ARTICLE OPEN

Machine learning on properties of multiscale multisource hydroxyapatite nanoparticles datasets with different morphologies and sizes

Ziteng Liu¹, Yinghuan Shi², Hongwei Chen¹, Tiexin Qin², Xuejie Zhou³, Jun Huo³, Hao Dong³, Xiao Yang⁴, Xiangdong Zhu⁴, Xuening Chen⁴, Li Zhang⁵, Mingli Yang⁶, Yang Gao²✉ and Jing Ma^{1,7}✉

Machine learning models for exploring structure-property relation for hydroxyapatite nanoparticles (HANPs) are still lacking. A multiscale multisource dataset is presented, including both experimental data (TEM/SEM, XRD/crystallinity, ROS, anti-tumor effects, and zeta potential) and computation results (containing 41,976 data samples with up to 9768 atoms) of nanoparticles with different sizes and morphologies at density functional theory (DFT), semi-empirical DFTB, and force field, respectively. Three geometric descriptors are set for the explainable machine learning methods to predict surface energies and surface stress of HANPs with satisfactory performance. To avoid the pre-determination of features, we also developed a predictive deep learning model within the framework of graph convolution neural network with good generalizability. Energies with DFT accuracy are achievable for large-sized nanoparticles from the learned correlations and scale functions for mapping different theoretical levels and particle sizes. The simulated XRD spectra and crystallinity values are in good agreement with experiments.

npj Computational Materials (2021)7:142; <https://doi.org/10.1038/s41524-021-00618-1>

INTRODUCTION

Hydroxyapatite, $\text{Ca}_{10}(\text{PO}_4)_6(\text{OH})_2$, is a kind of biomaterial widely distributed in human bone and teeth. Recently, hydroxyapatite nanoparticles (HANPs) have attracted intensive interest due to their biocompatibility and biological activity^{1–3}. HANPs could inhibit the proliferation of various tumor cells, such as hepatoma cells, osteosarcoma cells, lung cancer cells, and gastric cancer cells to some extent^{4,5}. More importantly, it was non-toxic to the normal tissue cells. Some research groups and co-authors of the present work had fabricated various HANPs with different morphology, size, and crystallinity, for which the in vitro and in vivo anti-melanoma effect was systematically characterized^{6,7}. By collecting these experiment data, including TEM/SEM, XRD/crystallinity (Supplementary Table S1), reactive oxygen species (ROS, Supplementary Table S2), anti-tumor effects (Supplementary Data, Source data are provided as a Source data file as shown in Data availability), and zeta potential (Supplementary Table S3), we found that the morphologies and physiochemical properties of HANPs played important roles in its anti-tumor ability (Supplementary Note 1 and Note 2). A high concentration of HANPs treatment in melanoma-bearing nude mice showed a strong inhibitory effect on tumor size and weight. Thus, HANPs hold the promise of acting as an excellent and safe biomaterial used in anti-tumor treatment.

Understanding the structure-property relation is of great importance in the rational design of HANPs with the desired structure and property. High throughput computations are highly desired to build the relationship between the different morphologies and properties of HANPs such as electrostatic potential surface (Supplementary Table S3), protein binding energy

(Supplementary Table S4), surface energy, and surface stress (Supplementary Table S5). One of the important properties of nanoparticles is surface energy, which plays a crucial role in nucleation and growth^{7–15}. Application of machine learning (ML) to rational nanomaterial design is gaining intensive attraction. In this work, we attempt to explore the correlation between the surface energy and the geometric features such as the ratio of length/diameter (L/D), the distance between the outer cap surface and HANP center (d_{cap}), and the number of Ca^{2+} coordination or contact ($N_{\text{Ca-contact}}$) of a large number of HANPs with different sizes and shapes through high throughput computations and machine learning techniques, as shown in Fig. 1.

Some popular explainable machine learning methods such as LightGBM, XGBoost, Support Vector Machine, GDRT, etc., were applied to predict the surface energies and surface stress. We also developed graph convolution neural network model to learn features for the molecular graphs, to predict surface energy more accurately, and to give a workflow and the learned scale function (λ) to get surface energies of large-sized HANPs with DFT accuracy. To better correlate with the experiment, image segmentation is used on the experimental TEM image to obtain the geometry parameters of different morphologies of HANPs for 3D HANPs model reconstruction. With these multi-source (experimental and computational) data and ML models, XRD patterns and crystallinities were simulated. The simulated results are in good agreement with the experimental ones. The survey of electrostatics potential surface (EPS) maps for different nanoparticles give a hint for predicting the nucleation tendency and binding abilities of nanoparticles with proteins, shedding insights into the understanding of anti-tumor effects and bio-mineralization of HANPs.

¹Key Laboratory of Mesoscopic Chemistry of Ministry of Education, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, P. R. China. ²State Key Laboratory for Novel Software Technology, Nanjing University, Nanjing, P. R. China. ³Kuang Yaming Honors School & Institute for Brain Sciences, Nanjing University, Nanjing, P. R. China. ⁴National Engineering Research Center for Biomaterials, Sichuan University, Chengdu, P. R. China. ⁵Institute of Atomic and Molecular Physics, Sichuan University, Chengdu, P. R. China. ⁶College of Biomedical Engineering, Sichuan University, Chengdu, P. R. China. ⁷Jiangsu Key Laboratory of Advanced Organic Materials, School of Chemistry and Chemical Engineering, Nanjing University, Nanjing, P. R. China. ✉email: gaoy@nju.edu.cn; majing@nju.edu.cn

Fig. 1 HANPs datasets constructed from experiments and calculations for machine learning of properties.

RESULTS

Construction of datasets

The present HANPs dataset contains both experimental data and theoretical data, as shown in Supplementary Table S5. Experimental data were collected from the published literatures with reference list given in Supplementary Note 1 (Supplementary Tables S1–S5, Supplementary Data). In comparison with the other well-organized nanoparticle datasets of enzyme binding energy, cellular uptake potentials by HEK293/A549 cells, ROS, logP, and zeta potentials in water/phosphate buffer, and catalytic properties of various nanoparticles^{16–18}, the HANPs dataset with a collection of various properties is rarely reported. The reported experimental data of HANPs are rather complicated with different crystal morphologies, surfaces, charges, and biological activity tested in different animals and cells in complex condition^{4,6,11}. In addition, the dataset size is very limited. It is hence difficult to build data-hungry machine learning models to directly predict properties like those done for other nanoparticles using either a set of nanoparticle descriptors or deep neural network (DNN) with many neurons in the layer¹⁸. Here we analyzed the experimental ROS and anti-tumor data using both Bayesian network and Apriori algorithm (Supplementary Note 2, Supplementary Fig. S1, Supplementary Tables S6 and S7). It is found that morphology of HANPs, sphere *versus* rod, is an important factor in both ROS indicator and in vitro inhibition rate.

Therefore, in the following subsections, we attempt to explore the correlation between the morphology and various properties such as surface energies and electrostatic potential. To achieve this target, the computational data of HANPs with different morphologies and sizes are produced by concurrence high throughput calculations at different theoretical levels of DFT (PBE/DND.44), semi-empirical DFTB, and force field (CVFF), respectively.

As shown in Fig. 2, HANPs with different morphologies such as sphere (S), rod (R), and needle (N), were generated automatically to satisfy the relationship between the crystal index and facets (Supplementary Note 3: Supplementary Fig. S2). To keep the generated HANPs neutral, some surface atoms were deleted randomly and automatically by using a homemade code. The random number was set for each fragment, namely, calcium ion, hydroxyl, and phosphorus oxygen tetrahedron of the HANPs structures. The excess parts were automatically deleted to meet the stoichiometry. Concurrent computation is applied to accelerate the data set construction process (Supplementary Note 4: Supplementary Figs. S3 and S4). Calculations were carried out at

different theoretical levels to establish the DFT, DFTB, and CVFF datasets, respectively. Each dataset contains HANPs of different sizes, i.e., with $N_{\text{atom}} \leq 176$ in DFT set, $N_{\text{atom}} \leq 968$ in DFTB set, and $N_{\text{atom}} \leq 9768$ in CVFF. As shown in Supplementary Table S5, most of DFT and DFTB data are corresponding to the sphere and rod-like nanoparticles, and the study of needle particle should resort to the CVFF calculations on the large-sized particles.

The surface energies predicted at different levels were validated in Supplementary Note 5, including the comparison of using different DFT functionals (Supplementary Fig. S5, Supplementary Tables S8 and S9); single point energy calculation versus ab initio molecular dynamics (AIMD) (Supplementary Fig. S6, Supplementary Table S10); DFT versus DFTB; different force fields and surface models (Supplementary Tables S11 and S12). The details of the whole HANPs datasets are described in Supplementary Note 6, containing file format and dataset architecture (Supplementary Fig. S7), selected HANPs of the DFT dataset (Supplementary Table S13, Supplementary Fig. S8), some samples of the DFTB dataset (Supplementary Table S14, Supplementary Fig. S9) and validation with DFT calculation (Supplementary Fig. S10), typical HANPs in the CVFF dataset (Supplementary Fig. S11) and validation with the DFTB calculation (Supplementary Fig. S12). The large-sized HANPs with different morphologies are also displayed in Supplementary Fig. S13. For those computation datasets, machine learning models were then built to predict the surface energies and surface stress.

Explainable machine learning model

It is of great importance to introduce explainable machine learning methods to understand the relationship between HANPs structure and properties. To describe the morphology differences of these HANPs, 6 feature descriptors, including ratio of length/diameter (L/D), distance between HANP center and (1 0 1) surface cap, d_{cap} , number of atoms, N_{atom} , SASA, SASA per number of supercell $N_{\text{supercell}}$, and number of Ca^{2+} coordination or contact within distance threshold $N_{\text{Ca-contact}}$, were selected as the input data (Fig. 3a). In the DFT database, 4033 configurations of HAP nanoparticles were used for machine learning models including MLP, SVR, XGBoost, LightGBM, GBRT, and Random Forest (RF) methods^{19–24}. Metrics of mean absolute errors (MAE), root mean square errors (RMSE), and coefficient of determination (R^2) were used to evaluate the model performance. All data samples were split into 8:1:1 ratio as the training, validation, and test set. The descriptors selection process together with the ‘last elimination method’²⁵ is shown in Supplementary Note 7 (Supplementary

Fig. 2 HANPs generation, multiscale datasets, and machine learning. **a** Nanoparticle generation from HA crystal structure with chemical stoichiometry. **b** Datasets at different theoretical levels. **c** Different machine learning (ML) methods applied to predict energies.

Tables S15–S18, Supplementary Fig. S14), from which 6-feature and 3-feature schemes were used to give good predictions of surface energies. Feature importance was obtained based on the average of multiple results on XGBoost, lightGBM, GBRT, and RF from the 6-feature and 3-feature scheme experiments, respectively (Supplementary Tables S19 and S20). The feature

importance for each descriptor and impressive performance of different machine learning methods are displayed in Fig. 3. The most important three features are L/D , d_{cap} , and $N_{\text{Ca-contact}}$, indicating that the shape characteristics are closely correlated to the surface stabilities. Since the surface energy is defined as the energy difference between the HANPs and the bulk divided by the

Fig. 3 Feature importance of machine learning on DFT dataset. **a** Feature importance of the selected features. **b** Machine learning of surface energies using six features. **c** Machine learning of surface energies using three features. **d** Comparison of different explainable ML methods. **e** Prediction of surface stress with both 6- and 3-feature schemes.

solvent-accessible surface area (SASA), it is not surprising that the SASA is not an important feature in predicting surface energy.

To treat larger sized systems, we resorted to the semi-empirical DFTB calculations. The DFTB dataset used for training contains 1850 optimized structures, which are classified into two groups: meta-stable nanoparticles with its surface energy over 2 J/m^2 and low energy nanoparticles of $<2 \text{ J/m}^2$. As shown in Fig. 4a, nanoparticles with higher surface energy have low calcium coordination number (<4) and nanoparticles with low surface energy have coordination number larger than 5.6. Specifically,

nanoparticles of needle shape ($L/D > 5$) with Ca coordination number of 4–6 usually have low surface energy. With the target of synthesizing stable nanoparticles with relatively low surface energy, we set a dataset with criteria of calcium coordination number >6 and $L/D > 5$. Descriptor selection was performed systematically, as shown in Supplementary Table S11. The importance of SASA is indicated by the close relation with the number of surface Ca^{2+} , called $N_{\text{surface_ca}}$. The increase in SASA and $N_{\text{surface_ca}}$ corresponds to unstable surface with the increased surface energy (Fig. 4b, Supplementary Fig. S15).

Fig. 4 Morphology and surface energy of HANPs on DFTB dataset. **a** Morphology and length diameter ratio, L/D ratio, distribution of stable and meta-stable structures. **b** The relationship between SASA, coordination number of surface calcium, and surface energy, along with two typical needle-like HANPs with low surface energy and growing direction toward (001) surface.

Interestingly, some needle-like nanoparticles growing in (001) crystal surface direction, as exemplified in Fig. 4b, have very small surface energies (N1: 0.64 J/m²; N2: 1.02 J/m² at DFTB level), approaching to the surface energy (DFT/PBE/DND4.4: 1.02 J/m²; DFTB: 0.64 J/m²) of (001) surface of HAP crystal. It will be shown that the (001) surface is crucial in HANPs not only because it has the lowest surface energy but also its XRD pattern is especially important for the crystallinity measurement.

However, the configuration spaces obtained from DFT and DFTB calculations are not large enough to cover all kinds of shapes. Only with the number of atoms increasing up to 9000, the needle-shaped nanoparticles can be formed with L/D up to 20. For such large-sized systems, force field method like CVFF was employed to yield 5940 optimized structures. It should be mentioned that surface energies predicted by DFT dataset cannot be directly compared with the CVFF dataset, since the energy of forcefield is the strain energy relative to the 'natural' parameters rather than the electronic energy in DFT (Supplementary Fig. S16). As shown in Fig. 5, both the 6-feature and 3-feature machine learning schemes in LightGBM give satisfactory performance in predicting surface energies of HANPs with different shapes within CVFF

dataset. Good performance of lightGBM on the training set, validation set, and test set are shown in Supplementary Table S21.

There are some experimental methods developed for measuring the surface energy, as summarized in Supplementary Note 5. Unfortunately, the experimentally measured surface energies of HANPs have not been reported yet. The predicted surface energies of nanoparticles are awaiting the future experimental tests. Another experimental endpoint property is surface stress²⁶ (Supplementary Note 8). Different from the surface energy, the surface stress is evaluated from the geometry distortion and compressibility, which is also predicted by machine learning method using both 6- and 3-feature schemes (Fig. 3e). As shown from Supplementary Table S22 and Supplementary Fig. S17, satisfactory performance is achieved for 6-feature scheme (MAE: 1.16 N/m; R^2 : 0.90) and 3-feature scheme (MAE: 1.47 N/m; R^2 : 0.84).

Multilevel attention graph convolutional neural network

Although the above-mentioned explainable ML methods performed well in all the computation datasets, deep learning method with predictive power and without pre-setting up the feature descriptors is highly desired in rational design of materials.

Fig. 5 Surface energy prediction on CVFF dataset with LightGBM. **a** Sphere HANPs. **b** Rod and plate HANPs. **c** Needle morphologies.

Here, we applied a multilevel attention graph convolutional neural network, called DeepMoleNet²⁷, to predict energies of HANPs. The atomic type, atomic number, out shell valence, van der Waals radius for each element, and atom node degree are used as node inputs and bond type, Gaussian expanded distance was selected as edge input, respectively (Supplementary Note 9: Supplementary Table S23). HANPs structures in both AIMD and DFT data sets including 27,033 data points were used in graph learning, among which 22,000 points were used as the training set, 2000 for validation set, and the rest nanoparticles as the test set. The learning results on DFT and DFTB datasets of DeepMoleNet are shown in Fig. 6a. Both lightGBM and deep learning methods show predicative power (Supplementary Tables S24–S26, Supplementary Fig. S18).

It was also shown in Fig. 6b that the present graph convolutional neural network has good transferability to the larger-sized particles in DFTB datasets through the training with small-sized systems of DFT and AIMD data points. To see the energy correlation between two different levels (and scales) more clearly, the comparison made in Fig. 6b is shown in two regions (left: low surface energy; right: higher than 2 J/m² at DFTB level). The transferability test through training with DFT data and predicting on DFTB dataset, represented as DFT→DFTB, displayed that the deep learning predicted surface energies at DFT level are in good correlation with the DFTB computation results. Since the comparison is made between the DFTB surface energies and the transferred ones from DFT data, MAE is not available in Fig. 6b.

The scale function, λ , is further learned from Fig. 6b, which presents a relationship between the DFT and DFTB surface energies. As illustrated in Fig. 6c, the introduction of the learned scale function into the deep learning flowchart on multi-scale data could yield the surface energies of the medium or even large-sized HANPs with DFT accuracy.

Interface with experiments: XRD, surface potential, and protein-binding ability

XRD simulations are useful to bridge the microscopic structure and the synthesized materials in experiments^{28–30}. Recently, Parameter Quantification Network³¹ (PQ-Net) was developed to predict scale factor, lattice parameter, and crystallite size of powder X-ray diffraction patterns from multi-phase systems of Ni-Pd/CeO₂-ZrO₂/Al₂O₃ catalytic material systems consisting of about 20,000 diffraction patterns. To apply the aforementioned ML models in real HANPs obtained from experimental synthesis, we realized the three-dimensional (3D) geometry reconstruction from two-dimensional (2D) TEM images by integrating the image segmentation techniques for extracting the objects in TEM images and machine learning algorithms (such as XGBoost and lightGBM) for analysis of these extracted objects (HANPs). After that, these

nanoparticle models, characterized by geometry parameters (L , D , d_{cap} , etc.) coming from TEM images (Fig. 7), were built upon satisfying the relationship between the faceted surfaces (Fig. 2).

We predicted XRD patterns for the selected HANPs, which are in good agreement with the experimental ones (Fig. 7). According to the XRD spectra, we further calculated the crystallinity of each nanoparticle. According to the experiments, HA-A and HA-E had far lower crystallinity than the other three HANPs, due to the absence of hydrothermal or calcining process⁶. Our simulated XRD and crystallinity results perfectly reflect this critical step differences during fabrication process. Furthermore, crystallinity of HANPs was also an important indicator of induced apoptosis of the tumor cells. Among the HANPs, HA-A had the lowest crystallinity but the highest inhibitory effect on the viability of melanoma cells.

Variations in molecular shape and surface charge of HANPs lead to different interactions with the protein target³² and self-aggregation. These different morphologies have distinct electrostatic potential surfaces (EPS), as shown in Fig. 8a and Supplementary Table S3. For the spherical nanoparticles, the weak positive and strong negative charges are distributed alternately, for both the charged (HA-S*) and neutral (HA-S) systems. In addition, the negatively charged spherical nanoparticles have repulsive electrostatic interactions between each other, making it relatively more difficult to aggregation and nucleation with the increase of concentration.

The HANPs can enter cells through endocytosis, induce apoptosis, and inhibit tumor metastasis. In vivo, the tumor cell membrane had locally positively charged regions which can permit the entrance of electronegatively charged particles. It was found that the negatively charged iron oxide nanoparticles were preferentially endocytosed by cervical carcinoma cells, as compared to the positively charged ones³³. Based on our simulation here, the small-sized HA-S with the strongly negatively charged surface would be easily taken by tumor cells, which was in good accordance with our experimental findings (Supplementary Table S3). Surface charge also plays a crucial role in the interaction with charged amino/proteins. It has been demonstrated that charged amino acids can regulate the nucleation of biomimetic HANPs (Supplementary Note 10).

As shown in Supplementary Fig. S19, EPS could be correlated with various properties such as surface energy, charge separation, and zeta potential. The surface potential in Supplementary Table S3, which was estimated from surface charge distributions in the presence of water solvent molecules, is related to but not identical with the zeta potential. The former is at the surface while the latter is located at the solid-liquid interface in close proximity to the solid surface. The geometry descriptors such as the length (L), shape (L/D), Ca coordination number, and (0 0 1) facet percentage

Fig. 6 Deep learning on internal and external datasets and the scale function bridging different scales. **a** Learning on internal dataset: training and testing on the same DFT or DFTB dataset, respectively. **b** Learning on external dataset: training on DFT dataset and then making predictions on DFTB dataset in different energy regions. **c** Prediction of DFT accuracy surface energies for the medium-sized nanoparticles with the number of atoms, N_{atom} , up to 1000 using the scale function learned from the energy relation between two different scales.

of nanoparticles are important factors to affect the EPS and zeta potential. In the analysis of zeta potential dataset, we found that the L , D , and L/D are correlated with the surface charge on the surface (Supplementary Table S3, Supplementary Figs. S19 and S20).

It is interesting to find in Fig. 8b that the appearance of the faceted (001) surfaces in nanoparticles leads to the charge separation on the surface. As mentioned above, the nanoparticle with the faceted surfaces, especially for the needle particle HA-N1, whose surface presents a stripe-like pattern with alternating positive and negative electrostatic potentials, has a low surface energy, even close to that of the crystal surface. To further display the relationship between the surface charge separation and the occurrence of faceted (001) surface and XRD, some nanoparticles with the evident faceted (001) surface have been illustrated in

Supplementary Fig. S21. With the increase in the particle size, the crystallinity is increased, especially for the rod-shaped nanoparticles. The surface energies of all the selected (001) facets terminated nanoparticles are very small and the charge separation takes place on the surface of rod nanoparticles.

DISCUSSION

In this study, both explainable machine learning methods and graph convolutional neural networks have been applied to predict the surface energies of various HANPs with different morphologies and sizes at different theoretical levels. Explainable machine learning methods showed predictive power in experimentally available surface energy and surface stress with just three features (L/D , d_{cap} , $N_{\text{ca-contact}}$), providing important insights to

Fig. 7 XRD simulations with the TEM-derived structures. Flowchart for XRD simulations: the image segmentation of TEM image, giving rise to 2D information. Then 3D models were built from geometry parameters such as L and D . Simulations of XRD and crystallinity are obtained from these 3D models. The surface energies are also predicted by LightGBM trained on DFTB dataset.

experimentalists in a fast way. However, when the learning problems become more complex, for example, the surfaces become more irregular with more surface atoms, it is not easy to find a uniform explainable model to give acceptable prediction accuracy for all data. The prediction was made by dividing the data into two groups, meta-stable HANPs with surface energies larger than 2 J/m² and low energy nanoparticles of <2 J/m². Developing proper descriptors for more complex structures of nanoparticles with the aid of both scientific intuitions and massive computations is still under way in our laboratory.

One the contrary, graph convolutional neural network gives high predictive power no matter how the surfaces change in a “black-box” manner. In this work, multilevel attention graph convolutional neural network for HANPs was designed and used for computational large nanomaterial graph modeling, which characterizes the structural diversity of nanomaterials and enables virtual screenings of nanoparticles. Good predictabilities were shown on surface energies of different morphologies and sizes with the introduction of the learned scale function. But it cannot provide clear physical insights underlying relationship between the features and the surface property.

To sum up, the current modeling strategy for HANPs is a universal tool for the rational material design of various nanoclusters without any laborious and complex feature engineering. The present ML flowchart and datasets will be further extended to the development of nanomaterials with desirable properties by fast virtual screening.

METHODS

Generation of nanoparticles with different morphologies and sizes

The calculations were carried out on HAP crystal structure and HAP habit with the exposed surfaces (Supplementary Fig. S1), at three different theoretical levels DFT (PBE/DND4.4), DFTB, and Consistent valence force field (CVFF), respectively.

According to the Wulff construction rule, the different morphologies of the nanoparticles are generated automatically by using the homemade codes in order to satisfy Eq. (1) and charge neutralization. It was stated that

a crystal that has the lowest surface energy is constructed by satisfying the following relation,

$$\frac{1}{d_{hkl}^2} = \frac{4}{3} \left(\frac{h^2 + hk + k^2}{a^2} \right) + \frac{l^2}{c^2}, \quad (1)$$

where (hkl) is the Miller index of the surface plane in crystal. The parameters of a and c are the lattice parameter of HAP. The parameter d_{hkl} is the distance between model center and (hkl) surface. In a similar way, we define the parameter d_{cap} , which is the distance between model center and outer faceted surface, for the HANPs. The equation was used for hexagonal crystal system material and the metal nanoparticles^{34,35}.

The surface energy, $E_{\text{surface}}^{\text{layer}}$ of a slab model cutting from a certain crystal surface is calculated according to Eq. (2).

$$E_{\text{surface}}^{\text{layer}} = \frac{E_{\text{slab}} - NE_{\text{bulk}}}{2A} \quad (2)$$

where E_{slab} is the energy of the optimized slab model with N layers, E_{bulk} is the energy of HAP crystal, and A is the surface of the exposed crystal surface. This equation is previously used for HAP, and gold surfaces^{36–39}. The slab models were cleaved from HAP crystal with thickness of 1 layer, 2 layers, and 3 layers, respectively. The slab vacuum is set to be 15 Å. The simulation models are displayed in Supplementary Fig. S5. All the theoretical simulations showed the same trend as the c (001) has the lowest surface energy. The m (100) surface and x (101) exhibit almost the same surface energy. At the level of PBE/DNP4.4 with Grimme correction, x (101) shows relative high surface energy than m (100) as shown in Supplementary Table S8. Our results also show consistency with other works (Supplementary Table S9).

Unlike the PBC model, the surface energies of HANPs are calculated according to Eq. (3).

$$E_{\text{surface}}^{\text{HANP}} = \frac{E_{\text{cluster}} - NE_{\text{bulk}}}{A} \quad (3)$$

Since the surfaces are irregular, we use the solvent-accessible surface area (SASA) of HANPs surface to describe the surface area A . The SASA is a very important feature descriptor used in the machine learning of surface energies.

The surface stress of a nanoparticle was evaluated from Eq. (4),

$$f = -\frac{3\Delta a}{2a\kappa}, \quad (4)$$

where a is the lattice constant of bulk HA, which is taken as the Ca-P distance in this work. Δa is the relative lattice constant change in HANPs, r is

Fig. 8 Electrostatic potential surface and implications for nucleation and complexation. **a** Electrostatic potential surface (EPS) of HANPs in different charge states, charged versus neutral. **b** EPS and surface energies for some selected morphologies, in which the appearance of the faceted (0 0 1) lowers the surface energy of a nanoparticle.

the mean particle diameter, and κ is the compressibility. In other words, Ca-P averaged distance is used to measure the relative lattice constant. The mean value of diameter, length, and d_{cap} is used as surface radius.

Explainable machine learning models

It is of great importance to introduce explainable machine learning methods to understand the physical process in HANPs properties. To describe the morphology differences of these HANPs, 6 feature descriptors, including ratio of length/diameter (L/D), distance between HANP center and (1 0 1) surface cap, d_{cap} , number of atoms, N_{atom} , SASA, SASA per number of supercell $N_{\text{supercell}}$, and number of Ca²⁺ coordination or contact within distance threshold $N_{\text{Ca-contact}}$, were initially selected as the input data, as shown in Fig. 3. Furthermore, the adoption of only three geometric descriptors, i.e., L/D , d_{cap} , $N_{\text{Ca-contact}}$ also works well for a large number of HANPs with different sizes and shapes. Explainable machine learning models including XGBoost, LightGBM, Random Forrest, and GRBT are used in this work. The performance of lightGBM on training set, validation set, and test set are recorded (Supplementary Table S21).

Multilevel attention graph convolutional neural network

Applying the feature engineering and machine learning approach with the fixed feature length to predict the target surface energy property may be problematic. Computational nanomaterial modeling and virtual nanomaterial screening in feature engineering are not always successful for nanomaterials¹⁷. Nanoparticles with irregular shapes and sizes arouse the challenge for finding the suitable feature descriptors. Complex feature engineering for considerable expertise chemistry knowledge is required in the traditional pattern-recognition or machine-learning way. Such feature engineering is important but labor-intensive and highlights the weakness of current learning algorithms: inability to extract and organize the discriminative information from the data^{40–42}.

Graph neural network has been widely used in many chemical applications^{43–57}. Various neural networks have been developed. Among these methods, DeepMoleNet is a kind of graph convolutional neural (GCN) combining multi-level attention with chemical descriptors through multi-task learning that achieves the state of the art performance on several datasets²⁷. The attractive idea is to apply GCN networks for

materials property prediction since chemical structures can be represented in the form of graphs with atoms for nodes and bonds for edges.

DeepMoleNet belongs to the MPNN architecture. That means that the computation complexity increased with the edge number, which limits its application toward larger graphs as they treat every graph as fully-connected graphs. In a more detailed way, DeepMoleNet calculation depends on edge network $A(e_{vw})h_w$, thus the time complexity per layer has an order of $O(m^2)$ with width c and the number of edges m strictly. Edges grow in the n^2 scale with nodes. For each input, it would be processed to the fully-connected state. It is rather difficult to treat large-sized molecules and nanoparticles. One way to speed up the calculation is to cut down the edge connections. According to our chemical intuition, it is reasonable to cut the fully connected atoms with center atom v in a cutoff radius and only consider those nearest neighbors, which influence the center core atom the most. The fully-connected MPNN achieved the best representation performance. Graph connections outside the radius should also be connected. However, to reduce the cost, atoms outside the radius were treated as environment, only parts nodes v' of that would be needed to simulate the real situation. If one cluster has 1000 atoms, the atoms within the radius are <6 . The edges would sharply decrease to save the computation cost. The strategy can be formalized in the following way.

$$m_v^{t+1} = \sum_{w \in [N_{\text{cutoff}}(v)]} M_t(h_v^t, h_w^t, e_{vw}) \quad (5)$$

Within the cutoff, the atom-atom connection takes the same as GCNN. Out of the cutoff, we randomly sample small part to represent the outer shell chemical environments. In this way, for the big molecules graph, the atom only strongly correlates with a certain small number of atoms (<10 atoms) within the cutoff radius. The calculation cost drops provided that the message received from neighbor part is sparse. The parameters used in this work are shown in Supplementary Table S23. The performance of DeepMoleNet on training set, validation set, and test set is provided (Supplementary Table S26).

Simulation of XRD and crystallinity

For the HANPs, spherical, rod-shaped, plate-like, and needle-like models with different morphologies were constructed by image segmentation of TEM pictures (Supplementary Note 10). The goal of image segmentation is to the partition of an image into various regions that each region is homogenous or similar in terms of some patterns. Thus, it is useful in differentiating the foreground from the background. During the segmentation procedure, the code assigns one label to each pixel of an image, so that pixels share similar characteristics are assigned as the same label. It is based on the assumption that the objects and the background in the image have a bimodal distribution. The threshold segmentation is one of the most popular methods for image segmentation. Threshold segmentation transforms input image to a binary image, e.g., by grouping the pixels with intensities, higher than the threshold value into one class, and the remaining pixels into another class⁵⁸. Then it would be much easier to judge the shapes of nanoparticles of different morphologies. Here, we employ the Otsu method⁵⁹ to determine the value of the global threshold and then perform image segmentation procedure according to this value. Afterward, the diameter and length parameters are obtained for the input of 3D structure generation, XRD simulation, and machine learning of surface energies. According to the XRD spectra, we calculated the crystallinity of each nanoparticle, according to the equation shown below,

$$X_c = (0.24/\beta_{002})^3, \quad (6)$$

where β_{002} is the half width of the (002) peak index. It should be mentioned that (002) index is also corresponding to the (001) surface property.

Electrostatic potential surface and surface potential prediction

Different nanoparticles' morphology can determine their binding specificity through electrostatic interactions. To simulate the surface potential in solvents, we embedded the HANPs into the explicit solvent molecules. Here, the solvents are water molecules. We then calculated the surface potential using the partial charges of atoms within 5 Å of the solvent.

DATA AVAILABILITY

Source data are provided with this paper and also at the website (<http://www.webace-3c.com/ATRMaterialDatabase/home/home>). Source data file for anti-tumor effects is provided as Supplementary Data. Computational data are also available upon request.

CODE AVAILABILITY

Source codes are available upon request.

Received: 7 April 2021; Accepted: 19 August 2021;

Published online: 08 September 2021

REFERENCES

1. Batchelar, D. L., Davidson, M., Dabrowski, W. & Cunningham, I. A. Bone-composition imaging using coherent-scatter computed tomography: assessing bone health beyond bone mineral density. *Med. Phys.* **33**, 904–915 (2006).
2. Sato, K. Mechanism of hydroxyapatite mineralization in biological systems. *J. Ceram. Soc. Jpn.* **115**, 124–130 (2007).
3. Malmberg, P. & Nygren, H. Methods for the analysis of the composition of bone tissue, with a focus on imaging mass spectrometry (TOF-SIMS). *Proteomics* **8**, 3755–3762 (2008).
4. Habibovic, P. et al. Comparative in vivo study of six hydroxyapatite-based bone graft substitutes. *J. Orthop. Res.* **26**, 1363–1370 (2008).
5. Kokubo, T. & Takadama, H. How useful is SBF in predicting in vivo bone bioactivity? *Biomaterials* **27**, 2907–2915 (2006).
6. Wu, H. et al. The in vitro and in vivo anti-melanoma effects of hydroxyapatite nanoparticles: influences of material factors. *Int. J. Nanomed.* **14**, 1177–1191 (2019).
7. Dorozhkin, S. V. Calcium Orthophosphates in Nature, Biology and Medicine. *Materials* **2**, 399–498 (2009).
8. Simmer, J. P. & Fincham, A. G. Molecular mechanisms of dental enamel formation. *Crit. Rev. Oral Biol. M.* **6**, 84–108 (1995).
9. Gajjeraman, S., Narayanan, K., Hao, J. J., Qin, C. L. & George, A. Matrix macromolecules in hard tissues control the nucleation and hierarchical assembly of hydroxyapatite. *J. Biol. Chem.* **282**, 1193–1204 (2007).
10. Doi, Y., Eanes, E. D., Shimokawa, H. & Termine, J. D. Inhibition of seeded growth of enamel apatite crystals by amelogenin and enamelin proteins in vitro. *J. Dent. Res.* **63**, 98–105 (1984).
11. Cai, Y. R. et al. Role of hydroxyapatite nanoparticle size in bone cell proliferation. *J. Mater. Chem.* **17**, 3780–3787 (2007).
12. Li, J. J., Yin, Y. J., Yao, F. L., Zhang, L. L. & Yao, K. D. Effect of nano- and micro-hydroxyapatite/chitosan-gelatin network film on human gastric cancer cells. *Mater. Lett.* **62**, 3220–3223 (2008).
13. Hong, Y. L. et al. Fabrication, biological effects, and medical applications of calcium phosphate nanoceramics. *Mat. Sci. Eng. R.* **70**, 225–242 (2010).
14. Mann, S. Molecular tectonics in biomineralization and biomimetic materials chemistry. *Nature* **365**, 499–505 (1993).
15. Vallet-Regi, M. & Gonzalez-Calbet, J. M. Calcium phosphates as substitution of bone tissues. *Prog. Solid State Chem.* **32**, 1–31 (2004).
16. Wang, W. Y. Universal nanohydrophobicity predictions using virtual nanoparticle library. *J. Cheminformatics* **11**, 6 (2019).
17. Yan, X. et al. In silico profiling nanoparticles: predictive nanomodeling using universal nanodescriptors and various machine learning approaches. *Nanoscale* **11**, 8352–8362 (2019).
18. Yan, X. L., Sedykh, A., Wang, W. Y., Yan, B. & Zhu, H. Construction of a web-based nanomaterial database by big data curation and modeling friendly nanostructure annotations. *Nat. Commun.* **11**, 2519 (2020).
19. Utgoff, P. E., Berkman, N. C. & Clouse, J. A. Decision tree induction based on efficient tree restructuring. *Mach. Learn.* **29**, 5–44 (1997).
20. Pedregosa, F. et al. Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
21. Mendonca, L. F., Vieira, S. M. & Sousa, J. M. C. Decision tree search methods in fuzzy modeling and classification. *Int. J. Approx. Reason.* **44**, 106–123 (2007).
22. Ke, G. L. et al. LightGBM: a highly efficient gradient boosting decision tree. *Adv. Neur. In.* **30**, 3146–3154 (2017).
23. Chen, T. Q. & Guestrin, C. XGBoost: a scalable tree boosting system. in *KDD'16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (2016).
24. Breiman, L. Random forests. *Mach. Learn.* **45**, 5–32 (2001).
25. Lu, S. H. et al. Accelerated discovery of stable lead-free hybrid organic-inorganic perovskites via machine learning. *Nat. Commun.* **9**, 3405 (2018).

26. Lamber, R., Wetjen, S. & Jaeger, N. I. Size dependence of the lattice parameter of small palladium particles. *Phys. Rev. B Condens. Matter* **51**, 10968–10971 (1995).
27. Liu, Z. et al. Transferable multilevel attention neural network for accurate prediction of quantum chemistry properties via multitask learning. *J. Chem. Inf. Model* **61**, 1066–1082 (2021).
28. Gopinath, S. & Philip, J. Preparation of metal oxide nanoparticles of different sizes and morphologies, their characterization using small angle X-ray scattering and study of thermal properties. *Mater. Chem. Phys.* **145**, 213–221 (2014).
29. Ingham, B. X-ray scattering characterisation of nanoparticles. *Crystallogr. Rev.* **21**, 229–303 (2015).
30. Yokaichiya, F. et al. Effects of doxorubicin on the structural and morphological characterization of solid lipid nanoparticles (SLN) using small angle neutron scattering (SANS) and small angle X-ray scattering (SAXS). *Phys. B* **551**, 191–196 (2018).
31. Dong, H. et al. A deep convolutional neural network for real-time full profile analysis of big powder diffraction data. *npj Comput. Mat.* **7**, 1–9 (2021).
32. Zhu, C. et al. The morphology of hydroxyapatite nanoparticles regulates cargo recognition in clathrin-mediated endocytosis. *Front. Mol. Biosci.* **8**, 39 (2021).
33. Villanueva, A. et al. The influence of surface functionalization on the enhanced internalization of magnetic nanoparticles in cancer cells. *Nanotechnology* **20**, 115103 (2009).
34. Kittel, C., McEuen, P. & McEuen, P. *Introduction to Solid State Physics* Vol. 8 (Wiley New York, 1996).
35. Zhu, B. E., Xu, Z., Wang, C. L. & Gao, Y. Shape evolution of metal nanoparticles in water vapor environment. *Nano. Lett.* **16**, 2628–2632 (2016).
36. Holec, D., Dumitraschkewitz, P., Vollath, D. & Fischer, F. D. Surface energy of Au nanoparticles depending on their size and shape. *Nanomaterials* **10**, 484 (2020).
37. Cheng, X. et al. Hydroxyl migration disorders the surface structure of hydroxyapatite nanoparticles. *Appl. Surf. Sci.* **416**, 901–910 (2017).
38. Wang, X., Wu, H., Cheng, X. J., Yang, M. L. & Zhang, L. Probing the surface activity of hydroxyapatite nanoparticles through their interaction with water molecules. *AIP Adv.* **10**, 065217 (2020).
39. Vollath, D., Fischer, F. D. & Holec, D. Surface energy of nanoparticles—influence of particle size and structure. *Beilstein J. Nanotechnol.* **9**, 2265–2276 (2018).
40. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444 (2015).
41. Jiang, D. J. et al. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J. Cheminformatics* **13**, 1–23 (2021).
42. Bengio, Y., Courville, A. & Vincent, P. Representation learning: a review and new. *Perspect. IEEE T. Pattern Anal. Mach. Intel.* **35**, 1798–1828 (2013).
43. Coley, C. W. et al. A graph-convolutional neural network model for the prediction of chemical reactivity. *Chem. Sci.* **10**, 370–377 (2019).
44. Eguchi, R. et al. Classification of alkaloids according to the starting substances of their biosynthetic pathways using graph convolutional neural networks. *BMC Bioinformatics* **20**, 1–13 (2019).
45. Gao, P., Zhang, J., Sun, Y. Z. & Yu, J. G. Toward accurate predictions of atomic properties via quantum mechanics descriptors augmented graph convolutional neural network: application of this novel approach in NMR chemical shifts predictions. *J. Phys. Chem. Lett.* **11**, 9812–9818 (2020).
46. Harada, S. et al. Dual graph convolutional neural network for predicting chemical networks. *BMC Bioinformatics* **21**, 1–13 (2020).
47. Hwang, D. et al. Comprehensive study on molecular supervised learning with graph neural networks. *J. Chem. Inf. Model.* **60**, 5936–5945 (2020).
48. Karamad, M. Orbital graph convolutional neural network for material property prediction. *Phys. Rev. Mater.* **4**, 093801 (2020).
49. Lee, C. K. Transfer learning with graph neural networks for optoelectronic properties of conjugated oligomers. *J. Chem. Phys.* **154**, 024906 (2021).
50. Nikitin, F., Isayev, O. & Strijov, V. DRACON: disconnected graph neural network for atom mapping in chemical reactions. *Phys. Chem. Chem. Phys.* **22**, 26478–26486 (2020).
51. Park, C. W. & Wolverton, C. Developing an improved crystal graph convolutional neural network framework for accelerated materials discovery. *Phys. Rev. Mater.* **4**, 063801 (2020).
52. Rathi, P. C., Ludlow, R. F. & Verdonk, M. L. Practical high-quality electrostatic potential surfaces for drug discovery using a graph-convolutional deep neural network. *J. Med. Chem.* **63**, 8778–8790 (2020).
53. Ruza, J. et al. Temperature-transferable coarse-graining of ionic liquids with dual graph convolutional neural networks. *J. Chem. Phys.* **153**, 164501 (2020).
54. Tsubaki, M., Tomii, K. & Sese, J. Compound-protein interaction prediction with end-to-end learning of neural networks for graphs and sequences. *Bioinformatics* **35**, 309–318 (2019).
55. Wang, S. H., Govindaraj, V. V., Gorris, J. M., Zhang, X. & Zhang, Y. D. Covid-19 classification by FGCNet with deep feature fusion from graph convolutional network and convolutional neural network. *Inf. Fusion* **67**, 208–229 (2021).
56. Wen, M. J., Blau, S. M., Spotte-Smith, E. W. C., Dwaraknath, S. & Persson, K. A. BonDNet: a graph neural network for the prediction of bond dissociation energies for charged molecules. *Chem. Sci.* **12**, 1858–1868 (2021).
57. Xuan, P., Pan, S. X., Zhang, T. G., Liu, Y. & Sun, H. Graph convolutional network and convolutional neural network based method for predicting lncRNA-disease associations. *Cells-Basel* **8**, 1012 (2019).
58. Liu, P. Y., Chen, Q. F. & Ma, J. Design of [2]rotaxane through image threshold segmentation of electrostatic potential image. *J. Comput. Chem.* **37**, 2228–2241 (2016).
59. Otsu, N. A threshold selection method from gray-level histograms. *IEEE trans. sys., man, cyber* **9**, 62–66 (1979).

ACKNOWLEDGEMENTS

This work was supported by the National Key Research and Development Program of China (2017YFB0702601), the National Natural Science Foundation of China (grant nos. 21873045, 22033004). We gratefully acknowledge the High Performance Computing Centre of Nanjing University for providing the IBM Blade cluster system and Nanxin Pharm Co., Ltd., Nanjing.

AUTHOR CONTRIBUTIONS

J.M. and Z.L. designed all the computations and algorithms. Z.L. and H.C. implemented the workflow and performed calculations. Y.G., Y.S., T.Q., and Z.L. developed database software and performed the image segmentation on images. X.Z. and X.Y. provided the TEM images. X.C. and Z.L. collected the experimental data. H.D., X.Z., and J.H. performed the EPS and surface potential calculations. All authors contributed to discussing the results as well as the writing and revising of the manuscript.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41524-021-00618-1>.

Correspondence and requests for materials should be addressed to Yang. Gao or Jing. Ma.

Reprints and permission information is available at <http://www.nature.com/reprints>

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021