MATERIALS CHEMISTRY
FRONTIERS

RESEARCH ARTICLE

View Article Online
View Journal | View Issue

Cite this: *Mater. Chem. Front.*, 2025, **9**, 3339

Explainable ensemble learning to predict anisotropic nanomaterial band gap using atomic-scale structural descriptors

Ziqi Wang[a] and Kenry *[abc]

Received 29th July 2025,
Accepted 6th October 2025

DOI: 10.1039/d5qm00559k

rsc.li/frontiers-materials

Predicting the electronic band gap of nanomaterials is essential for discovering and developing novel nanostructures with tailored properties for a myriad of applications, including biomedical and pharmaceutical applications. Band gap predictions are commonly performed using computational modeling approaches such as molecular dynamics simulations and density functional theory calculations. However, the high computational cost and extensive infrastructural requirements of these methods have impeded their wider adoption and consequently, more rapid and efficient discovery of high-performance nanomaterials. In this contribution, we demonstrate the use of explainable ensemble supervised learning to accelerate the prediction of the electronic band gap of anisotropic nanomaterials. We systematically assess the capacity of several base models and a stacking model in predicting the band gap of more than 300 polyhedral nanomaterials with varying atomic-scale structural attributes. By coupling ensemble learning with explainable feature selection, we achieve outstanding performance in predicting nanomaterial band gap, with R^2 values above 0.96 and MSE below 0.004. We anticipate that this work can further catalyze the development of machine learning and other artificial intelligence approaches to streamline the prediction of the band gap and other electronic properties of nanomaterials.

Introduction

One of the most important fundamental properties of nanomaterials is their electronic band gap. It is the energy difference between the conduction band and the valence band, and the size of the band gap determines if a nanomaterial is an electrical conductor, semiconductor, or insulator.[1–3] In addition, the band gap size modulates the optical and thermal behaviors of the nanomaterials.[4–6] Consequently, by controlling the size of the band gap, nanomaterials with tunable physical properties can be designed and developed for specific applications, particularly for optoelectronics, biosensing, bioimaging, and phototherapy.[7–10]

The electronic band gap of a nanomaterial is influenced by many factors. These include internal factors like the type and arrangement of atoms and the presence of lattice defects and dopants as well as external factors like environmental pressure and temperature. Due to the central role of electronic band gap in modulating the many physical properties of nanomaterials, numerous computational approaches have been developed over the years to estimate electronic band gap. These methods include molecular dynamics simulations,[11–13] density functional theory (DFT) calculations,[14,15] quantum Monte Carlo methods,[16,17] and coupled cluster theory calculations.[18,19] While important insights into nanomaterial band gap have been obtained through these approaches, their high computational cost and extensive infrastructural requirements have prevented their wider adoption to facilitate a faster and more streamlined discovery of nanomaterials with the desired band gap and other characteristics.

With a vast amount of experimental and simulated data on nanomaterial properties generated over the past decades, recent years have seen an increasing implementation of artificial intelligence and machine learning approaches to accelerate material discovery.[20–26] Compared to conventional methods like DFT, machine learning models can be easily trained on huge datasets containing many nanomaterial properties to accurately predict the band gap of nanomaterials at a much higher speed with minimal computational resources.[27–29] One of the earliest studies on the use of machine learning for band gap analysis reported the implementation of support vector regression and artificial neural network to predict the band gap

[a] *Department of Pharmacology and Toxicology, R. Ken Coit College of Pharmacy, University of Arizona, Tucson, AZ 85721, USA. E-mail: kenry@arizona.edu*
[b] *Clinical and Translational Oncology Program and Skin Cancer Institute, University of Arizona Cancer Center, University of Arizona, Tucson, AZ 85721, USA*
[c] *BIO5 Institute, University of Arizona, Tucson, AZ 85721, USA*

of compound semiconductors based on elemental predictors.[30] Separately, a machine learning model leveraging support vector classification and regression was constructed to predict the band gap of inorganic solids based on compositional descriptors.[31] As opposed to the band gap values calculated using DFT, the machine-learning-predicted values were much closer to the experimentally derived values. More recently, machine learning classification and regression models were developed to predict the band gap of perovskite oxides based on geometric and atomic descriptors.[32]

It is crucial to highlight that, to date, many studies demonstrating the use of machine learning for band gap estimations have focused predominantly on either the development of novel algorithms or the discovery of combinations of elements with specific band gap.[33–35] Furthermore, these works have centered largely on certain material classes like perovskites and two-dimensional materials.[36–41] For the limited number of studies on band gap analysis revolving around less explored materials, such as metal-based nanostructures, the emphasis has been on isotropic nanoparticles or "black box" machine learning models.[42–44] Despite the significance of anisotropic nanomaterials like polyhedral nanostructures,[45–50] which have facet-dependent surface configurations and shape-governed physicochemical properties, to our knowledge, no study on the use of explainable machine learning to predict the band gap of anisotropic metallic nanomaterials has been reported.

In this contribution, we implemented an explainable ensemble learning approach to accelerate the prediction of the electronic band gap of more than 300 polyhedral silver (Ag) nanomaterials based on atomic-level structural descriptors. The predictive capacity of numerous base models and a stacking model was interrogated using datasets with varying number of features. Interpretable supervised learning based on SHapley Additive exPlanations (SHAP) values was employed for feature selection. Through a systematic analysis, we identified the most optimal combination of supervised learning models and structural descriptors to realize highly accurate and reliable electronic band gap predictions.

Results and discussion

The workflow of our study is illustrated in Fig. 1. Briefly, a dataset on simulated anisotropic polyhedral Ag nanomaterials with various structural attributes and electronic properties was first acquired from a publicly available database (**https://data.csiro.au/collection/csiro:23472**). The original dataset was then preprocessed to yield 347 entries of nanomaterials with 20 atomic-level structural attributes (*e.g.*, zonohedron of nanomaterials, number of atoms, average radius, anisotropy, number of surface facets, number of Ag–Ag bonds, and so on) as the input descriptors and electronic band gap as the target output (Excel file S1). This preprocessed dataset was next randomly split into 70% training and cross-validation dataset and 30% testing dataset. Since we focused primarily on regression analysis, the training and testing performances of all models were evaluated according to mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R^2). In parallel, through SHAP values, we determined the importance of specific structural attributes in influencing the decisions of the better performing models. Eventually, the most important attributes were identified to construct separate training and testing datasets with reduced number of descriptors, which were then employed for model training and testing, respectively.

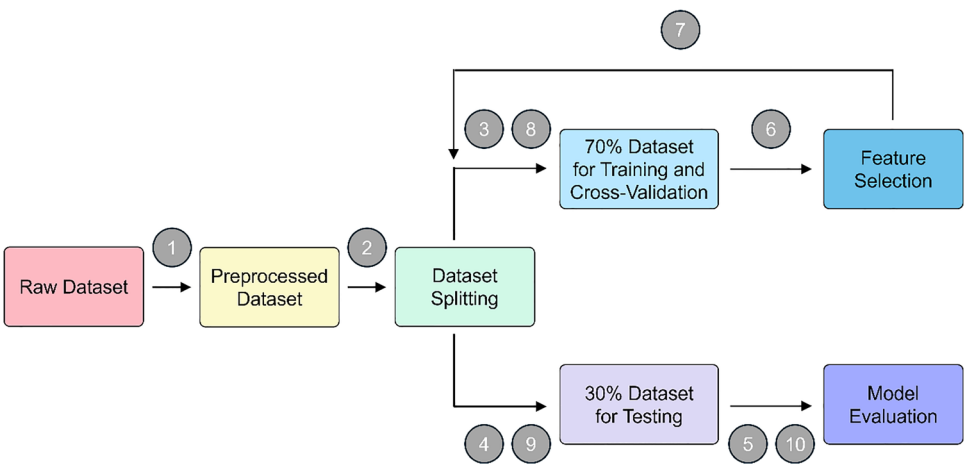

Fig. 1 Schematic illustrating the workflow of supervised learning analysis adopted in this study. The raw dataset was first preprocessed (step 1), followed by splitting the resultant dataset into 70% and 30% for model training/validation and model testing, respectively (step 2). The training/validation dataset containing the complete set of 20 features was then used to optimize all algorithm and model hyperparameters based on 10-fold cross-validation (step 3). The tuned hyperparameters were implemented to evaluate the performance of all trained models against the testing dataset (steps 4 and 5). Next, based on the training/validation dataset, all 20 features were ranked according to their SHAP values, and eight most important features were identified (step 6). Separate training/validation and testing datasets containing only the eight most essential features were constructed (step 7) and subsequently employed for model training and hyperparameter tuning (step 8) and model testing (steps 9 and 10).

To gain an insight into the atomic-scale structural properties of the polyhedral nanomaterials used in this study, we first grouped the nanostructures according to their geometric shape and the number of flat faces. Here, 12 distinct polyhedrons were identified, notably cuboctahedrons, decahedrons, great rhombicuboctahedrons, hexoctahedrons, icosahedrons, octahedrons, rhombic dodecahedrons, small rhombicuboctahedrons, tetrahedrons, tetrahexahedrons, trapezohedrons, and trisoctahedrons. Some of the structural properties of these polyhedral nanostructures were next statistically analyzed (Fig. 2 and Fig. S1). We noted that, apart from the number of surface facets and anisotropy (Fig. 2a and b), there was no statistically significant difference between all polyhedral nanostructures with respect to their other eight atomic-scale structural properties. Specifically, the 12 types of polyhedral nanostructures had comparable number of atoms, number of

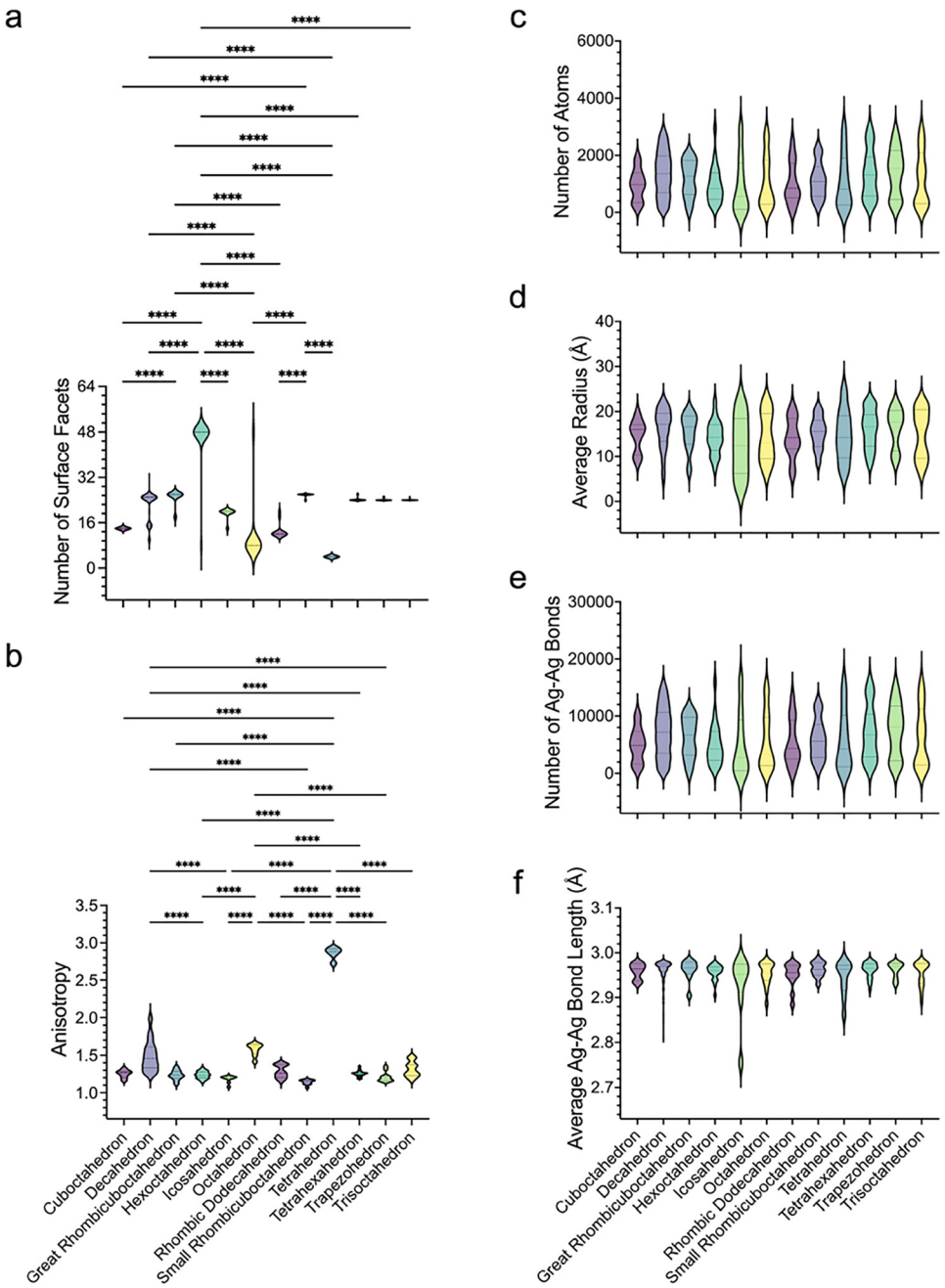

Fig. 2 Structural properties of the different polyhedral Ag nanomaterials evaluated in this study. (a) Number of surface facets, (b) anisotropy, (c) number of atoms, (d) average radius, (e) number of Ag–Ag bonds, and (f) average Ag–Ag bond length. n = 7 for cuboctahedron, 193 for decahedron, 9 for great rhombicuboctahedron, 22 for hexoctahedron, 9 for icosahedron, 17 for octahedron, 13 for rhombic dodecahedron, 12 for small rhombicuboctahedron, 10 for tetrahedron, 24 for tetrahexahedron, 11 for trapezohedron, and 20 for trisoctahedron. **** indicates $p < 0.0001$ based on the nonparametric Kruskal–Wallis test followed by Dunn's multiple comparisons test.

bulk atoms, number of surface atoms, number of FCC atoms, average radius, number of Ag–Ag bonds, average Ag–Ag bond length, and average Ag coordination number (Fig. 2c–f and Fig. S1). In terms of the number of surface facets, hexoctahedral nanostructures had the highest median value, while tetrahedrons had the lowest median value (Fig. 2a). Additionally, hexoctahedrons and octahedrons had the widest distributions of the number of surface facets, although the median value of hexoctahedrons was much higher than that of octahedrons. In terms of anisotropy, tetrahedrons had the highest median value, while the decahedral nanostructures had the largest distribution of values (Fig. 2b).

We next sought to examine if the same trend in the structural properties of the polyhedral nanomaterials would be reflected in their electronic band gap. To this end, we also statistically evaluated the band gap of the 12 types of nanostructures (Fig. 3). Similarly, we noted that there was no statistically significant difference between all polyhedral nanomaterials in terms of their electronic band gap (Fig. 3a). Interestingly, some polyhedral nanostructures like icosahedrons had a very wide distribution of band gap, ranging from about 0.6 to about 3 eV. Visualization of the two-dimensional spatial distribution of the polyhedral nanomaterials revealed that most of the nanostructures had band gap values ranging from slightly more than 0.5 to 1 eV (Fig. 3b). In addition, nanostructures with band gap values of higher than 1 eV were predominantly decahedrons. Collectively, all these indicate that the polyhedral nanomaterials examined in this study had comparable electronic band gap values.

Although our statistical analysis revealed no strong correlation between the electronic band gap and atomic-scale structural properties of polyhedral nanomaterials, we were motivated to evaluate if the nanomaterial band gap could be predicted based on a particular set of structural attributes using machine learning.

One of the most widely employed applications of supervised learning is to predict target outputs based on certain number of input descriptors. In this part of the study, we were motivated to leverage supervised learning to predict the band gap of polyhedral nanomaterials based on their atomic-scale structural attributes (Fig. 4 and Fig. S2). In particular, we sought to train, validate, and test several supervised learning models using a set of 20 structural descriptors and assessed their performances based on quantitative metrics, *i.e.*, MSE, RMSE, MAE, MAPE, and R^2 values. Five supervised learning algorithms, *i.e.*, linear regression, random forest, extreme gradient boosting, *k*-nearest neighbors, and neural network, were selected to build individual models. These algorithms were selected due to their distinct learning characteristics. Specifically, linear regression and neural network rely on model-based learning, random forest and extreme gradient boosting capitalize on ensemble learning, and *k*-nearest neighbors relies on instance-based learning. In addition, a stacking model comprising an ensemble of three base models of extreme gradient boosting, *k*-nearest neighbors, and neural network was constructed. Here, the predictions from individual base models served as the input features for the stacking model. Against the training dataset and with optimized hyperparameters, the stacking model emerged as the best performing model (Fig. S2). In fact, it outperformed all base models with the highest R^2 value of 0.983 and the lowest MSE of 0.003, RMSE of 0.053, and MAE and MAPE of 0.022. For the base models, extreme gradient boosting exhibited the best performance with an R^2 value of 0.979, MSE of 0.004, RMSE of 0.059, MAE of 0.025, and MAPE of 0.024. The *k*-nearest neighbors and neural network models were the next best performing models with R^2 values of 0.978 and 0.976, respectively, and RMSE of 0.06 and 0.063, respectively.

Against the testing dataset, we observed that there was a slight decline in the predictive performance of all models in

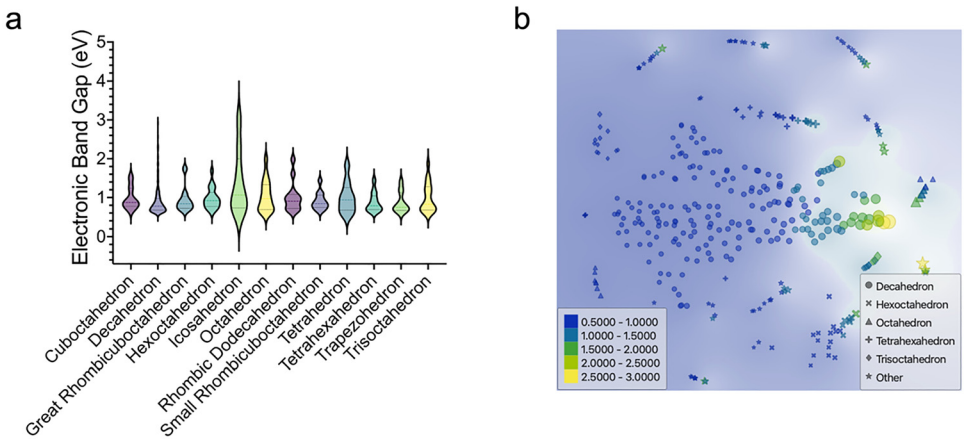

Fig. 3 Distribution of the electronic band gap of the different polyhedral Ag nanomaterials evaluated in this study. (a) Quantitative distribution of the electronic band gap of the polyhedral Ag nanomaterials. (b) Two-dimensional spatial distribution of the polyhedral Ag nanomaterials as visualized through *t*-SNE. Shape and color of the icons represent polyhedral shape and electronic band gap, respectively. n = 7 for cuboctahedron, 193 for decahedron, 9 for great rhombicuboctahedron, 22 for hexoctahedron, 9 for icosahedron, 17 for octahedron, 13 for rhombic dodecahedron, 12 for small rhombicuboctahedron, 10 for tetrahedron, 24 for tetrahexahedron, 11 for trapezohedron, and 20 for trisoctahedron.

a

Model	MSE	RMSE	MAE	MAPE	R^2
k-Nearest Neighbors	0.003	0.058	0.029	0.029	0.961
Stacking	0.003	0.059	0.022	0.021	0.959
Random Forest	0.004	0.061	0.032	0.033	0.957
Neural Network	0.004	0.063	0.019	0.017	0.954
Linear Regression	0.005	0.068	0.031	0.033	0.946
Extreme Gradient Boosting	0.005	0.069	0.035	0.036	0.944

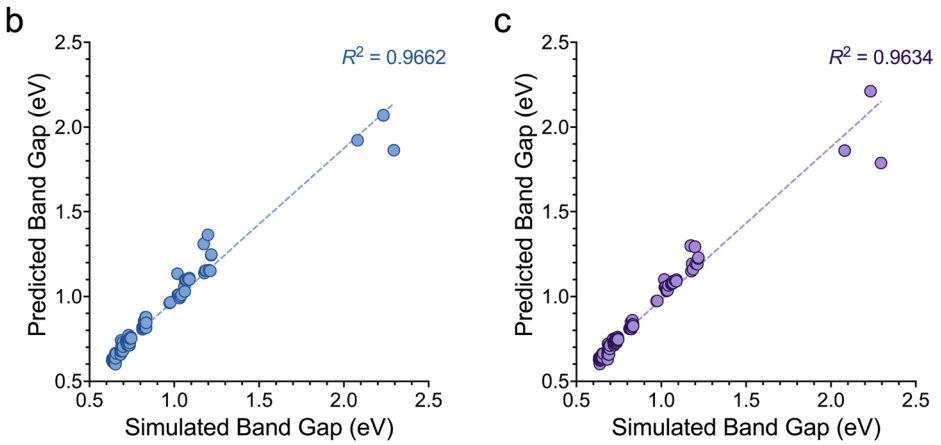

Fig. 4 Ensemble-learning-assisted prediction of nanomaterial band gap using full set of structural features. (a) Table summarizing the performance metrics of the different supervised learning models against the testing dataset. (b) and (c) Comparison of the predicted band gap values against the simulated band gap values for (b) *k*-nearest neighbors and (c) stacking models.

terms of R^2 and RMSE (Fig. 4a). For instance, the R^2 value of the stacking model decreased from 0.983 to 0.959 with the use of the testing dataset. Its RMSE, on the other hand, increased from 0.053 to 0.059. Similarly, the R^2 value of extreme gradient boosting dropped from 0.979 to 0.944 and its RMSE increased from 0.059 to 0.069. The other error metrics of this model also increased considerably. For *k*-nearest neighbors, its R^2 value decreased from 0.978 to 0.961, although its RMSE improved from 0.06 to 0.058. The R^2 values of the other three models also decreased against the testing dataset. It is, nevertheless, important to highlight that, with R^2 values ranging from 0.944 to 0.961 and with RMSE ranging from 0.058 to 0.069, the base and stacking models evaluated here still showed outstanding performance in predicting the electronic band gap of the polyhedral nanomaterials.

Comparison of the band gap values predicted by some of the supervised learning models against the simulated band gap values further verified the outstanding predictive capacity of these models (Fig. 4b and c). Specifically, for the base *k*-nearest neighbors model and the stacking model, the R^2 values quantifying the degree of linear correlation between the predicted and simulated band gap were 0.9662 and 0.9634, respectively. The strong predictive capacity of these supervised learning models was particularly evident in analyzing narrow band gap polyhedral nanomaterials with band gap values between 0.5 and 1 eV.

To gain an insight into the influence of various atomic-scale structural descriptors on band gap predictions, we acquired the SHAP values of the descriptors considered by the base and stacking models during the training and validation processes (Fig. 5). We noted that, for the stacking model, which was the best performing model, the number of atoms was the most important feature modulating the model output (Fig. 5a). The number of Ag–Ag bonds, average radius, average Ag coordination number, average Ag–Ag bond length, number of surface atoms, and number of FCC atoms were the next most important features. Like the stacking model, for extreme gradient boosting, the number of atoms was the most important feature (Fig. 5b). This was then followed by the average radius, number of bulk atoms, number of surface atoms, and number of FCC atoms, which collectively were the top five most essential features. The number of atoms was also one of the most significant features for both *k*-nearest neighbors (Fig. 5c) and neural network (Fig. 5d), although this feature ranked lower than the number of Ag–Ag bonds for *k*-nearest neighbors and lower than the average Ag coordination number, average Ag–Ag bond length, average radius, and the number of surface atoms for neural network. For *k*-nearest neighbors, the five highest ranked features were the number of Ag–Ag bonds, number of atoms, number of FCC atoms, number of bulk atoms, and number of surface atoms.

Taken together, through SHAP value analysis, we noted that the decisions of the stacking and base models with stronger predictive capacity were consistently impacted by several of the same descriptors. Notably, some of these structural attributes

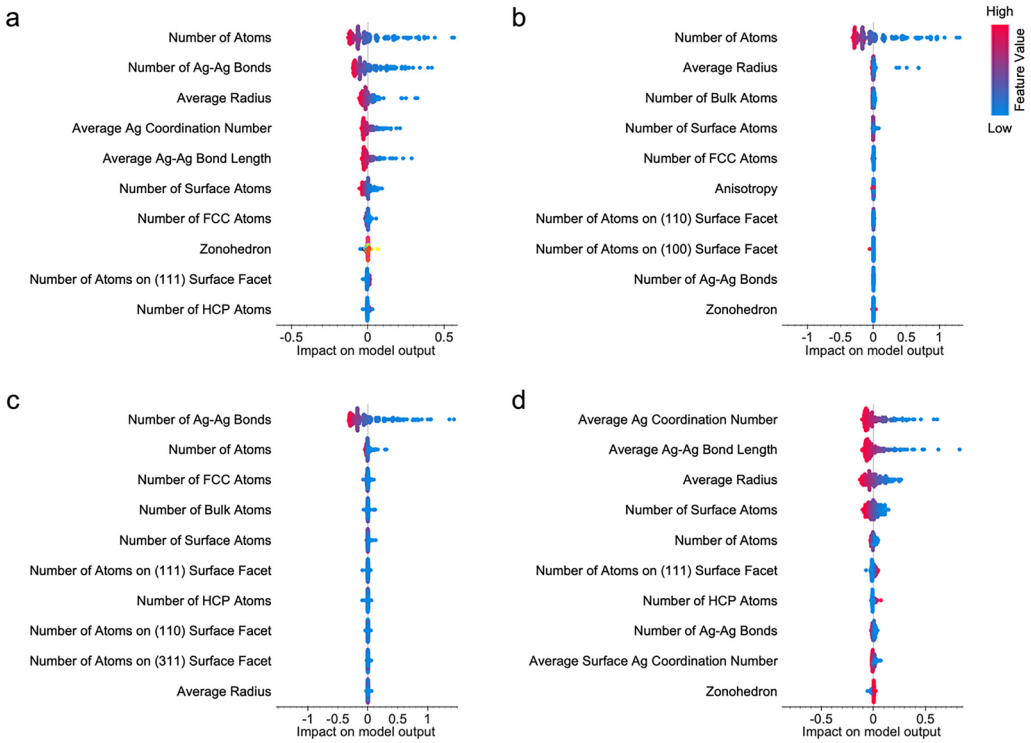

Fig. 5 Feature importance scoring and ranking. (a)–(d) SHAP values and ranking of the ten most influential structural features considered by the best performing supervised learning models during the training and validation processes: (a) stacking, (b) extreme gradient boosting, (c) *k*-nearest neighbors, and (d) neural network models.

were the number of atoms, number of surface atoms, number of bulk atoms, number of FCC atoms, number of Ag–Ag bonds, average radius, average Ag–Ag bond length, and average Ag coordination number. All these suggest that the electronic band gap of the polyhedral nanomaterials is highly correlated to their size, as reported in a previous study,[42] where the nanomaterial band gap is inversely proportional to the nanomaterial size. We were, therefore, intrigued to probe if the predictive performance of all supervised learning models would be affected if only the most essential structural descriptors were included in our analysis.

In this part of our study, we sought to re-evaluate the performance of the base and stacking models in predicting nanomaterial band gap using only the most significant atomic-scale structural descriptors (Fig. 6 and Fig. S3). To start with, we modified the training and testing datasets to yield the same nanomaterial entries, but with only eight structural attributes (instead of the complete 20 structural attributes). Against the modified training dataset and with tuned hyperparameters, the stacking model remained the best performing model during the training and validation processes (Fig. S3). Specifically, of the six supervised learning models, the stacking model had the highest R^2 value of 0.985 and the lowest MSE and RMSE of 0.002 and 0.049, respectively. Its MAE and MAPE of 0.021 were also the lowest of all absolute error values analyzed. Neural network emerged as the best performing base model with an R^2 value of 0.981 as well as MSE and RMSE of 0.003 and 0.056, respectively. The R^2 values of extreme gradient boosting, *k*-nearest neighbors, and random forest were 0.979, 0.978, and 0.977, respectively. These three models had the same MSE of 0.004.

Like our observations on the predictive performance of the models against the testing dataset containing 20 structural features, except for random forest, there was a decline in the predictive performance of the models in terms of their R^2 values with the use of a modified testing dataset (Fig. 6a). Nonetheless, most of the models experienced improvements in terms of their RMSE, MAE, and MAPE. Intriguingly, against the testing dataset containing eight structural features, random forest had the highest R^2 value of 0.98 and the lowest MSE and RMSE of 0.002 and 0.041, respectively. Comparison of the band gap values predicted by the random forest model against the simulated values revealed a high R^2 value of 0.9815 (Fig. 6b). The next best performing model was neural network, with an R^2 value of 0.97 and MSE and RMSE of 0.003 and 0.05, respectively. While the stacking model was not the best performing model, it still had a high R^2 value of 0.968 and low MSE and RMSE of 0.003 and 0.052, respectively. Furthermore, its MAE and MAPE of 0.019 and 0.018, respectively, were among the lowest. Assessment of the degree of linear correlation between the simulated band gap values and those predicted by the stacking model showed an R^2 value of 0.9735 (Fig. 6c), illustrating its strong ability in predicting polyhedral nanomaterial band gap.

In assessing the predictive capacity of the supervised learning models based on the complete and reduced sets of

a

Model	MSE	RMSE	MAE	MAPE	R^2
Random Forest	0.002	0.041	0.023	0.025	0.980
Neural Network	0.003	0.050	0.017	0.017	0.970
Stacking	0.003	0.052	0.019	0.018	0.968
Extreme Gradient Boosting	0.003	0.058	0.028	0.028	0.960
k-Nearest Neighbors	0.003	0.058	0.029	0.029	0.960
Linear Regression	0.006	0.076	0.032	0.033	0.933

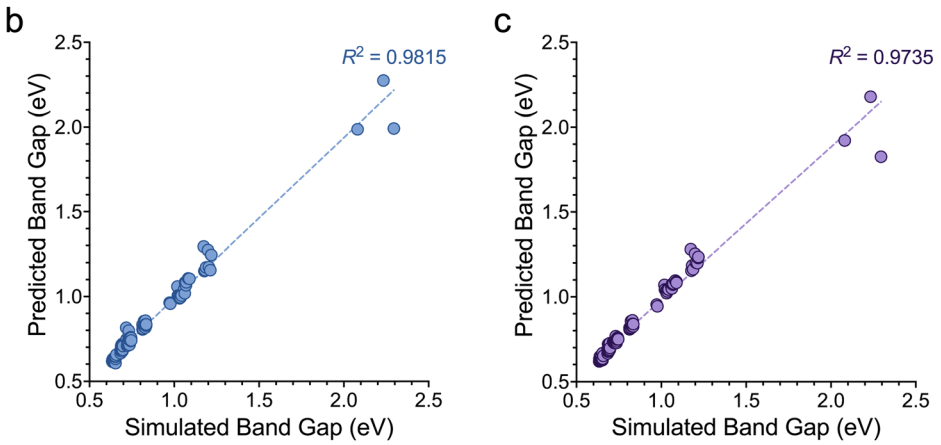

Fig. 6 Ensemble-learning-assisted prediction of nanomaterial band gap using reduced set of structural features. (a) Table summarizing the performance metrics of the different supervised learning models against the testing dataset. (b) and (c) Comparison of the predicted band gap values against the simulated band gap values for (b) random forest and (c) stacking models.

structural attributes, we noted the advantage of the implementation of ensemble learning coupled with explainable feature selection. For instance, during the training and validation processes, the R^2 value of the stacking model, which was the best performing model, increased from 0.983 to 0.985 as the dataset with 20 structural descriptors was replaced with that with eight descriptors (Fig. S2 and S3). In parallel, its MSE and RMSE improved from 0.003 and 0.053 to 0.002 and 0.049, respectively. The R^2 values and RMSE of neural network and random forest also improved with the use of training dataset with reduced number of features. Likewise, enhancements of the model predictive capacity in terms of R^2 values and RMSE with the use of a smaller dataset were reflected in the model testing performance. For example, the R^2 value of the stacking model increased from 0.959 to 0.968 and its RMSE decreased from 0.059 to 0.052 (Fig. 4a and 6a). The three base models of random forest, neural network, and extreme gradient boosting also experienced similar improvements in their R^2 values and RMSE when the modified testing dataset with a smaller number of features was used in place of the initial testing dataset.

While we noted the merit of employing only the most important structural descriptors in our predictive tasks, it is worth mentioning that many of the improvements in quantitative metrics were less than 10%. This might seem insignificant in certain cases. However, the advantage of this approach may become more apparent when dealing with larger datasets with more structural descriptors. Specifically, with a reduced number of structural descriptors, the computational time required to optimize algorithm and model hyperparameters can be considerably decreased. This is especially significant when resource-intensive algorithms like extreme gradient boosting and neural network are used. Therefore, with an increase in the dataset size and number of features, even a small enhancement in the predictive performance of supervised learning, coupled with a substantial reduction in computational cost, afforded by explainable feature selection will be beneficial.

Conclusion

Herein, we explored the use of explainable ensemble learning to predict the electronic band gap of more than 300 anisotropic polyhedral nanomaterials based on their atomic-scale structural descriptors. Using two datasets with different number of features, where one dataset comprised 20 structural attributes while the other had only eight structural attributes, we systematically assessed the performance of several base models and a stacking model in band gap predictions. We demonstrated that, irrespective of the number of features in the datasets, the predictive capacity of supervised learning during training and validation could be strengthened using a stacking model.

In parallel, the interpretability of supervised learning models could be improved using SHAP values, which in turn could be leveraged to identify the most essential features affecting predictive outcomes. We eventually showed that by combining ensemble learning and SHAP-value-guided feature selection, we could achieve outstanding performance in predicting nanomaterial band gap, with R^2 values above 0.96 and MSE below 0.004. Nevertheless, it is important to note that in this study, we employed only a single dataset containing 347 entries of nanomaterials with 20 structural attributes. This may potentially limit the generalizability of our findings. Any future work, therefore, may be extended to include larger and more diverse datasets containing more nanomaterial entries, nanomaterial types, and structural attributes. Additionally, it may be interesting to directly compare the predictive performance of different supervised learning models against that of established approaches, which is currently missing from this work. Taken together, despite the limitations of our study, we expect that it will further encourage the implementation of machine learning and other artificial intelligence approaches to streamline the analysis of the electronic properties of nanomaterials.

Methods

Acquisition and preprocessing of dataset

The original dataset of the simulated polyhedral nanomaterials used in this study was acquired from CSIRO Data Access Portal (**https://data.csiro.au/collection/csiro:23472**). These structures were optimized using density functional tight-binding method with self-consistent charges.[42] Self-consistent density functional calculations of weakly confined neutral atoms within the generalized gradient approximation were used to generate the reference density. The two-center tight-binding matrix elements within the DFT level were accounted for with a minimal valence basis set. Using a conjugate gradient methodology, forces on every atom were minimized to be less than 10^{-4} a.u. ($\sim$ 5 meV Å^{-1}) to fully relax all structures.

The raw dataset with 425 entries of nanomaterials was then preprocessed to generate a dataset comprising 347 entries of nanomaterials with 20 structural attributes as the input descriptors and electronic band gap as the output target (Excel file S1). The 20 structural attributes of the nanomaterials are zonohedron of nanomaterials, number of atoms, number of bulk atoms, number of surface atoms, average radius, anisotropy, number of atoms on (100) surface facet, number of atoms on (111) surface facet, number of atoms on (110) surface facet, number of atoms on (311) surface facet, number of surface facets, average Ag coordination number, average bulk Ag coordination number, average surface Ag coordination number, average Ag–Ag bond length, number of Ag–Ag bonds, total number of FCC atoms, total number of HCP atoms, total number of ICOS atoms, and total number of DECA atoms.

Statistical analysis

Statistical analysis was performed using GraphPad Prism 10.5.0 (GraphPad Software Inc., United States). All data were first evaluated for normality based on the Shapiro-Wilk test. The Kruskal–Wallis test followed by Dunn's multiple comparisons test were next used to assess nonparametric data. **** indicates $p < 0.0001$.

t-Distributed stochastic neighbor embedding (t-SNE) analysis

t-SNE analysis was performed using Orange Data Mining (University of Ljubljana, Slovenia). Data was normalized and preprocessing based on principal component analysis was applied. Here, 15 principal components were selected, and Euclidean distance metric was used.

Supervised learning analysis and feature selection

Supervised learning analysis and feature selection were carried out using Orange Data Mining (University of Ljubljana, Slovenia). To minimize biased evaluations, the preprocessed datasets were first randomly partitioned into 70% training (243 entries) and 30% testing (104 entries) datasets. Five supervised learning algorithms, *i.e.*, linear regression, random forest, extreme gradient boosting, *k*-nearest neighbors, and neural network, were selected for constructing individual models. In addition, a stacking model comprising an ensemble of three base models of extreme gradient boosting, *k*-nearest neighbors, and neural network was employed. The specific weight of each individual base model within the stacking ensemble was not manually assigned as the meta-model implicitly learned the most optimal way to combine the predictions of different base models during the training process. The hyperparameters of all algorithms and models (Table S1) were optimized using the training datasets and 10-fold cross-validation to generate the best performance metrics (Tables S2 and S3). The performance of the supervised learning models was quantified in terms of mean squared error (MSE), root mean squared error (RMSE), mean absolute error (MAE), mean absolute percentage error (MAPE), and coefficient of determination (R^2). Based on the optimized hyperparameters, the testing performance of all models was then evaluated. Using the SHapley Additive exPlanations (SHAP) values, feature ranking and selection were performed.

Author contributions

K. conceived, designed, and supervised the study. Z. W. conducted supervised learning analysis. All authors wrote, read, revised, and approved the submission of the manuscript.

Conflicts of interest

There are no conflicts of interest to declare.

Data availability

The data supporting this article have been included as part of the supplementary information (SI). Supplementary information is available. See DOI: **https://doi.org/10.1039/d5qm00559k**.

Acknowledgements

The authors would like to acknowledge the departmental start-up fund of Kenry from the Department of Pharmacology and Toxicology, R. Ken Coit College of Pharmacy, University of Arizona.

References

1 A. M. Smith and S. Nie, Semiconductor Nanocrystals: Structure, Properties, and Band Gap Engineering, *Acc. Chem. Res.*, 2010, **43**(2), 190–200.
2 C.-Z. Ning, L. Dou and P. Yang, Bandgap engineering in semiconductor alloy nanomaterials with widely tunable compositions, *Nat. Rev. Mater.*, 2017, **2**(12), 17070.
3 C. S. Boland, Y. Sun and D. G. Papageorgiou, Bandgap Engineering of 2D Materials toward High-Performing Straintronics, *Nano Lett.*, 2024, **24**(41), 12722–12732.
4 X. Cui, Q. Ruan, X. Zhuo, X. Xia, J. Hu, R. Fu, Y. Li, J. Wang and H. Xu, Photothermal Nanomaterials: A Powerful Light-to-Heat Converter, *Chem. Rev.*, 2023, **123**(11), 6891–6952.
5 J. E. Daniel, C. M. Jesby, K. E. Plass and M. E. Anderson, Multinary Tetrahedrite (Cu12–x–yMxNySb4S13) Nanoparticles: Tailoring Thermal and Optical Properties with Copper-Site Dopants, *Chem. Mater.*, 2024, **36**(7), 3246–3258.
6 A. Abareshi, M. M. Shahidi and N. Salehi, Comparison of structural, optical, and thermal properties in MoS2 based nanocomposites into cancer therapy, *J. Mater. Sci.: Mater. Med.*, 2025, **36**(1), 33.
7 A. Sitt, I. Hadar and U. Banin, Band-gap engineering, optoelectronic properties and applications of colloidal heterostructured semiconductor nanorods, *Nano Today*, 2013, **8**(5), 494–513.
8 A. Chaves, J. G. Azadani, H. Alsalman, D. R. da Costa, R. Frisenda, A. J. Chaves, S. H. Song, Y. D. Kim, D. He and J. Zhou, *et al.*, Bandgap engineering of two-dimensional semiconductor materials, *npj 2D Mater. Appl.*, 2020, **4**(1), 29.
9 Y. Zhang, X. Zhu and Y. Zhang, Exploring Heterostructured Upconversion Nanoparticles: From Rational Engineering to Diverse Applications, *ACS Nano*, 2021, **15**(3), 3709–3735.
10 J. C. Ranasinghe, A. Jain, W. Wu, K. Zhang, Z. Wang and S. Huang, Engineered 2D materials for optical bioimaging and path toward therapy and tissue engineering, *J. Mater. Res.*, 2022, **37**(10), 1689–1713.
11 C. Fang, W.-F. Li, R. S. Koster, J. Klimeš, A. van Blaaderen and M. A. van Huis, The accurate calculation of the band gap of liquid water by means of GW corrections applied to plane-wave density functional theory molecular dynamics simulations, *Phys. Chem. Chem. Phys.*, 2015, **17**(1), 365–375.
12 N. Jaykhedkar, R. Bystrický, M. Sýkora and T. Bučko, Understanding the structure-band gap relationship in SrZrS3 at elevated temperatures: a detailed NPT MD study, *J. Mater. Chem. C*, 2022, **10**(33), 12032–12042.
13 F. Creazzo, Engineering of MoSe2 and WSe2 Monolayers and Heterostructures by DFT-Molecular Dynamics Simulations, *ACS Appl. Mater. Interfaces*, 2025, **17**(27), 39676–39693.
14 Á. Morales-García, R. Valero and F. Illas, An Empirical, yet Practical Way To Predict the Band Gap in Solids by Using Density Functional Band Structure Calculations, *J. Phys. Chem. C*, 2017, **121**(34), 18862–18866.
15 M. Brütting and H. Bahmann, Kümmel, S. Predicting fundamental gaps accurately from density functional theory with non-empirical local range separation, *J. Chem. Phys.*, 2024, **160**(18), 181101.
16 Y. Yang, V. Gorelov, C. Pierleoni, D. M. Ceperley and M. Holzmann, Electronic band gaps from quantum Monte Carlo methods, *Phys. Rev. B*, 2020, **101**(8), 085115.
17 H. Shin, K. Gasperich, T. Rojas, A. T. Ngo, J. T. Krogel and A. Benali, Systematic Improvement of Quantum Monte Carlo Calculations in Transition Metal Oxides: sCI-Driven Wavefunction Optimization for Reliable Band Gap Prediction, *J. Chem. Theory Comput.*, 2024, **20**(18), 8175–8189.
18 A. Dittmer, R. Izsák, F. Neese and D. Maganas, Accurate Band Gap Predictions of Semiconductors in the Framework of the Similarity Transformed Equation of Motion Coupled Cluster Theory, *Inorg. Chem.*, 2019, **58**(14), 9303–9315.
19 E. Moerman, H. Miranda, A. Gallo, A. Irmler, T. Schäfer, F. Hummel, M. Engel, G. Kresse, M. Scheffler and A. Grüneis, Exploring the accuracy of the equation-of-motion coupled-cluster band gap of solids, *Phys. Rev. B*, 2025, **111**(12), L121202.
20 J. E. Saal, A. O. Oliynyk and B. Meredig, Machine Learning in Materials Discovery: Confirmed Predictions and Their Underlying Approaches, *Annual Rev. Mater. Res.*, 2020, **50**, 49–69.
21 E. O. Pyzer-Knapp, J. W. Pitera, P. W. J. Staar, S. Takeda, T. Laino, D. P. Sanders, J. Sexton, J. R. Smith and A. Curioni, Accelerating materials discovery using artificial intelligence, high performance computing and robotics, *npj Comput. Mater.*, 2022, **8**(1), 84.
22 Kenry, Machine Learning-Assisted Clustering of Nanoparticle-Binding Peptides and Prediction of Their Properties, *Adv. Theory Simul.*, 2023, **6**(6), 2300122.
23 Kenry, Machine-learning-guided quantitative delineation of cell morphological features and responses to nanomaterials, *Nanoscale*, 2024, **16**(42), 19656–19668.
24 Z. Wang and Kenry, Machine-learning-guided identification of protein secondary structures using spectral and structural descriptors, *Biomater. Sci.*, 2025, **13**(11), 2973–2982.
25 S. Dhoble, T.-H. Wu and Kenry, Decoding Nanomaterial-Biosystem Interactions through Machine Learning, *Angew. Chem., Int. Ed.*, 2024, **63**(16), e202318380.
26 C. Sahli, Kenry. Enhancing Nanomaterial-Based Optical Spectroscopic Detection of Cancer through Machine Learning. ACS, *Mater. Lett.*, 2024, **6**(10), 4697–4709.
27 T. Ko, T. Park, M. Kim and K. Min, Enhancing predictions of experimental band gap using machine learning and knowledge transfer, *Mater. Today Commun.*, 2024, **41**, 110717.
28 B. Liu, Y. Yan and M. Liu, Harnessing DFT and machine learning for accurate optical gap prediction in conjugated polymers, *Nanoscale*, 2025, **17**(13), 7865–7876.

29 C. Ezeakunne, B. Lamichhane and S. Kattel, Integrating density functional theory with machine learning for enhanced band gap prediction in metal oxides, *Phys. Chem. Chem. Phys.*, 2025, **27**(10), 5338–5358.
30 T. Gu, W. Lu, X. Bao and N. Chen, Using support vector regression for the prediction of the band gap and melting point of binary and ternary compound semiconductors, *Solid State Sci.*, 2006, **8**(2), 129–136.
31 Y. Zhuo, A. Mansouri Tehrani and J. Brgoch, Predicting the Band Gaps of Inorganic Solids by Machine Learning, *J. Phys. Chem. Lett.*, 2018, **9**(7), 1668–1673.
32 A. Talapatra, B. P. Uberuaga, C. R. Stanek and G. Pilania, Band gap predictions of double perovskite oxides using machine learning, *Commun. Mater.*, 2023, **4**(1), 46.
33 C. E. Belle, V. Aksakalli and S. P. Russo, A machine learning platform for the discovery of materials, *J. Cheminf.*, 2021, **13**(1), 42.
34 S. Shermukhamedov, D. Mamurjonova, T. Maihom and M. Probst, Structure to Property: Chemical Element Embeddings for Predicting Electronic Properties of Crystals, *J. Chem. Inf. Model.*, 2024, **64**(15), 5762–5770.
35 F. Dinic, I. Neporozhnii and O. Voznyy, Machine learning models for the discovery of direct band gap materials for light emission and photovoltaics, *Comput. Mater. Sci.*, 2024, **231**, 112580.
36 V. Gladkikh, D. Y. Kim, A. Hajibabaei, A. Jana, C. W. Myung and K. S. Kim, Machine Learning for Predicting the Band Gaps of ABX3 Perovskites from Elemental Properties, *J. Phys. Chem. C*, 2020, **124**(16), 8905–8918.
37 A. Sabagh Moeini, F. Shariatmadar Tehrani and A. Naeimi-Sadigh, Machine learning-enhanced band gaps prediction for low-symmetry double and layered perovskites, *Sci. Rep.*, 2024, **14**(1), 26736.
38 F. Gou, Z. Ma, Q. Yang, H. Du, Y. Li, Q. Zhang, W. You, Y. Chen, Z. Du and J. Yang, *et al.*, Machine Learning-Assisted Prediction and Control of Bandgap for Organic–Inorganic Metal Halide Perovskites, *ACS Appl. Mater. Interfaces*, 2025, **17**(12), 18383–18393.
39 S. Manti, M. K. Svendsen, N. R. Knøsgaard, P. M. Lyngby and K. S. Thygesen, Exploring and machine learning structural instabilities in 2D materials, *npj Comput. Mater.*, 2023, **9**(1), 33.
40 M. T. Dau, M. Al Khalfioui, A. Michon, A. Reserbat-Plantey, S. Vézian and P. Boucaud, Descriptor engineering in machine learning regression of electronic structure properties for 2D materials, *Sci. Rep.*, 2023, **13**(1), 5426.
41 J. Wang, Z. Li, M. Li, W. Jiao, Y. Luo, H. Liu and Y. Fang, Accurate prediction of band gap of two-dimensional monolayer materials via transfer learning, *Mater. Today Phys.*, 2025, **56**, 101774.
42 B. Sun, M. Fernandez and A. S. Barnard, Machine Learning for Silver Nanoparticle Electron Transfer Property Prediction, *J. Chem. Inf. Model.*, 2017, **57**(10), 2413–2423.
43 P. Sinha, A. Joshi, R. Dey and S. Misra, Machine-Learning-Assisted Materials Discovery from Electronic Band Structure, *J. Chem. Inf. Model.*, 2024, **64**(22), 8404–8413.
44 A. Saeed and M. A. Farrukh, Haque, H. M. u.; Javaid, D. Advanced Machine Learning Algorithms for Accurate Prediction of Band Gaps in Rare Earth Metal Oxide Nanoparticles, *ES Energy Environ.*, 2025, **28**, 1314.
45 J. Fang, S. Liu and Z. Li, Polyhedral silver mesocages for single particle surface-enhanced Raman scattering-based biosensor, *Biomaterials*, 2011, **32**(21), 4877–4884.
46 M. H. Huang, S. Rej and S.-C. Hsu, Facet-dependent properties of polyhedral nanocrystals, *Chem. Commun.*, 2014, **50**(14), 1634–1644.
47 L. Yang, J. Feng, Y. Ding, J. J. Bian and G. F. Wang, An analytical description for the elastic compression of metallic polyhedral nanoparticles, *AIP Adv.*, 2016, **6**, 8.
48 S. P. McDarby, C. J. Wang, M. E. King and M. L. Personick, An Integrated Electrochemistry Approach to the Design and Synthesis of Polyhedral Noble Metal Nanoparticles, *J. Am. Chem. Soc.*, 2020, **142**(51), 21322–21335.
49 A. Kanwal, B. Saif, A. Muhammad, W. Liu, J. Liu, H. Ren, P. Yang and Z. Lei, Hemoglobin-Promoted Growth of Polyhedral Gold Nanoparticles for the Detection of Glucose, H2O2, and Ascorbic Acid, *ACS Appl. Nano Mater.*, 2023, **6**(6), 4734–4746.
50 K. E. Hermann, Nanoparticles with cubic symmetry: classification of polyhedral shapes, *J. Phys.: Condens. Matter*, 2024, **36**(4), 045303.