Eur. Phys. J. Appl. Phys. **97**, 11 (2022)

https://doi.org/10.1051/epjap/2022210262

THE EUROPEAN PHYSICAL JOURNAL APPLIED PHYSICS

Regular Article

Gaussian mixture model for the unsupervised classification of AgCu nanoalloys based on the common neighbor analysis*

Cesare Roncaglia*

Dipartimento di Fisica dell'Università di Genova, via Dodecaneso 33, Genova 16146, Italy

Received: 18 November 2021 / Received in final form: 10 January 2022 / Accepted: 11 January 2022

Abstract. In this short communication we describe the results obtained from the application of the Gaussian mixture model, a popular unsupervised learning algorithm, to some modified data sets gained after the global optimizations of three different AgCu nanoalloys. In particular we highlight both positive and negative aspects of such an approach to this kind of data. We show indeed that thanks to the Common Neighbor Analysis we are still able to describe nanoalloys well enough to exploit a physically meaningful separation in different structural families, even with a very low-dimensional representation. On the other hand, we show that the imposition of an energy cutoff over the data set is a delicate matter since it forces us to find a tradeoff between having a large set of data and having clean data.

1 Introduction

Understanding the nature of nanoalloys is a fundamental step towards a better comprehension of the world at the nanoscale. The improvement of fabrication and microscopic techniques, as well as the discovery of very important physical and chemical properties of single and bi-metallic nanoparticles (optical [1], catalytic [2], biomedical [3] properties, and many more), led the scientific community to increase the interest with respect to this very exciting field. Similarly, thanks to the impressive improvement of computational techniques and resources of the last decades, also the field of machine learning has received considerable attention. Only in the last few years, however, there has been an attempt to connect these two apparently distant areas [4,5], for the application of these algorithms requests special care, especially in cases in which only a small amount of data is available. Here in this work we try to lay the foundations of a new bridge between these two worlds, since we believe that this could open up many interesting possibilities. In particular here we show and discuss the application of a probabilistic unsupervised learning algorithm used to detect different groups of nanoalloy structures inside a data set of structures. We decided to use AgCu as a system, since it has already exhibited its important properties in catalysis and plasmonics and it is therefore interesting to have new theoretical insights about it; but also because the previous simulations of this system have recovered experimental results as well as more sophisticated calculations (such as density functional theory calculations) [6–8], so that we are more confident about the numerical results obtained by the atomistic force field used in this paper. It is important to note however that we strongly believe that the method used in this work is intrinsically general so that the results can be applied to a large variety of systems.

We note that in the following the term *cluster* is always used to denote a group of nanoparticle structures as obtained by an unsupervised learning algorithm.

The material is organized as follows. Section 2 is dedicated to the theoretical methods and models; in Section 3 the results obtained are shown and discussed. Finally Section 4 concludes the article.

2 Methods

2.1 Atomistic force field and global optimizations

In order to describe interactions between silver and copper atoms, we used an atomistic force field, which is known as Gupta potential [9] and that can be derived from the second moment approximation to the tight-binding model [10]. In this form, the potential energy of a nanoparticle can be expressed as a sum of repulsive and attractive terms, over all atoms. We used parameters from reference [11]. Global optimizations were then performed with the basin hopping algorithm, using our own code [12].

2.2 Clustering

Before performing the clustering algorithm that we will describe below, we first implemented the Common Neighbor Analysis (CNA) [13] in order to describe in a simpler manner our data set of AgCu nanoalloys, rather than using all their $3N$ spatial coordinates. For each pair of nearest-neighbor atoms, the CNA defines a *signature* consisting of a triplet of integer numbers rst:

* Special Issue on 'Nanoalloys: Kinetic and Environmental Behaviour', edited by Pascal Andreazza, Riccardo Ferrando and Liu Xiaoxuan

* e-mail: roncaglia@fisica.unige.it

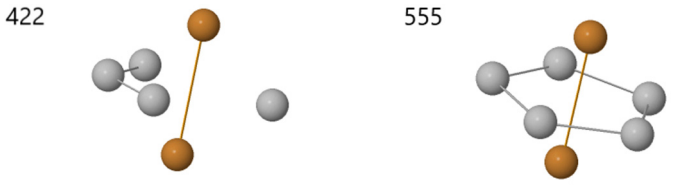

Fig. 1. 422 signature, typical of FCC fragments with local HCP zones (stacking faults, twin planes), and 555 signature, typical of atom pairs along 5-fold symmetry axes (such as those of icosahedra and decahedra).

- r - The number of common nearest neighbors of the pair.
- s - The number of bonds between those r atoms.
- t - The length of the longest chain of bonds that can be made out of the s bonds present.

For a given nanoalloy, we define its *signature* order parameter as the number of nearest-neighbor pairs presenting such *signature* divided by the total number of nearest-neighbor pairs in the nanoalloy. Following reference [14], here we decided to use first a two-dimensional description based on the (555,422) pair of signatures, shown in Figure 1. Thus, we have reduced the dimension of each nanoparticle description from $3N$ to 2.

In order to perform clustering, we decided to use the Gaussian Mixture Model (GMM), implemented with Scikit-learn [15], a very popular approach in unsupervised learning [16,17]. This probabilistic model assumes that the observed data set has been generated according to some Gaussian probability distributions with unknown parameters (weights, means and covariance matrices). The goal of the model is then to find an approximation of these parameters, according to some criteria. In particular, a multi-dimensional Gaussian probability density has the form

$$N(x;\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)} \quad (1)$$

where μ and Σ are the mean and the covariance matrix of the distribution, respectively, and d is the dimension. If the data set consists of $x_1, x_2, ..., x_n$, the problem can be formulated in terms of maximizing the log-likelihood of a model defined by a mixture of K multi-dimensional Gaussian distributions, that is:

$$\ell(\theta) = \log \prod_{i=1}^{n} \sum_{k=1}^{K} \pi_k N(x_i;\mu_k,\Sigma_k) \quad (2)$$

where $\theta = (\pi_1, ..., \pi_K, \mu_1, ..., \mu_K, \Sigma_1, ..., \Sigma_K)$ is the vector of all parameters and where π_k is the weight of the k-th distribution in the mixture model. The solution is given by the vector θ^* which maximizes ℓ. Notice that θ has $K(d^2 + 3d + 2)/2 - d$ components, so that a reduction of the dimension d is of some interest. In practice the problem is solved by the Expectation-Maximization algorithm, which iteratively finds an approximate solution to $\theta^*(K)$. In order to select the best possible K, that is the best number of clusters in the data set (i.e. the number of different families in which it is possible to assign each point), some rule must be followed. Here we decided to use the Bayesian Information Criterion (BIC) [18], which is defined as follows:

$$\mathrm{BIC} = p \cdot \log(n) - 2\log(\hat{L}) \quad (3)$$

where n is the number of points in the data set, p the number of parameters of the mixture model and $\hat{L}$ is the maximized value of the likelihood function. The best number of clusters K is the one that minimizes this score. We note that this choice is somehow arbitrary, so that in principle a different scoring function could lead to a different result.

3 Results

We considered AgCu nanoalloys composed of $N = 200$ atoms, for different compositions. In order to construct our data sets, we performed global optimizations and collected various structures coming from the exploration and local minimization of the potential energy surface defined by the Gupta potential. In particular here we selected three different compositions: Ag_0Cu_{200}, $Ag_{90}Cu_{110}$ and $Ag_{104}Cu_{96}$. For all these three compositions we calculated the CNA 422 and 555 signature order parameters for all nanoparticles. Finally, we found the best clustering for each case, by looking at the BIC scores given by the different GMM fits. As already discussed in reference [14], it can be useful to impose an energy cutoff over the entire data set in order to exclude very disordered structures (i.e. high energy local minima, almost liquid-like nanoalloys) that could be very difficult to distinguish one from another even at human-eye level. Here we discuss the choice of different cutoffs.

3.1 Ag_0Cu_{200}

For pure copper, our lowest-energy structure is a Marks decahedron. The exploration of the potential energy surface in this case was able to bring out also other structures that we will describe below. In particular thanks to our basin hopping code we could collect some hundreds of local minima visited during such exploration. The representation of these local minima in the two dimensional space of 422 and 555 signature order parameters is given in Figure 2a. GMM was then applied for different values of K (i.e. the number of gaussian distributions to describe this data set). The BIC scores for such models is reported in Figure 2b. The optimal number of clusters is $K = 4$, the one that minimizes the BIC score. In the left panel it is shown the result of the fit for four distributions. Different colors represent different clusters, and so they represent different structural families. Manual inspection of this result revealed that:

- the *black cluster* – fcc structures with some islands in stacking faults (which we will denote as fcc/hcp)
- the *purple cluster* – decahedral structures
- the *yellow cluster* – decahedral structures with one anti-Mackay crust, and multiple fivefold axes [19]

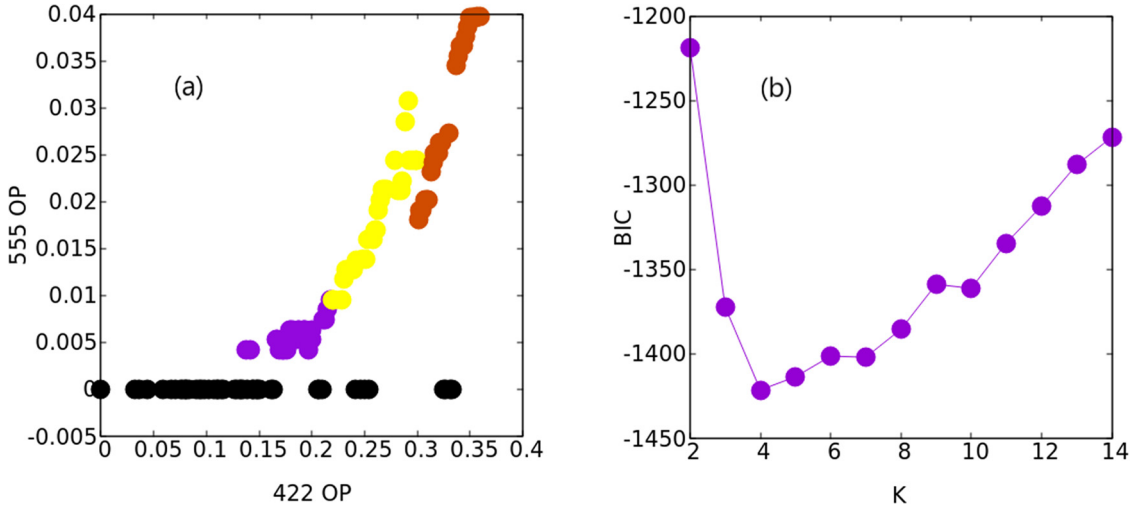

Fig. 2. Ag_0Cu_{200} (a) local minima representation in the two dimensional space of 422 and 555 signature order parameters and (b) the BIC score for different values of K.

– the *red cluster* – icosahedral structures with incomplete Mackay crust.

For this data set, a cutoff of 1.5 eV above the global minimum was set. This operation reduced the number of local minima from 163 to 126. In this case we could not proceed with the standard choice of 0.5 eV because the resulting number of structures to be considered would have been too small.

3.2 $Ag_{90}Cu_{110}$

For this composition, our global minimum is a Cu@Ag icosahedral nanoalloy with an anti-Mackay crust [20–22]. The same procedure for the previous case led to a separation of the local minima in three different clusters:

– icosahedral Cu@Ag nanoalloys with Mackay crust with some chiral distortions
– icosahedral Cu@Ag nanoalloys with mixed Mackay and anti-Mackay crust
– icosahedral Cu@Ag nanoalloys with anti-Mackay crust.

For this data set, a cutoff of 0.5 eV above the lowest local minimum was set. In this case the number of structures was reduced from 178 to 27. We noticed that increasing the cutoff to 3 eV (thus considering 148 structures) made the clustering worse in terms of clarity and distinction between the different structural families: the addition of high-energy isomers led indeed to an unavoidable dispersion of the essential properties of each cluster. In particular two clusters obtained with the 0.5 eV cutoff were merged together. However, thanks to this higher cutoff it was also possible to detect other two families otherwise discarded: fcc/hcp and multi-decahedral Cu@Ag nanoalloys. From this evidence it is clear that the choice of the cutoff before the application of the GMM is crucial for the final result. If the cutoff is too small, the data set will be represented in a clearer way, but still there will be a chance of having too few structures to analyze from a reasonable statistical perspective. On the other hand, while having the benefit of including more structures, a too large cutoff could lead to a wrong division in the plane of the signatures since high-energy local minima tend to blur the distributions. It is therefore necessary to find a tradeoff between between these two needs. We believe that a cutoff of the order of 0.5 eV is in general a good choice for the global optimization of nanoalloys in this size range, and that it is far more preferable to force the algorithm responsible for the exploration of the potential energy surface to increase the number of collected local minima rather than increase the cutoff in order to include more structures.

3.3 $Ag_{104}Cu_{96}$

For this composition, our lowest energy structure is a Cu@Ag fcc/hcp nanoalloy, as shown in Figure 3. With a cutoff of 0.5 eV, the same procedure applied in the previous cases led to a separation of the data set in two clusters:

– fcc/hcp Cu@Ag nanoalloys
– icosahedral Cu@Ag nanoalloys with chiral crust [23, 24].

In particular, using this cutoff reduced the data set from 188 to 19, and the two clusters are really one cluster and one outlier structure (i.e. one point in the CNA plane which is very far from the others). As it can be seen in Figure 4, different choices for the cutoff force us to only consider a fraction of the potential energy surface. In particular the benefit that comes from this choice of selecting only structures that are lower in energy than some reference, allows us not to consider those barriers which often divide two regions representing different structural families in the CNA space. In this case it is evident that the choice of 0.5 eV is high enough to get a point from the other potential well separated by a barrier. It must be noted however that this trend for the potential energy

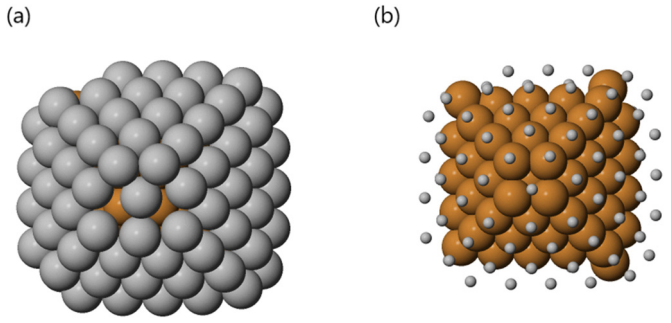

Fig. 3. $Ag_{104}Cu_{96}$ (a) top view of the global minimum and (b) the same view but with smaller silver atoms to better see the Cu core.

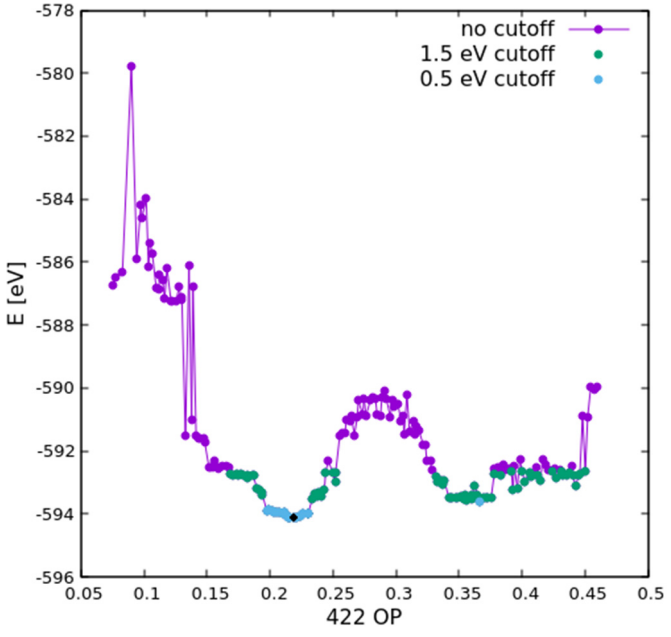

Fig. 4. $Ag_{104}Cu_{96}$ evolution of the potential energy as a function of the 422 signature order parameter. Different colors refer to different cutoffs; the black point is the lowest local minimum.

as a function of some order parameter, although quite general, may be different or even absent for some global optimizations. In these cases the plot given in Figure 4 may look much more complicated, thus making even more difficult the interpretation of the results of the clustering (for example, the potential wells could be wider and they could be separated by smaller barriers).

4 Conclusions

In this short communication we showed that the application of a probabilistic model for unsupervised learning of nanoparticles is indeed a real possibility. In particular, thanks to global optimization searches, we first collected three different data sets of AgCu nanoalloys with same size but different compositions, and then we applied successfully this Machine Learning technique in order to separate such sets in different structural families. We also discussed the choice of imposing an energy cutoff to our data, a delicate operation which is, at the present stage, almost necessary.

Further work in this direction can be important for different reasons. The primary importance of a successful application of an unsupervised learning algorithm of this kind, is that it automatically separates data in a physically meaningful way, and it does it without having to look at them. This allows to save time otherwise spent at manually inspecting many structures. Another important advantage is that, especially for large data sets (for example those of Molecular Dynamics), a human classification could fail at detecting some subtle differences in nanoparticles. We believe that a careful design of a more complex description of nanoparticles, still based on CNA but with the use of a larger number of signatures, could solve this problem by making the automatic classification of nanoalloys even better in terms of precision and stability.

The author acknowledges support from the PRIN 2017 project UTFROM of the Italian MIUR, from the Progetto di Eccellenza of the Physics Department of the University of Genoa, networking support from the IRN Nanoalloys of CNRS and thanks Prof. Riccardo Ferrando for useful discussions.

References

1. K.B. Mogensen, K. Kneipp, J. Phys. Chem C **118**, 28075 (2014)
2. M.B. Gawande, A. Goswami, F.X. Felpin, T. Asefa, X. Huang, R. Silva, X. Zou, R. Zboril, R.S. Varma, Chem. Rev. **116**, 3722 (2016)
3. K. McNamara, S.A.M. Tofail, Adv. Phys. X **2**, 54 (2017)
4. R. Jinnouchi, H. Hirata, R. Asahi, J. Phys. Chem. C **121**, 26397 (2017)
5. H. Kurban, Chem. Phys. **545**, 111143 (2021)
6. C. Langlois, Z.Y. Li, J. Yuan, D. Alloyeau, J. Nelayah, D. Bochicchio, R. Ferrando, C. Ricolleau, Nanoscale **4**, 3381 (2012)
7. M. Snellman, N. Eom, M. Ek, M.E. Messing, K. Deppert, Nanoscale Adv. **3**, 3041 (2021)
8. D. Bochicchio, R. Ferrando, Nano Lett. **10**, 4211 (2010)
9. R.P. Gupta, Phys. Rev. B **23**, 6265 (1981)
10. F. Cyrot-Lackmann, F. Ducastelle, Phys. Rev. B **4**, 2406 (1971)
11. F. Baletto, C. Mottet, R. Ferrando, Phys. Rev. Lett. **90**, 135504 (2003)
12. G. Rossi, R. Ferrando, J. Phys.: Condens. Matter **21**, 084208 (2009)
13. D. Faken, H. Jónsson, Comput. Mater. Sci. **2**, 279 (1994)
14. C. Roncaglia, D. Rapetti, R. Ferrando, Phys. Chem. Chem. Phys. **23**, 23325 (2021)
15. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg et al., J. Mach. Learn. Res. **12**, 2825 (2011)
16. C.M. Bishop, *Pattern Recognition and Machine Learning*, Information Science and Statistics (Springer, 2006)
17. K.P. Murphy, *Machine Learning: A Probabilistic Perspective* (The MIT Press, 2012)

18. T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer Series in Statistics (Springer, 2009)
19. G. Rossi, R. Ferrando, Nanotechnology **18**, 225706 (2007)
20. D. Bochicchio, R. Ferrando, Phys. Rev. B **87**, 165435 (2013)
21. E. Panizon, R. Ferrando, Nanoscale **8**, 15911 (2016)
22. D. Nelli, R. Ferrando, Nanoscale **11**, 13040 (2019)
23. M. Settem, A.K. Kanjarla, Comput. Mater. Sci. **184**, 109822 (2020)
24. M. Settem, J. Alloys Compd. **844**, 155816 (2020)

Cite this article as: Cesare Roncaglia, Gaussian mixture model for the unsupervised classification of AgCu nanoalloys based on the common neighbor analysis, Eur. Phys. J. Appl. Phys. **97**, 11 (2022)