxicity prediction. *Comput. Toxicol.* **16**, 100133 (2020).
50. Scalia, G., Grambow, C. A., Pernici, B., Li, Y.-P. & Green, W. H. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *J. Chem. Inf. Model.* **60**, 2697–2717 (2020).
51. Tetko, I. V. *et al.* Critical assessment of QSAR models of environmental toxicity against Tetrahymena pyriformis: Focusing on applicability domain and overfitting by variable selection. *J. Chem. Inf. Model.* **48**, 1733–1746 (2008).
52. Sushko, I. *et al.* Applicability domains for classification problems: Benchmarking of distance to models for Ames mutagenicity set. *J. Chem. Inf. Model.* **50**, 2094–2111 (2010).
53. Sahigara, F. *et al.* Comparison of different approaches to define the applicability domain of QSAR models. *Molecules* **17**, 4791–4810 (2012).
54. Hanser, T., Barber, C., Marchaland, J. F. & Werner, S. Applicability domain: Towards a more formal definition. *SAR QSAR Environ. Res.* **27**, 865–881 (2016).
55. Berenger, F. & Yamanishi, Y. A distance-based Boolean applicability domain for classification of high throughput screening data. *J. Chem. Inf. Model.* **59**, 463–476 (2018).
56. Kim, S. *et al.* A band-gap database for semiconducting inorganic materials calculated with hybrid functional. *Sci. Data* **7**, 1–6 (2020).
57. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett.* **77**, 3865–3868 (1996).
58. Heyd, J., Scuseria, G. E. & Ernzerhof, M. Hybrid functionals based on a screened Coulomb potential. *J. Chem. Phys.* **118**, 8207–8215 (2003).
59. Haastrup, S. *et al.* The Computational 2D Materials Database: High-throughput modeling and discovery of atomically thin crystals. *2D Mater.* **5**, 42002 (2018).
60. Gjerding, M. N. *et al.* Recent progress of the Computational 2D Materials Database (C2DB). *2D Mater.* **8**, 44002 (2021).
61. Rasmussen, A., Deilmann, T. & Thygesen, K. S. Towards fully automated GW band structure calculations: What we can learn from 60.000 self-energy evaluations. *NPJ Comput. Mater.* **7**, 1–9 (2021).
62. Rosen, A. S. *et al.* Machine learning the quantum-chemical properties of metal-organic frameworks for accelerated materials discovery. *Matter* **4**, 1578–1597 (2021).
63. Borysov, S. S., Geilhufe, R. M. & Balatsky, A. V. Organic materials database: An open-access online database for data mining. *PLoS ONE* **12**, e0171501 (2017).
64. Olsthoorn, B., Geilhufe, R. M., Borysov, S. S. & Balatsky, A. V. Band gap prediction for large organic crystal structures with machine learning. *Adv. Quant. Technol.* **2**, 1900023 (2019).
65. Choudhary, K., DeCost, B. & Tavazza, F. Machine learning with force-field-inspired descriptors for materials: Fast screening and mapping energy landscape. *Phys. Rev. Mater.* **2**, 083801 (2018).
66. 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. *Phys. Rev. B* **93**, 1–12 (2016).
67. Liang, J. & Zhu, X. Phillips-inspired machine learning for band gap and exciton binding energy prediction. *J. Phys. Chem. Lett.* **10**, 5640–5646 (2019).
68. Satsangi, S., Mishra, A. & Singh, A. K. Feature blending: An approach toward generalized machine learning models for property prediction. *ACS Phys. Chem. Au* **2021**, 5 (2021).
69. Meredig, B. *et al.* Combinatorial screening for new materials in unconstrained composition space with machine learning. *Phys. Rev. B* **89**, 094104 (2014).
70. Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
71. Seabold, S. & Perktold, J. Statsmodels: Econometric and statistical modeling with python. in *Proceedings of the 9th Python in Science Conference* vol. 5761 (2010).

72. Sheridan, R. P., Feuston, B. P., Maiorov, V. N. & Kearsley, S. K. Similarity to molecules in the training set is a good discriminator for prediction accuracy in QSAR. *J. Chem. Inf. Comput. Sci.* **44**, 1912–1928 (2004).
73. Mavračić, J., Mocanu, F. C., Deringer, V. L., Csányi, G. & Elliott, S. R. Similarity between amorphous and crystalline phases: The case of TiO2. *J. Phys. Chem. Lett.* **9**, 2985–2990 (2018).
74. Olivares-Amaya, R. *et al.* Accelerated computational discovery of high-performance materials for organic photovoltaics by means of cheminformatics. *Energy Environ. Sci.* **4**, 4849–4861 (2011).
75. Setyawan, W., Gaume, R. M., Lam, S., Feigelson, R. S. & Curtarolo, S. High-throughput combinatorial database of electronic band structures for inorganic scintillator materials. *ACS Comb. Sci.* **13**, 382–390 (2011).
76. Jain, A., Shin, Y. & Persson, K. A. Computational predictions of energy materials using density functional theory. *Nat. Rev. Mater.* **1**, 1–13 (2016).
77. Wang, T., Tan, X., Wei, Y. & Jin, H. Accurate bandgap predictions of solids assisted by machine learning. *Mater. Today Commun.* **29**, 102932 (2021).
78. Satsangi, S., Mishra, A. & Singh, A. K. Feature blending: An approach toward generalized machine learning models for property prediction. *ACS Phys. Chem. Au* **2**, 16–22 (2022).
79. Schütt, K. T., Sauceda, H. E., Kindermans, P. J., Tkatchenko, A. & Müller, K. R. SchNet—a deep learning architecture for molecules and materials. *J. Chem. Phys.* **148**, 241722 (2018).
80. Geilhufe, R. M. & Olsthoorn, B. Identification of strongly interacting organic semimetals. *Phys. Rev. B* **102**, 205134 (2020).
81. Lakshminarayanan, B., Pritzel, A. & Blundell, C. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. in *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).
82. Rasmussen, C. E. Gaussian processes in machine learning. in *Summer school on machine learning* 63–71 (2003).
83. Politis, D. N. & Romano, J. P. Large sample confidence regions based on subsamples under minimal assumptions. *Ann. Stat.* **22**, 2031–2050 (1994).
84. Wager, S., Hastie, T. & Efron, B. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *J. Mach. Learn. Res.* **15**, 1625–1651 (2014).
85. Der Kiureghian, A. & Ditlevsen, O. Aleatory or epistemic? Does it matter?. *Struct. Saf.* **31**, 105–112 (2009).
86. Kendall, A. & Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? in *Advances in Neural Information Processing Systems* (eds. Guyon, I. et al.) vol. 30 (Curran Associates, Inc., 2017).
87. Zhuo, Y., Mansouri-Tehrani, A. & Brgoch, J. Predicting the band gaps of inorganic solids by machine learning. *J. Phys. Chem. Lett.* **9**, 1668–1673 (2018).
88. Marchenko, E. I. *et al.* Database of two-dimensional hybrid perovskite materials: Open-access collection of crystal structures, band gaps, and atomic partial charges predicted by machine learning. *Chem. Mater.* **32**, 7383–7388 (2020).
89. Koenker, R. & Hallock, K. F. Quantile regression. *J. Econ. Perspect.* **15**, 143–156 (2001).
90. Riley, P. *Three pitfalls to avoid in machine learning* (2019).
91. He, Y., Cubuk, E. D., Allendorf, M. D. & Reed, E. J. Metallic metal-organic frameworks predicted by the combination of machine learning methods and ab initio calculations. *J. Phys. Chem. Lett.* **9**, 4562–4569 (2018).
92. Noh, J. *et al.* Inverse design of solid-state materials via a continuous representation. *Matter* **1**, 1370–1384 (2019).
93. Korolev, V., Mitrofanov, A., Eliseev, A. & Tkachenko, V. Machine-learning-assisted search for functional materials over extended chemical space. *Mater. Horizons* **7**, 2710–2718 (2020).
94. Zhao, Y. *et al.* High-throughput discovery of novel cubic crystal materials using deep generative neural networks. *Adv. Sci.* **8**, 2100566 (2021).
95. Ren, Z. *et al.* An invertible crystallographic representation for general inverse design of inorganic crystals with targeted properties. *Matter* **5**, 314–335 (2022).
96. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. https://arXiv.org/1802.03426 (2018).
97. McInnes, L., Healy, J., Saul, N. & Grossberger, L. UMAP: Uniform manifold approximation and projection. *J. Open Sourc. Softw.* **3**, 861 (2018).
98. Larsen, P. M., Pandey, M., Strange, M. & Jacobsen, K. W. Definition of a scoring parameter to identify low-dimensional materials components. *Phys. Rev. Mater.* **3**, 34003 (2019).

Acknowledgements

We thank LetPub (www.letpub.com) for its linguistic assistance during the preparation of this manuscript.

Author contributions

V.K. and I.N. performed calculations and analysis. P.P. supervised the project. All authors discussed the results and contributed to preparing the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to V.K.

Reprints and permissions information is available at www.nature.com/reprints.

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