# Recognition of the three-dimensional structure of small metal nanoparticles by a supervised artificial neural network

Timothée Fages, Franck Jolibois, Romuald Poteau

**Noname manuscript No.**
(will be inserted by the editor)

# Recognition of the three-dimensional structure of small metal nanoparticles by a supervised artificial neural network

**Timothée Fages, Franck Jolibois, Romuald Poteau(*)**

the date of receipt and acceptance should be inserted later

**Abstract** Catalytic characteristics of metal nanoparticles heavily depend on their global shapes and sizes as well as on the structure and environment of catalytic sites. On the computational chemistry side, calculations of thermodynamic and kinetic data involve a high calculation cost which can be significantly lowered by the use of a trained machine learning (ML) model. This paper outlines a preliminary approach that aims at classifying the shape of the metal core of nanoparticles. Four different supervised Artificial Neural Networks (ANN) were trained, tested and submitted to a challenging dataset. They are based on two different structural descriptors, Coulomb Matrices (CM) and Radial Distribution Functions (RDF). Each model is trained with hundreds of 3D models of nanoparticles that belong to eleven structural classes. The best model classifies a NP according to its discretized RDF profile and its first derivative. 100% accuracy is reached on the test stage and up to 70% accuracy is obtained on the challenging dataset. It is mainly made of compounds that have global shapes significantly different from the training set. But some non obvious structural patterns make then related to the eleven classes learned by the ANNs. Such strategy could easily be adapted to the recognition of NPs based on experimental neutron or X-ray diffraction data.

**Keywords** Artificial Intelligence · Artificial Neural Network · Transition Metal Nanoparticles · Structure Descriptors

LPCNO (IRSAMC), Université de Toulouse; INSA, UPS, CNRS (UMR 5215); Institut National des Sciences Appliquées, 135 avenue de Rangueil, F-31077 Toulouse, France; (*) romuald.poteau@univ-tlse3.fr

## 1 Introduction

The structure of nanomaterials can directly influence their physical and chemical properties, that can be of interest for applications in various fields ranging from biology, medicine, optoelectronics, catalysis, energy, etc [1]. Among nano-objects, colloidal transition metal nanoparticles (TMNPs) exhibit unique properties, often located between those of bulk materials and small clusters [2, 3], and related to their size, shape, surface composition, surface or core defects. Thanks to the art of chemical synthesis, the metal core of TMNPs exhibit a fascinating variety of shapes, most of them being in fact Platonic (Figure 1), Archimedean, or Catalan solids, or even concave or convex polyhedra. Magic numbers and other structural characteristics of such NPs are listed for example in refs. 4 or 5. Noble-metal nanocrystals can be considered as a paragon of this structural versatility, which is ruled out by kinetic or thermodynamic effects involved by synthesis conditions [6, 7].

Nanocatalysis has now become a major application which involves the use of nanomaterials of metals, oxides or semiconductors for transforming molecules into added-value compounds [1, 8]. TMNPs are valuable complement to conventional homogeneous and heterogeneous catalysts [9], potentially showing high activities and selectivities. The interest in colloidal TMNPs not only relies on their high surface area-to-volume ratio [3], but also on the high concentration of potential and different active metal sites. It is not only a matter of global shape, apexes, edges, and of diversity of crystallographic planes within the same nanocrystal, but a special site can also be related to the very nature of the surface species (hydrides, ligands, ...) that stabilize the NP [10]. In other words, and in line with the so-called Sabatier principle, it is possible to modulate the catalytic activity of a nanocatalyst by a modification of

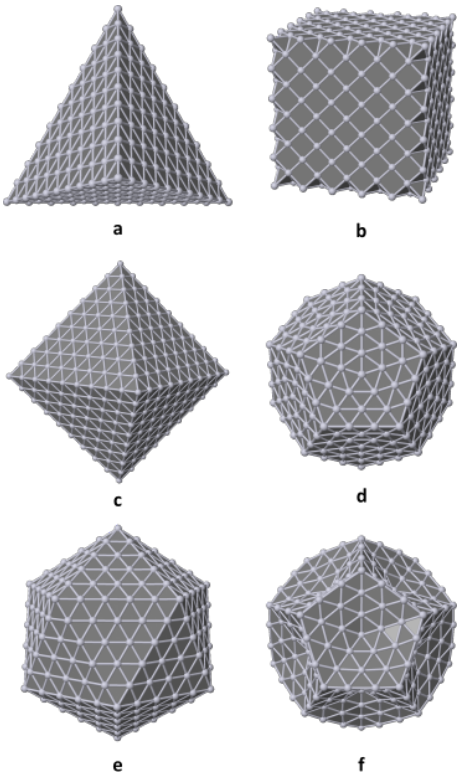


**Fig. 1** Platonic solids : 286-atoms fcc regular tetrahedron, **a**; 666-atoms fcc cube, **b**; 670-atoms fcc octahedron, **c**; 427-atoms dodecahedron, **d**; 561-atoms icosahedron, **e**; optimized 427-atoms dodecahedron, **f** (geometries **a**-**e** are not relaxed, interatomic distances are set to 2.7 Å, a typical metal-metal bond length; Sutton-Chen potential for Pt used to optimize compound **f**).

its surface composition [11]. In this context, and in opposition with trial-and-error approaches, the rational design of heterogeneous or nano- catalysts relies on correlations between descriptors and catalytic performance data like activity and selectivity. Thanks to density functional theory (DFT), it is possible to derive electronic structure descriptors and atomic structure descriptors. Catalytic activities and selectivities can be understood and predicted by associating DFT energy descriptors with microkinetic models, in the framework of the Sabatier-Balandin volcano plots [12–14], using Brønsted-Evans-Polanyi and scaling relationships [15]. This is for example the case of the search for alternatives to platinum electrodes for the hydrogen evolution reaction (HER). According to the seminal study of Nørskov and co-workers [16], the best catalysts are characterized by a dissociative hydrogen adsorption standard Gibbs free energy close to zero, *i.e.* the closer to zero $\Delta_d G^\circ(\mathrm{H}_2)$, the higher the measured exchange current $i_0$ involved by hydrogen evolution. A lot of recent studies still rely on this strategy, which involves to multiply expensive electronic structure calculations on a wide variety of materials and active sites.

Yet, a new paradigm has emerged, based upon machine learning (ML) approaches. It is especially the case in the heterogeneous catalysis domain, with approaches transferable to nanocatalysis [17]. Regarding HER, and still in the framework of the Sabatier principle which states in this case that hydrogen should neither bind too weakly nor too strongly, it has been shown on $MoS_2$ and AuCu nanoclusters that it is possible to train a model to predict $\Delta_d G^\circ(\mathrm{H}_2)$ energies for an arbitrary site based on its structural description [18]. Most of ML applications to atomistic systems need to use relevant structural local or global descriptors, such as Coulomb Matrices (CM) [19], Many-Body Tensor Representation (MBTR) [20], Atom-centered Symmetry Functions (ACSF) [21] and the Smooth Overlap of Atomic Positions (SOAP) [22] (see also ref. 23 for a comparison and discussion). The CM index is appealing by its simplicity. However, it is not invariant with respect to permutation of atomic indices, and different numbers of atoms result in different dimensionalities of the Coulomb matrices. The Radial Distribution Function (RDF) $g(r)$ is another good fingerprint for materials science that encodes information about the whole atomic structure. It is experimentally obtained from high energy X-ray diffraction. It is particularly well suited to discriminate *in situ* different crystalline structures and different shapes of NPs (see refs. 24, 25 and references therein). It is unique, continuous, differentiable with respect to atomic coordinates, invariant with respect to rotation, translation and nuclear permutation. It is however unable to distinguish between chiral compounds and it depends on the inter-atomic distances, *i.e.* it depends upon atom types. An implementation has been proposed by von Lilienfeld and coll. for machine learning models of molecular species [26]. X-ray absorption near-edge structure (XANES) spectroscopy can also be used to determine particle sizes, structural motifs, and shapes in NPs. A recent coupling of XANES with a supervised artificial neural network (ANN) showed that the ANN extends the sensitivity of XANES by enabling the determination of particle sizes and shapes [27]. The present work is related to the classification tasks in ML. It requires the use of ML algorithms that learn how to assign a class label to examples from the problem domain. It finds several remarkable applications, such as mastering the game of Go [28], image classification [29], or even cancer classification [30]. We present here an application of ML to the structural analyzis of metal NPs. After a presentation of the methods and of the various data sets, we will apply four supervised ANNs, namely one CM ANN and three RDF ANNs, to classify the shape of nanoparticles described as 3D models, such as those plotted in Figure 1.

## 2 Methods and training set

*3D models.* All cartesian coordinates of the various NPs considered in this work were generated using the in-house *polyhedra* software [24].

*Energy potentials.* Sutton-Chen potentials [31] and analytical gradients are implemented in *polyhedra*, with parameters optimized for platinum [32] and silver [33]. It has been used in some cases to relax the geometries generated from geometrical considerations.

*Coulomb Matrix (CM).* The matrix elements of a Coulomb matrix $\boldsymbol{C}$ [19] are given by:

$$C_{ij} = \begin{cases} 0.5Z_i^{2.4} & \forall\, i = j \\ \frac{Z_i Z_j}{|\boldsymbol{R}_j - \boldsymbol{R}_i|} & \forall\, i \neq j \end{cases} \quad (1)$$

where $Z_i$ is the nuclear charge of element *i*. Off-diagonal elements correspond to the Coulomb repulsion between atoms *i* and *j*, while diagonal elements encode a polynomial fit of atomic energies to nuclear charge. To avoid large values on the diagonal, all nuclear charges were set up to 1. To partially overcome the ill-defined ordering of atoms, first they were sorted in two subsets of subsurface and surface atoms, second the matrix is permuted in such a way that its rows and columns in each subset are ordered by their norm [19, 34]. In order to have the same dimensionality $d$ for all systems, matrices are completed with zeros, up to the number of atoms of the largest NP ($d_{\mathrm{CM}} = 189$ in this study).

*Theoretical Radial Distribution Functions (RDF).* The RDF function $g(r)$ has been calculated from atomic coordinates, *i.e.* directly in the real space, using (it is different from von Lilienfeld's approach [26]):

$$g(r) = A \sum_i \sum_j \left[ \frac{b_i b_j}{< b >^2} \delta(r - r_{ij}) \right] \quad (2)$$

where $r_{ij}$ is the interatomic distance between two atoms $i$ and $j$ belonging to the model crystal, $b_i$ is the scattering power of atom $i$, $< b >$ is the average scattering power of the sample and $A$ is a parameter for the amplitude of the signal. In the case of X-rays, $b_i$ is simply the number of electrons of atom *i* [35]. Eq. 2 is the so-called chemists definition [36] (another formula is often used, where $g'(r) = g(r)/r$). The delta function, $\delta(r - r_{ij})$ is replaced by a Gaussian distribution function of the form:

$$\delta(r - r_{ij}) = \frac{1}{\sqrt{2\pi}\sigma(r_{ij})} \exp\left[ -\frac{1}{2} \left( \frac{r - r_{ij}}{\sigma(r_{ij})} \right)^2 \right] \quad (3)$$

We have considered the $r$-independent formulation of the peak width $\sigma(r_{ij})$, which has been is set up to a constant value $\sigma_0 = 0.2$ Å. Note that in its usual definition, and for the purpose of comparison with the experimental RDF function $g_{\mathrm{exp}}(r)$ – which is the sine Fourier transformation of the normalized scattering intensity $S(Q)$ provided by X-Ray or neutron diffraction – the $4\pi r^2 \rho_0$ term is subtracted from eq. 2, where $\rho_0$ is the average number density of the material. Since we were only interested in a global fingerprint of TMNPs for structural analyzis, this term has not been taken into account in the present study. RDF profiles of two identical structures made of different atom types (e.g. Ag and Pt) will not coincide given that equilibrium bond lengths differ. To circumvent this issue, we considered $g(\tilde{r})$ where $\tilde{r} = r/R_{\mathrm{NN1}}$, and $R_{\mathrm{NN1}}$ is the position of the first peak of the RDF profile that characterizes nearest neighbours. $g(\tilde{r})$ is then discretized, with a $\tilde{r}$-step of 0.002 for $\tilde{r} \leq 3$, which is increased to 0.02 for $\tilde{r} > 3$. It means that a strong weight is put on the bonding scheme of an atom with its neighbors that belong to a sphere of radius $3R_{\mathrm{NN1}}$ centered on each atom. Moreover, the intensity of the first peak is normalized to 1 in order to make the intensity of the first peaks weakly dependent on the NP size. To reinforce the information, the first derivative, $g'(\tilde{r}) = dg(\tilde{r})/d\tilde{r}$, was also used to train two ANNs, namely RDF2-ANN and RDF3-ANN (*vide infra*).

*Training set.* Eleven classes were considered, with two to four different sizes according to the class. For some of them, the sizes correspond to the well-known structural magic numbers in cluster science [5, 37–39]. The 30 resulting compounds are shown in Figure 2. Each class is identified by an acronym, with -C and -S that stand for cubic and spherical shapes. The training set is made of (i) Platonic NPs: fcc and bcc cubes (FCC-C and BCC-C), fcc octahedra (OH), fcc regular tetrahedra (RTD), dodecahedra (DD), Mackay icosahedra (IC) [40]; (ii) Archimedean NP: fcc cuboctahedra (CB); (iii) spherical fcc, hcp and $\beta$-Mn NPs (FCC-S, HCP-S, BMN-S); (iv) pentagonal bipyramid decahedron (DC). On the contrary to the ten other classes made only from building principles, dodecahedra were fully optimized given the very low packing efficiency of such structure before optimization. The resulting geometries are concave (see also Figure 1). As evidenced by high-energy X-Ray diffraction technique, several NPs exhibit a $\beta$-Mn-type crystal structure [41, 42], which turns out to be a very interesting polytetrahedral packing of atoms. They are for example encountered in Frank-Kasper phases [43, 44] and in some bare transition metal clusters, such as 55-atoms species [45].

*Characteristics of the ANNs.* The supervised neural network models are multi-layer perceptrons (MLP) imple-

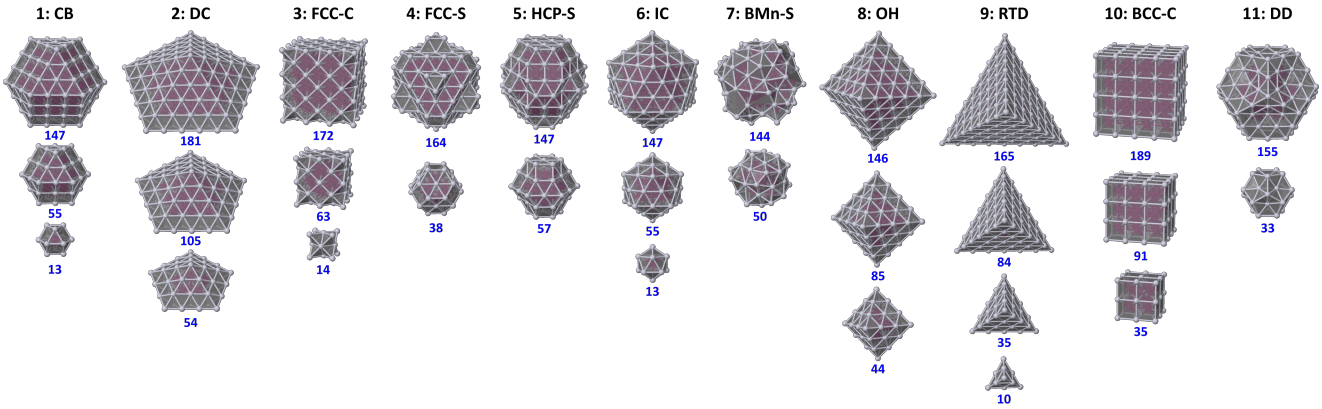


**Fig. 2** Training set made of 30 structures divided into eleven classes. Core atoms appear in magenta below surface facets. CB: cuboctahedron; FC: decahedron; FCC-C: cubic fcc shape; FCC-S: spherical piece of fcc system; HCP-S: spherical piece of hcp system; IC: icosahedron; BMn-S: spherical piece of $\beta$-Mn system; OH: octahedron; RTD: regular tetrahedron; BCC-C: cubic bcc shape; DD: regular dodecahedron (optimized). Core atoms appear in dark magenta, beneath the triangular and square facets.

mented in the scikit-learn project [46], which trains using back-propagation. The Adam stochastic optimizer was chosen given its robustness for relatively large datasets [47]. For classification, MLP minimize the cross-entropy loss function. For a given compound **j**, ANNs returns $\mathcal{P}_j(C_k)$ – with $k = 1, 11$ –, the probability number per class $C_k$. After an adjustment stage, the ANNs' architecture was set up to three hidden layers of 40 neurons each. It appeared appropriate to avoid overfitting or underfitting.

Four different implementations for the recognition of the shape of 3D models of NPs have been achieved: (i) CM-ANN, based on flattened Coulomb matrices (schematized in Figure S1 for the 105-atoms decahedron); (ii) RDF1-ANN, based on discretized $g(\tilde{r})$ RDF profiles; (iii) RDF2-ANN based on the discretized first derivate $g'(\tilde{r})$ of RDF profile; (iv) RDF3-ANN based on concatenated – and still discretized – $g(\tilde{r})$ and $g'(\tilde{r})$ fingerprints (schematized in Figure S2). RDF*i*-ANNs are designed in the same spirit as the shape recognition based on XANES experimental spectra [27], the machinery of this application being also based on the discretization of the continuous and differentiable experimental XANES signal $I(E)$ as well as on the discretization of its theoretical counterpart.

*Data augmentation and training of the ANN.* The training set summarized in Figure 2 is obviously too small to train an ANN, a process which usually requires several thousands of data. To enlarge the datasets, 1000 structures, together with their CM, their RDF function $g(\tilde{r})$ as well as $g'(\tilde{r})$, have been generated for each of the 30 nanoclusters used for training. Each new geometry is obtained by a random displacement by 0.1 Å in all directions of each atom of the parent geometry.

Ten structures per class were separated to make the test set. The remaining 990 structures were randomly split in a 80%-structures per class subset, used to train the ANN, and in a 20%-structures per class subset, used to validate the ANN performance. $g(\tilde{r})$ RDF profiles of selected fully symmetric and randomly modified fcc structures are plotted in Figure S3. One can see that these random displacements preserve the structural fingerprint of the structures. Such training will also make the ANNs able to classify the shape of the metal core of ligand-protected NPs, which interatomic metal-metal bond lengths, sensitive to electron donation or back-bonding effects of ligands and to surface coverage, differ from bare NPs.

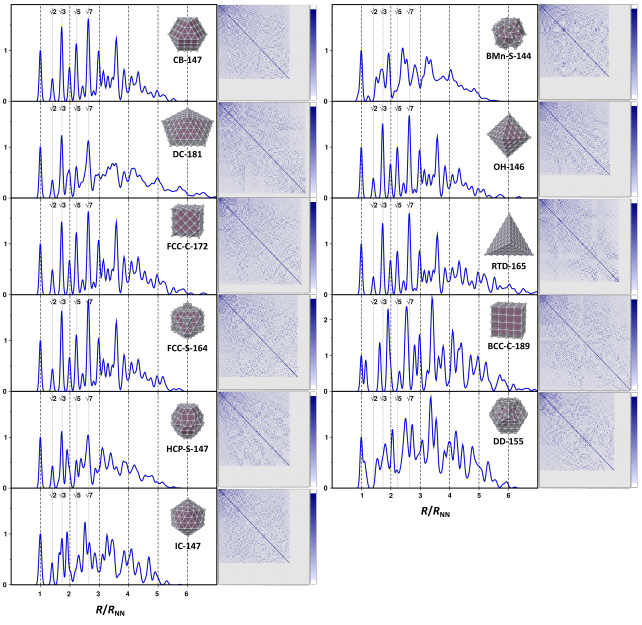


**Fig. 3** RDF functions $g(\tilde{r})$ and coulomb matrices for the largest NP in each class. Same labels as in Figure 2.

*Cross-validation.* The stratified 5-folds cross-validator of scikit-learn was used [48]. It provides an average perfor-

mance of the models, with folds made by preserving the percentage of samples for each class.

*Challenging dataset and scoring of the related performance of the ANNs.* Another dataset will be presented in section 3. It aims at evaluating the performance of the ANNs when confronted with structures that significantly differ from the training set: truncated NPs, alloys, other metal than Pt with different interatomic bond lengths, coalesced NPs, new bcc shape different from the only learned bcc cubic shape, oblate and prolate structures whereas the training set is made of structures that exhibit only small deviations from spherical shape, NPs larger than those of the training set, closely-related structures (cuboctahedra vs. truncated octahedra). Some of them can be seen as decoys, made to deceive the optimized ANNs. In other words, the capability of the ANNs to identify some a priori non obvious structural patterns is evaluated through this dataset, hereafter named a challenging dataset. It is not straightforward, in this context, to evaluate the performance of the ANNs defined and optimized in this study. Performance is often conveniently assessed through RMS deviations between target and estimated values. It is less obvious to define in such structure recognition application. This is why a gross scoring has been defined by considering for each compound $s = \sum_j \mathcal{P}_j(C_x)/N$, where $N$ is the total number of compounds submitted to the ANN (CM-ANN: $N = 27$; RDF-ANN$i$: $N = 32$) and $\mathcal{P}_j(C_x)$ is the probability associated to the expected class $C_x$.

## 3 Results

*Validation of the ANN performance.* It is done on the 11 $\times$ 20% validation subsets. The cross-validation and test scores for each of the three schemes are reported in Table 1. The RDF-based ANNs perform very well. They never fail to identify the classes the structures of the validation sets belong to. After analysis, it turns out that the CM-ANN is unable to classify 13-atoms cuboctahedra and icosahedra in the appropriate CB and IC classes. So far, it can be considered as an insignificant disagreement. Given the similarity of $g(\tilde{r})$ and $g'(\tilde{r})$ for fcc structures (see examples in Figure S4 and Figure S5), the ability of all RDF-based ANNs to discriminate between classes CB, FCC-C, FCC-S and OH is quite rewarding. On the contrary, the patterns of their CM (see Figure S6) seem different enough to unambiguously assign these compounds to our shape classification on the basis of this descriptor.

Other scores obtained with CM-ANN and RDF3-ANN are reported in Table S1 as a function of their architecture, *i.e.* the number of hidden layers and the number of neurons per hidden layer. A good validation is already obtained with 1 hidden layer made of 11 neurons, whereas 2 neurons is obviously not enough. Regarding CM-ANNs, increasing the number of neurons does not help improving the 0.96 test score.

| | CM | RDF1 | RDF2 | RDF3 |
|---|---|---|---|---|
| cv score | 0.97 | 1.00 | 1.00 | 1.00 |
| test score | 0.96 | 1.00 | 1.00 | 1.00 |

**Table 1** Test and cross-validation (cv) scores of the four considered ANN (CM and RDFi ANNs are defined in section 2).

*Classification of similar or new shapes.* The 32 structures submitted to the four ANNs are reported in Figure 4. On the contrary to the training and validation sets, most of the structures were fully optimized without any symmetry constraints, using Sutton-Chen potentials. This is what makes structure **1** ($Pt_{147}$) different from the IC-147 compound shown in Figure 2. The 32 structures are classified in nine new classes, that we shall now briefly review. The **1**-**8** Mackay icosahedra-based structures are either magic number clusters, *i.e.* with closed atomic shells (**1**, **4**, **7**, and the bimetallic core-shell **8**) or partially filled atomic shells (**2** and **3**). Compounds **5** and **6** are short metal rods made of bound icosahedral $Pt_{13}$ and $Pt_{55}$ clusters. Structures **9**-**12** are rectangular cuboid, spherical, oblate and prolate spheroid fcc clusters, respectively. Two trigonal bipyramids, **13** and **14**, were also added to the challenging dataset. Structures **15**-**21** are five-fold twinned NPs [49]. These so-called Ino's [50] and Marks' [51] decahedra have non spherical shapes, built from decahedra. Two Boerdijk-Coxeter-Bernal (BCB) helices [52–55], **22** and **23**, were also added to the challenging dataset. They are made of linearly stacked regular fcc tetrahedra, they are chiral, and they experimentally appear as either left-handed spirals or right-handed ones. Compounds **24** and **25** are bcc rhombic dodecahedra. Such structures are Catalan solids. The in-house *polyhedra* software also allows to excavate polyhedra with *emporte-pièces*. Concave cubic and icosahedral NPs **26**, **27** and **29** were made with square-pyramidal and icosahedral tips, respectively. Compound **28** is made by digging through a $Pt_{147}$ icosahedron with a cylinder. Finally, three fcc-based truncated polyhedra were added to the challenging dataset, namely two truncated octahedra (**30** and **31**) and a truncated regular tetrahedron (**32**). Given that structures **4**, **26**, **29**, **31** and **32** contain more atoms than the largest structure of the training set (*i.e.* 189 atoms), and owing to the imposed limitation of CM-ANN to $d_{\mathrm{CM}} = 189$ (see section 2), this machine cannot classify them.

The expected classification of the 32 compounds to the eleven classes known by the ANNs is subjectively color-coded in Figure 5 (green for *a priori* easy classifications,

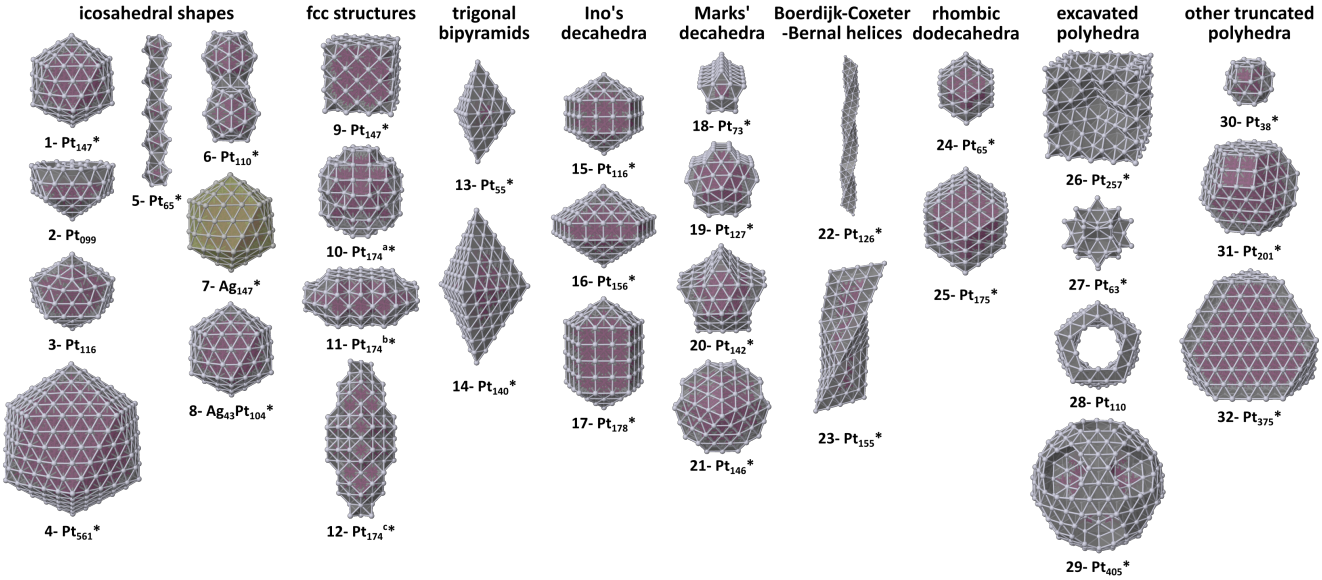


**Fig. 4** Structures of the challenging dataset. All structures marked with a * were fully optimized with Sutton-Chen potentials optimized for Pt and Ag (see Methods and datasets section). Core atoms appear in dark magenta, beneath the triangular and square facets.

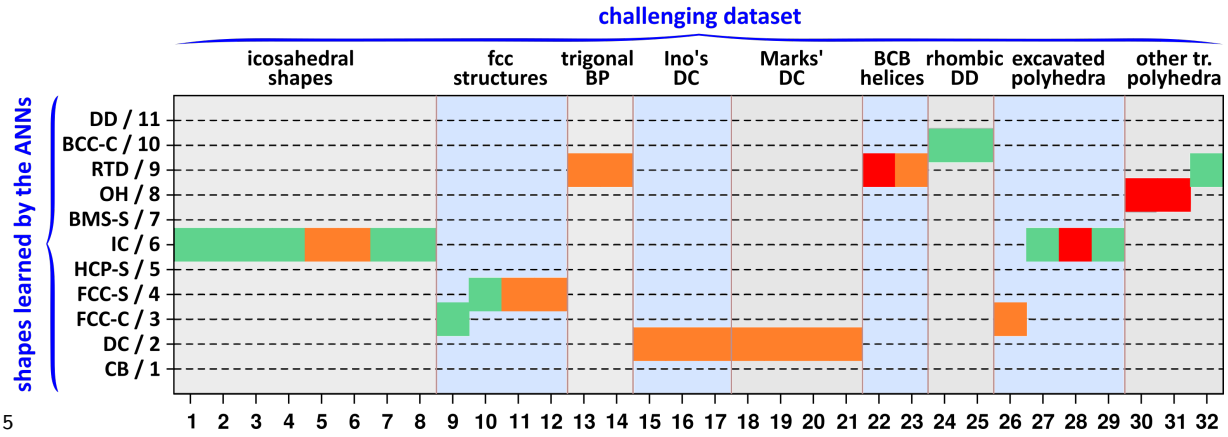


**Fig. 5** Expected classification $C_x$ of the 32 compounds that belong to the challenging dataset, into one of the eleven classes known by the ANNs. A color-coded subjective point of view is proposed: green/orange/red = easy/not that obvious/challenging classification (see justification in the text). Same classes as in Figure 4 (BP: bipyramids, DC: decahedra, BCB: Boerdijk-Coxeter-Bernal, DD: dodecahedra, tr.: truncated).

orange for less obvious classifications, red for challenging classifications). We did not include shapes that can be assigned to all eleven learned classes. As already said in section 2, some of these classes (CB, HCP-S, BMN-S, DD) are decoys that may deceive ANNs. Compounds **1**-**4**, **7**-**8** should be easily identified as icosahedra. Given their elongated shape and the repetition pattern, it may be uneasy for the CM-ANN to assign them to the IC class. Although the ANNs have only been confronted with cubic shapes, it should not be uneasy to assign compound **9** to the FCC-C class. The number of atoms of this species ($n$: 147), identical to the IC and CB magic numbers, should not be a problem. Whereas compound **10** is just another FCC-S shape, the significantly oblate and prolate shapes of **11** and **12** may perturb the ANNs. It is obvious for a reader familiar with polyhedra that the trigonal bipyramids **13** and **14** are three-fold twinned regular tetrahedra (RTD class). However, it could be less easy for the CM-ANN, and even for the RDF$i$-ANNs that may wrongfully assign them to another fcc class. Ino's decahedra (**15**-**17**) and Marks' decahedra (**18**-**21**) are five-fold twinned nanoparticle built by appropriate truncations of the edges of parent decahedra. Let us remind that NPs that exhibit decahedra shapes are five-fold multiply twinned (MTP) fcc NPs [49]. Similarly to icosahedra, they are composed of tetrahedral sub-units joined along twin boundaries and sharing axes of five-fold symmetry. In both types or structures, tetrahedra are inherently strained due to twinning. The illustration that five regular fcc tetrahedra form an imperfect decahedron with a gap of 7.35° is provided in Figure S7. Given these

comments, Ino's and Marks' decahedra may not be that easy to assign to the DC class. By the way, a reader unfamiliar with MTPs could also hesitate between the DC and IC classes. But a conveniently trained ANN is expected to do better than that. If you visualize the quite long $Pt_{126}$ BCB helix (**22**) with a molecular viewer and even if you zoom in enough and if you rotate this compound, you will see that it is not that easy to identify that this structure is an helical stacking of $Pt_{10}$ regular tetrahedra. Given also that the present ANNs have been faced to several fcc classes that do not differ that much from each other, this RTD assignment could be challenging (it is marked in red in Figure 5). The larger one, **23**, is less elongated, and the $Pt_{35}$ RTD pattern is probably more obvious in RDF profiles. The excavated fcc cube **26** is a concave structure that misses a lot of atoms. On top of that, its cubic core is also reduced to a very low number of atoms. An ANN could thus be confused with other fcc classes that are significantly different from a cube. Given that **27** has a $Pt_{55}$ IC core, its classification to the IC class should not be a problem. On the contrary, **28** does not have an IC core, and that could make its classification to the IC class very challenging. It is not immediately evident that compound **29** is an excavated icosahedron. However, it could be more obvious on the basis of its RDF profile. Its core is nothing else than **27** and the numerous surface atoms probably reinforce the IC pattern. Compounds **30** and **31** are highlighted in red in Figure 5, given the structural proximity between these truncated octahedra and cuboctahedra. The assignment could fall in either of the CB and OH classes. Finally, given that the apexes of the parent regular tetrahedron of **32** were moderately truncated, its assignment to the RTD class is expected to be obvious. Given this biased analysis, a gross score $s$ between ∼0.5 and ∼0.8 would be acceptable, whereas $s \gtrsim 0.8$ would be a remarkable outcome.

The classification of structures **1**-**32** to the eleven classes are summarized in Figure 6 and more detailed in Tables S2 to S5. The performance of each ANN can be caught at a glance: the higher the number of dark green cells, the better the performance. With a gross score $s : 0.16$, the CM-ANN miserably fails. The inability to identify compounds **2** and **3** as belonging to the IC class or compound **10** to the FCC-S class are among the most prominent failures of CM-ANN. This optimized neuron network actually shows a propensity for the DC and BMN-S classes. With gross scores between $s : 0.56$ (RDF2-ANN) and $s : 0.67$ (RDF3-ANN), the RDF*i*-ANNs are by far more adapted to this recognition task. They all succeed in assigning icosahedral species **1**-**7** to the IC class, excepted for the biggest one, **4**. This may be because long-range information of the $g(\tilde{r})$ and $g'(\tilde{r})$ profiles perturb the recognition. The core-shell icosahedron **8** is identified as a decahedron by RDF1- and RDF2- ANNs, whilst it is – luckily ? – assigned to the IC class by RDF3-ANN. Yet, the RDF profiles of decahedra and icosahedra are significantly different, and the RDF fingerprint of the core-shell AgPtNP is close to the single metal $Pt_{147}$ NP, optimized or not (see Figure S8). It is hard to understand the failure of RDF1-ANN. It is probably a data-fitting issue of the training process, that led to a bad weighting of some neural connections. It also turns out that among the compounds allegedly easy to assign, **29** is never well identified and $g'(\tilde{r})$ seem to alter the assignation of **9** to the FCC-C class (RDF2-ANN: $\mathcal{P}_9(\mathrm{FCC-C}) : 0.24$, RDF3-ANN: $\mathcal{P}_9(\mathrm{FCC-C}) : 0.02$).

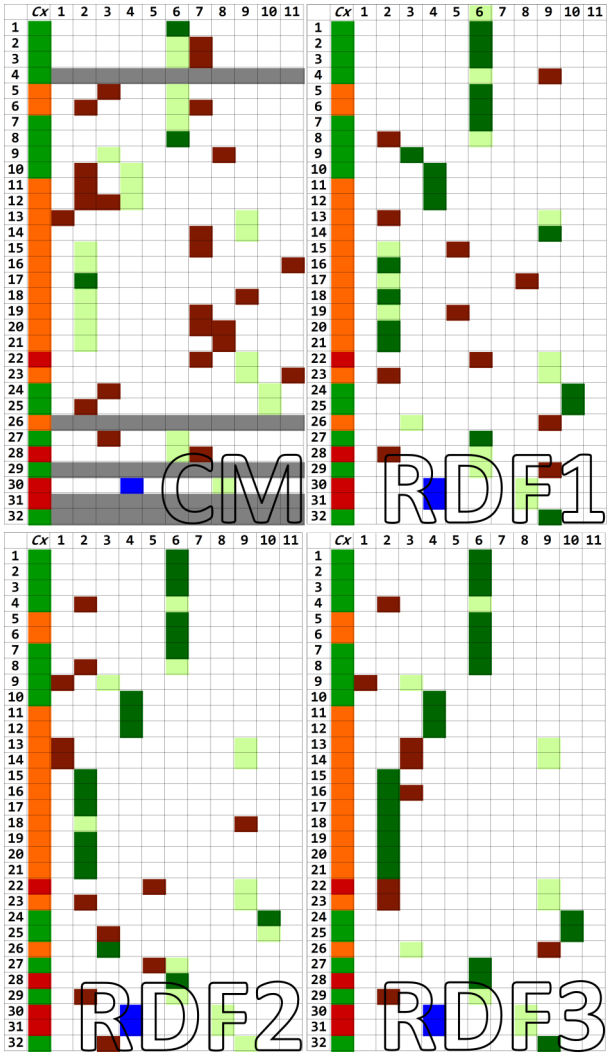


**Fig. 6** Classification of structures **1**-**32** to the eleven classes learned by the ANNs (columns: classes index, same as in Figures 2 and 4; CM and RDFi ANNs are defined in section 2, see also Table 1). Green: expected result; dark green: right class identification by the ANN; burgundy: wrong identification. The probabilities $\mathcal{P}_i(C_k)$ calculated for each class $C_k$ by the ANNs are given in the SI, Tables S2-S5. $C_x$: biased expectation on ANNs' performance, same color code as in Figure 5. Blue cells highlight answers that were not expected but that can be considered as acceptable.

Let us now consider the challenging species (in red in Figure 5). As expected, no ANN is able to assign the long BCP helix **22** to the RTD class, although a non-zero $\mathcal{P}_{22}(\mathrm{RTD})$ is found by RDF1-ANN and RDF2-ANN (0.026 and 0.020). But RDF-ANN assigns it to the HCP-S class, which has nothing in common with RTDs. Interestingly, RDF1-ANN and RDF3-ANN classify it in the DC or IC classes. Given the similar twinning patterns of tetrahedra in RTDs, ICs and DCs, it can be considered as a quite relevant identification. The RDF first derivative helps identifying the drilled icosahedron **28** to the IC class, whilst RDF1-ANN assigns it to the DC class. Finally, none of these RDF*i*-ANNs is able to see that compounds **30** and **31** are made from octahedra. They all agree to say that it is a spherical piece of an fcc packing (FCC-S class), which is after all an acceptable answer. With this tolerance being shown, the $s$ score increases to 0.20, 0.67, 0.63 and 0.73 for CM-, RDF1-, RDF2- and RDF3- ANNs, respectively.

## 4 Conclusion and outlook

We built the geometries of 30 fcc, bcc, hcp and polytetrahedral metal clusters and nano-clusters $M_n$ ($n$ = 13-189). Four different ANNs, based on Coulomb matrices or RDF profiles and their first derivative, learned to classify each compound **j** in eleven different classes $C_k$ with a given probability $\mathcal{P}_j(C_k)$. The CM-ANN and RDF*i*-ANNs all brilliantly passed the validation test, *i.e.* they successfully classified shapes very close to the training set. We then assessed the ability of the 4 ANNs to properly assign 32 additional structures to the eleven classes. Let us attempt to introduce an analogy with the cats *vs.* dogs classification, although *true* image recognition is usually best achieved in the framework of convolutional neural networks, a specialized kind of neural network for processing data that have a known grid-like topology [56, 29]. With some of these new 32 shapes, to some extent we checked the ability of the ANNs to extrapolate to other felines, big cats and canids, some of them having significant physical disabilities or wearing camouflages. Regarding the fcc-type NPs, we also evaluated the ability these ANNs to discriminate between very similar structures such as cuboctahedra and truncated octahedra. Despite such strong requirement, the RDF-based ANNs do well, with scores close to 0.6 or even 0.7 – a somewhat acceptable performance –, whereas the CM-based ANN is not adapted at all to this task, even in the "easy" cases. Given that such scoring is not related to the test set but to the challenging test, it shall be underlined that it cannot be analyzed as an overfitting or underfitting evidence. With very few exceptions, such as compound **8**, the mistakes made by the RDF*i*-ANNs can usually be rationalized in terms of atomic arrangements, a further indication that ANNs do not suffer from pathological over- or underfitting. The best ANN, namely RDF3-ANN, uses both discretized RDF profiles and their first derivative counterpart. May be owing to the explicit identification of critical points.

With this application of ANNs to structural chemistry, we are far from any DFT-based and descriptors-based assistance by machine learning to the rational design and optimization of NPs with tailored-made properties. This is a first step, that could easily find a direct application to the analysis of experimental RDF, obtained by *in situ* high-energy synchrotron X-ray diffraction or other WAXS experiments [25, 41], after taking the average number density of the material, $\rho_0$, into account [24]. Experimental RDFs would be submitted to ANNs trained on theoretical RDF profiles, possibly supplemented with a collection of well-resolved and identified experimental RDF profiles. The RDF3-ANN could also take part to the *in silico* optimization of nanocatalysts, among other descriptors (coordination numbers, chemical hardness, *d*-band center, adsorption energies, etc...). Other structural descriptors are needed to introduce coordination modes (on-top, face capping, edge-bridging...) of the surface species that stabilize metal NPs. We are currently working along these lines.

**Acknowledgements** We acknowledge the CALcul en MIdi-Pyrénées HPC (CALMIP-Olympe, grant P0611) for generous allocations of computer time. Université Paul Sabatier-Toulouse, INSAT and CNRS are also thanked for financial support. FJ and RP also thank Béatrice Laurent-Bonneau and Olivier Roustant for helpful and stimulating discussions. This article is dedicated to Fernand Spiegelmann on the occasion of his retirement. Short private message from RP: "*Fernand, je ne te remercierai jamais assez de m'avoir donné la passion de ce métier, d'avoir su aiguiser ma curiosité scientifique, d'avoir contribué à ma formation large en chimie physique et théorique, et de m'avoir appris que développer à bon escient ses propres outils est une démarche féconde*".

## References

1. G. Schmid (ed.), *Nanoparticles. From theory to application*, 2nd edn. (Wiley-VCH, Weinheim, Germany, 2010). DOI 10.1002/3527602399
2. C. Burda, X.B. Chen, R. Narayanan, M.A. El-Sayed, Chem. Rev. **105**(4), 1025 (2005). DOI 10.1021/cr030063a
3. J.P. Wilcoxon, B.L. Abrams, Chem. Soc. Rev. **35**(11), 1162 (2006). DOI 10.1039/b517312b
4. T. Mori, T. Hegmann, J. Nanopart. Res. **18**(10) (2016). DOI 10.1007/s11051-016-3587-7
5. F.H. Kaatz, A. Bultheel, Nanoscale Res. Lett. **14**(1) (2019). DOI 10.1186/s11671-019-2939-5
6. Y. Shi, Z. Lyu, M. Zhao, R. Chen, Q.N. Nguyen, Y. Xia, Chem. Rev. **121**(2), 649 (2020). DOI 10.1021/acs.chemrev.0c00454
7. W. Niu, W. Zhang, S. Firdoz, X. Lu, J. Am. Chem. Soc. **136**(8), 3010 (2014). DOI 10.1021/ja500045s
8. P. Serp, K. Philippot (eds.), *Nanomaterials in Catalysis* (Wiley-VCH, Weinheim, 2013). DOI 10.1002/9783527656875
9. V. Polshettiwar, R.S. Varma, Green Chem. **12**(5), 743 (2010). DOI 10.1039/b921171c

10. L.M. Martínez-Prieto, B. Chaudret, Acc. Chem. Res. **51**, 376 (2018). DOI 10.1021/acs.accounts.7b00378
11. I. del Rosal, R. Poteau, *Nanoparticles in Catalysis: Advances in Synthesis and Applications* (Wiley-VCH, 2021), chap. Sabatier principle and surface properties of small Ruthenium nanoparticles and clusters. Case studies., pp. 331–348
12. I. Chorkendorff, J.W. Niemantsverdriet, *Concepts Of Modern Catalysis And Kinetics* (Wiley-VCH, 2003). DOI 10.1002/3527602658
13. J.K. Nørskov, T. Bligaard, J. Rossmeisl, C.H. Christensen, Nat. Chem. **1**(1), 37 (2009). DOI 10.1038/NCHEM.121
14. J.K. Nørskov, F. Abild-Pedersen, F. Studt, T. Bligaard, Proc. Natl. Acad. Sci. USA **108**(3), 937 (2011). DOI 10.1073/pnas.1006652108
15. F. Abild-Pedersen, J. Greeley, F. Studt, J. Rossmeisl, T.R. Munter, P.G. Moses, E. Skúlason, T. Bligaard, J.K. Nørskov, Phys. Rev. Lett. **99**, 016105 (2007). DOI 10.1103/PhysRevLett.99.016105
16. J.K. Nørskov, T. Bligaard, A. Logadottir, J.R. Kitchin, J.G. Chen, S. Pandelov, U. Stimming, J. Electrochem. Soc. **152**(3), J23 (2005). DOI 10.1149/1.1856988. URL `http://dx.doi.org/10.1149/1.1856988`
17. P. Schlexer Lamoureux, K.T. Winther, J.A. Garrido Torres, V. Streibel, M. Zhao, M. Bajdich, F. Abild-Pedersen, T. Bligaard, ChemCatChem **11**(16), 3581 (2019). DOI 10.1002/cctc.201900595
18. M.O.J. Jäger, E.V. Morooka, F. Federici Canova, L. Himanen, A.S. Foster, npj Comput. Mater. **4**(1), 1 (2018). DOI 10.1038/s41524-018-0096-5. URL `https://www.nature.com/articles/s41524-018-0096-5`
19. M. Rupp, A. Tkatchenko, K.R. Müller, O.A. von Lilienfeld, Phys. Rev. Lett. **108**(5), 058301 (2012). DOI 10.1103/PhysRevLett.108.058301. URL `https://link.aps.org/doi/10.1103/PhysRevLett.108.058301`
20. H. Huo, M. Rupp, arXiv:1704.06439 [cond-mat, physics:physics] (2018). URL `http://arxiv.org/abs/1704.06439`. ArXiv: 1704.06439
21. J. Behler, J. Chem. Phys. **134**(7), 074106 (2011). DOI 10.1063/1.3553717
22. A.P. Bartók, R. Kondor, G. Csányi, Phys. Rev. B **87**(18), 184115 (2013). DOI 10.1103/PhysRevB.87.184115. URL `https://link.aps.org/doi/10.1103/PhysRevB.87.184115`
23. L. Himanen, M.O. Jäger, E.V. Morooka, F.F. Canova, Y.S. Ranawat, D.Z. Gao, P. Rinke, A.S. Foster, Comput. Phys. Commun. **247**, 106949 (2020). DOI 10.1016/j.cpc.2019.106949
24. L. Cusinato, I. del Rosal, R. Poteau, Dalton Trans. **46**, 378 (2017). DOI 10.1039/C6DT04207D
25. J.A. Vargas, V. Petkov, E.S.A. Nouh, R.K. Ramaamorthy, L.M. Lacroix, R. Poteau, G. Viau, P. Lecante, R. Arenal, ACS Nano **12**, 9521 (2018). DOI 10.1021/acsnano.8b05036
26. O.A. von Lilienfeld, R. Ramakrishnan, M. Rupp, A. Knoll, Int. J. Quant. Chem. **115**(16), 1084 (2015). DOI 10.1002/qua.24912. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/qua.24912`
27. J. Timoshenko, D. Lu, Y. Lin, A.I. Frenkel, J. Phys. Chem. Lett. **8**(20), 5091 (2017). DOI 10.1021/acs.jpclett.7b02364. URL `https://doi.org/10.1021/acs.jpclett.7b02364`
28. D. Silver, A. Huang, C.J. Maddison, A. Guez, L. Sifre, G. van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, D. Hassabis, Nature **529**(7587), 484 (2016). DOI 10.1038/nature16961
29. A. Krizhevsky, I. Sutskever, G.E. Hinton, Commun. ACM **60**(6), 84 (2017). DOI 10.1145/3065386
30. S. Cascianelli, I. Molineris, C. Isella, M. Masseroli, E. Medico, Sci. Rep. **10**(1), 14071 (2020). DOI 10.1038/s41598-020-70832-2
31. A.P. Sutton, J. Chen, Philos. Mag. Lett. **61**(3), 139 (1990). DOI 10.1080/09500839008206493
32. R. Huang, Y.H. Wen, G.F. Shao, S.G. Sun, J. Phys. Chem. C **117**(8), 4278 (2013). DOI 10.1021/jp312048k
33. J.W. Hewage, W.L. Rupika, F.G. Amar, Eur. Phys. J. D **66**(11), 282 (2012). DOI 10.1140/epjd/e2012-20691-6
34. K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O.A. von Lilienfeld, A. Tkatchenko, K.R. Müller, J. Chem. Theor. Comput. **9**(8), 3404 (2013). DOI 10.1021/ct400195d
35. T. Proffen, S. Billinge, J. Appl. Crystallogr. **32**(3), 572 (1999). DOI 10.1107/S0021889899003532
36. V. Korsunsky, Coord. Chem. Rev. **199**(1), 55 (2000). DOI 10.1016/S0010-8545(99)00171-X. URL `http://www.sciencedirect.com/science/article/pii/S001085459900171X`
37. R.V. Hardeveld, F. Hartog, Surf. Sci. **15**(2), 189 (1969). DOI 10.1016/0039-6028(69)90148-4
38. B.K. Teo, N.J.A. Sloane, Inorg. Chem. **24**(26), 4545 (1985). DOI 10.1021/ic00220a025
39. T.P. Martin, Phys. Rep. **273**, 199 (1996)
40. A.L. Mackay, Acta Crystallogr. **15**(9), 916 (1962). DOI 10.1107/s0365110x6200239x
41. F. Dassenoy, M.J. Casanove, P. Lecante, M. Verelst, E. Snoeck, A. Mosset, T.O. Ely, C. Amiens, B. Chaudret, J. Chem. Phys. **112**(18), 8137 (2000). DOI 10.1063/1.481414
42. O. Margeat, M. Respaud, C. Amiens, P. Lecante, B. Chaudret, Beilstein J. Nanotechnol. **1**, 108 (2010). DOI 10.3762/bjnano.1.13
43. F.C. Frank, J.S. Kasper, Acta Crystallogr. **11**(3), 184 (1958). DOI 10.1107/s0365110x58000487
44. F.C. Frank, J.S. Kasper, Acta Crystallogr. **12**(7), 483 (1959). DOI 10.1107/s0365110x59001499
45. T. Rapps, R. Ahlrichs, E. Waldt, M.M. Kappes, D. Schooss, Angew. Chem. Int. Ed. **52**(23), 6102 (2013). DOI 10.1002/anie.201302165
46. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, E. Duchesnay, J. Mach. Learn. Res. **12**, 2825 (2011)
47. D.P. Kingma, J. Ba. Adam: A method for stochastic optimization (2014). URL `http://arxiv.org/abs/1412.6980`. Cite arxiv:1412.6980Comment: Published as a conference paper at the 3rd International Conference for Learning Representations, San Diego, 2015
48. R. Kohavi, in *Intl. Jnt. Conf. AI* (Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1995), IJCAI 95, pp. 1137–1143
49. H. Hofmeister, in *Encyclopedia of Nanoscience and Nanotechnology*, vol. 3, ed. by H.S. Nalwa (American Scientific Publishers, Stevenson Ranch, 2004), pp. 431–452
50. S. Ino, J. Phys. Soc. Jpn. **27**(4), 941 (1969). DOI 10.1143/jpsj.27.941
51. L.D. Marks, J. Cryst. Growth **61**(3), 556 (1983). DOI 10.1016/0022-0248(83)90184-7
52. A. Boerdijk, Philips Res. Rep. **7**, 303 (1952)
53. H.S.M. Coxeter, *Regular Complex Polytopes*, 2nd edn. (Cambridge University Press, 1991)
54. J.D. Bernal, Nature **185**(4706), 68 (1960). DOI 10.1038/185068a0
55. J.J. Velaźquez-Salazar, R. Esparza, S.J. Mejiá-Rosales, R. Estrada-Salas, A. Ponce, F.L. Deepak, C. Castro-Guerrero, M. José-Yacamań, ACS Nano **5**(8), 6272 (2011). DOI 10.1021/nn202495r. URL `http://dx.doi.org/10.1021/nn202495r`

56. I. Goodfellow, Y. Bengio, A. Courville, *Deep Learning* (MIT Press, 2016). http://www.deeplearningbook.org