Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.

Supplementary Information

Regression and clustering algorithms for AgCu nanoalloys: from energy predictions to structure recognition

Cesare Roncaglia[1], Daniele Rapetti[1] and Riccardo Ferrando[2]

[1] *Dipartimento di Fisica dell'Università di Genova, via Dodecaneso 33, Genova 16146, Italy*

[2] *Dipartimento di Fisica dell'Università di Genova and CNR-IMEM, via Dodecaneso 33, Genova 16146, Italy*

1 DFT results on the mixing energy

We confirm our claims about the negativity of the mixing energy here by reporting the plot of DFT calculations at some compositions in fig. 1. The DFT results confirm the overall behavior of the mixing energy giving values that are even more negative than those obtained by the Gupta potential. In our DFT calculations we used the PBE exchange-correlation functional[1]. Calculations were performed by the Quantum Espresso software[2] with the two pseudopotentials Ag.pbe-n-kjpaw_psl.1.0.0.UPF and Cu.pbe-dn-kjpaw_psl.1.0.0.UPF taken from http://www.quantum-espresso.org. The structures were relaxed with a simple cubic cell of side 30 Å and cut offs for wave functions and charge density were chosen to be 45 Ry and 236 Ry respectively.

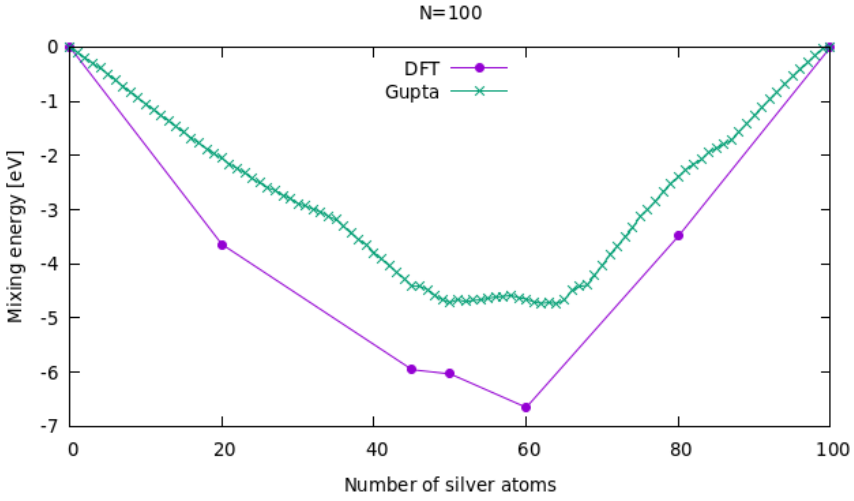


Figure 1: DFT calculations of the mixing energy for AgCu nanoalloys with $N = 100$ atoms for $m = 0, 20, 45, 50, 60, 80$ and 100 silver atoms, along with mixing energies calculated by the Gupta potential.

2 Relationship between K-means classification and mixing energy

For $N = 100$, we studied the relationship between the mixing energy and the classification as given by the unsupervised learning algorithm K-Means. We find that each structural family corresponds to a well-defined interval of mixing energies, according to the fact that, as described at the end of subsection 3.1 in main text, different clusters correspond to different intervals of compositions, without any random jump in the compositions. The plot of the mixing energy with different colors for different clusters is represented in fig. 2.

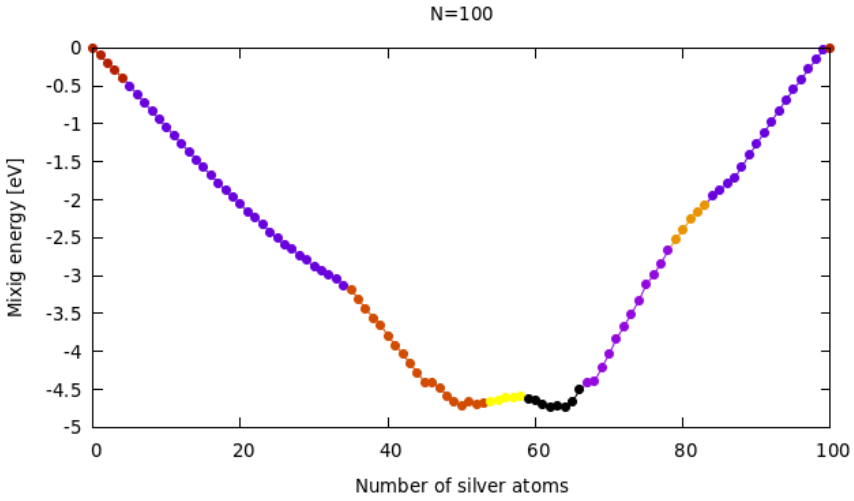


Figure 2: The mixing energy profile for AgCu nanoalloys with $N = 100$, with different colors. Each color refers to a different cluster as given by K-means, as described in subsection 3.1 of the main text.

3 Machine Learning models parameters

3.1 $N = 100$ - Regression

Here we list the parameters for the models fitting the mixing energy, as given by the Scikit-Learn implementation of SVR. In particular the model with gaussian kernel, is:

$$E(x) = \sum_{i=1}^{N_{sv}} (\alpha_i - \alpha_i^*) e^{-\gamma(x-x_i)^2} + b \tag{1}$$

where the x_i are the coordinates of the N_{sv} support vectors (in this case they are integers corresponding to some particular compositions, i.e. number of silver atoms), $\alpha_i - \alpha_i^*$ are the dual coefficients and b is the intercept.

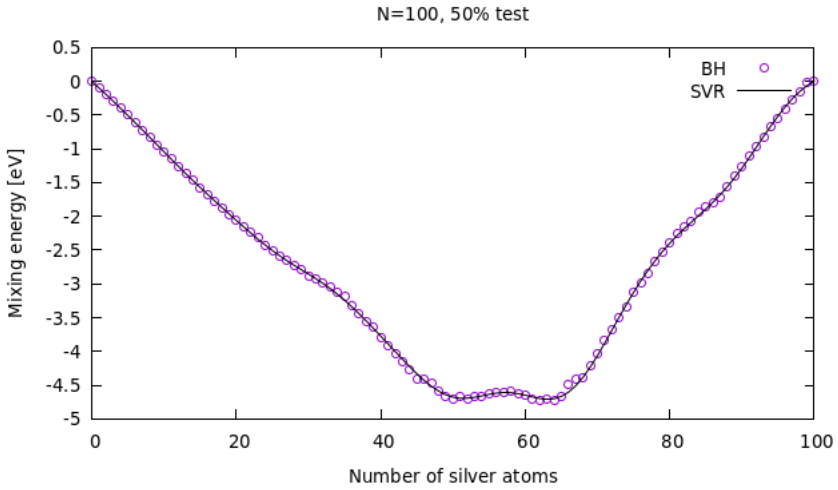
3.1.1 80-20 splitting

The best hyperparameters are: C=50, ϵ=0.01, γ=0.01
The number of support vectors is $N_{sv} = 48$. They are:
x_i = 68, 46, 100, 99, 5, 37, 42, 77, 27, 50, 63, 86, 84, 88, 79, 0, 30, 96, 38, 35, 44, 10, 2, 58, 71, 65, 15, 33, 19, 57, 62, 41, 48, 87, 64, 69, 75, 66, 49, 45, 51, 25, 90, 39, 91, 54, 73, 83
The non zero coefficient in the support vector expansion are:
$\alpha_i - \alpha_i^*$ = -50, 50, -25.0491547, 50, 10.9612307, -50, 5.09702542, -46.75985597, 50, -50, 50, -8.39496924, 50, -23.36440404, 5.65337911, 13.069987, -50, -38.5463711, -39.97069796, 50, -13.78177396, -3.49775963, -19.0974567, 50, 50, 19.00456702,

2.16461365, 12.81171681, 0.41309314, -31.71899986, -50, 7.15579358, 50, -50, -50, -15.72424457, 50, 50, -44.39422285, -50, 50, -23.43549626, 50, 50, 10.25674516, -10.22770178, -39.35329699, -3.27174596
The value of the intercept is: b = -2.20347226

3.1.2 50-50 splitting

The best hyperparameters are: C=1000, ϵ=0.005, γ=0.005
The number of support vectors is $N_{sv} = 41$. They are:
x_i = 13, 3, 17, 38, 8, 79, 65, 36, 88, 56, 100, 54, 50, 68, 46, 69, 61, 98, 80, 41, 58, 48, 90, 57, 95, 63, 85, 37, 29, 52, 21, 23, 87,99, 74, 86, 82, 20, 71, 92, 51
The non zero coefficient in the support vector expansion are:
$\alpha_i - \alpha_i^* =$ 260.48117464, 28.88777125, -284.6944565, -572.74259947, -111.07578008, -449.53082293, -6.93715845, 1000, -735.14710421, 91.02484193, -475.96794941, -1000, -1000, -1000, 1000, 1000, -1000, -242.63080093, 58.3001304, -1000, 1000, 465.90717086, 1000, 1000, -1000, -133.75181755, 1000, 264.67775736, -632.53134579, -961.29718532, -1000, 1000, -1000, 1000, -1000, -891.39900903, 1000, 426.25304986, 1000, 902.17413338, 1000
The value of the intercept is: b = -1.9449444



3.2 $N = 100$ - Clustering

The means of each cluster found by the K-means algorithm implemented by Scikit-Learn, when K=7, are:

$$m_1 = (0.473568, 0.140969)$$
$$m_2 = (0.373829, 0.062394)$$
$$m_3 = (0.193622, 0.009112)$$

$$m_4 = (0.277986, 0.139389)$$
$$m_5 = (0.434499, 0.133748)$$
$$m_6 = (0.259078, 0.044254)$$
$$m_7 = (0.447368, 0.171053)$$

in the two dimensional space of 422 and 555 signature order parameters.

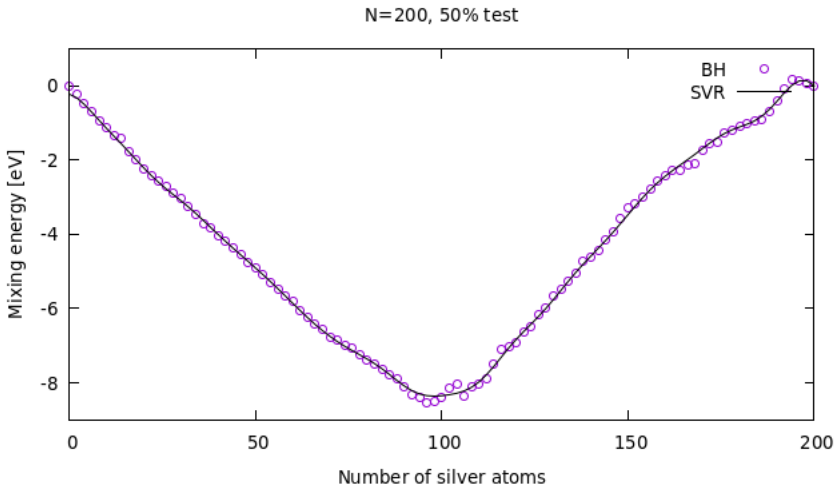
3.3 N =200 - Regression

3.3.1 80-20 splitting

The best hyperparameters are: C=50, ϵ=0.01, γ=0.001
The number of support vectors is $N_{sv} = 76$. They are:
x_i = 56, 10, 92, 144, 158, 12, 102, 30, 178, 50, 68, 130, 170, 104, 184, 8, 14, 80, 64, 154, 54, 116, 82, 128, 0, 124, 192, 156, 152, 174, 36, 138, 134, 22, 100, 38, 132, 118, 84, 58, 32, 90, 96, 122, 136, 142, 188, 42, 28, 86, 70, 46, 88, 172, 94, 162, 164, 4, 48, 6, 106, 198, 74, 24, 176, 76, 112, 18, 44, 98, 40, 110, 180, 146, 150, 108
The non zero coefficient in the support vector expansion are:
$\alpha_i - \alpha_i^*$ = 50, 50, -50, -50, 50, 2.98026768, 50, 50, 50, 50, -50, 50, -50, 50, -50, -50, 50, -46.53661166, -50, -31.87828558, 50, 50, 50, -50, 43.19106288, -50, 50, 50, 50, -50, -50, 50, -50, -50, 50, 34.6107961, 50, 50, 50, 50, 50, -23.38110505, -50, -1.73081023, 21.21297419, -50, -50, -3.47336526, 50, 50, -50, -17.11382658, 50, 50, -50, -44.17617668, -50, -10.15446417, -50, -50, -50, -2.64365895, -24.67181788, -50, 50, 50, -50, -6.69679316, -50, -50, -50, -50, 19.78770691, -50, 50, 40.67410745
The value of the intercept is: b = -2.99463209

3.3.2 50-50 splitting

The best hyperparameters are: C=10, ϵ=0.001, γ=0.005
The number of support vectors is $N_{sv} = 50$. They are:
x_i = 94, 8, 176, 108, 58, 192, 194, 104, 154, 52, 32, 140, 162, 200, 36, 68, 84, 64, 180, 124, 2, 12, 172, 138, 78, 118, 90, 18, 160, 38, 6, 42, 122, 152, 190, 184, 178, 88, 196, 114, 144, 112, 188, 22, 128, 126, 66, 110, 30, 46
The non zero coefficient in the support vector expansion are:
$\alpha_i - \alpha_i^*$ = -10, -10, 10, -2.25791822, 7.08168716, 10, 10, 10, -10, -5.63445682, 0.21546008, -10, 4.90514214, 1.48849082, -10, 10, -3.54070112, -5.51255695, 4.16583064, -10, 10, 10, -4.71619456, 10, -3.6501727, 10, -2.10673057, 5.00442999, -1.17792878, 0.46365526, -4.75441234, 5.65938074, -6.00800092, 10, -10, 8.52785891, -10, 10, -5.2285789, 10, -2.28382033, -10, -10, -10, -4.11658301, 10, -8.32573316, -10, 10, 1.80185267
The value of the intercept is: b = -3.86084462



3.4 $N = 200$ - Clustering

The means of each cluster found by the K-means algorithm implemented by Scikit-Learn, when K=5, are:

$$m_1 = (0.300443,\ 0.030700)$$
$$m_2 = (0.446679,\ 0.075048)$$
$$m_3 = (0.180112,\ 0.005834)$$
$$m_4 = (0.358179,\ 0.049046)$$
$$m_5 = (0.239096,\ 0.001020)$$

in the two dimensional space of 422 and 555 signature order parameters.

3.5 m =64, n =36

GMM is a probabilistic model. The algorithm implemented by Scikit-Learn infers the parameters from a mixture of gaussian distributions, i.e. centers and covariance matrices, and their weights.

3.5.1 Full data set

The best seven two dimensional gaussian distributions are described by the following parameters.

Centers:

$$\mu_1 = (0.11572449,\ 0.11809805)$$
$$\mu_2 = (0.40272652,\ 0.12600916)$$
$$\mu_3 = (0.05052110,\ 0.29737415)$$

$$\mu_4 = (0.25706954,\ 0.14213907)$$
$$\mu_5 = (0.18111922,\ 0.08351079)$$
$$\mu_6 = (0.43967932,\ 0.17555897)$$
$$\mu_7 = (0.34365538,\ 0.16413533)$$

Covariance matrices:

$$\Sigma_1 = \begin{pmatrix} 6.59619905 \cdot 10^{-4} & 8.13361909 \cdot 10^{-5} \\ 8.13361909 \cdot 10^{-5} & 1.19122195 \cdot 10^{-3} \end{pmatrix}$$
$$\Sigma_2 = \begin{pmatrix} 2.55754116 \cdot 10^{-4} & 9.67081250 \cdot 10^{-5} \\ 9.67081250 \cdot 10^{-5} & 6.61988936 \cdot 10^{-5} \end{pmatrix}$$
$$\Sigma_3 = \begin{pmatrix} 5.36384822 \cdot 10^{-4} & -9.81284325 \cdot 10^{-4} \\ -9.81284325 \cdot 10^{-4} & 3.04791709 \cdot 10^{-3} \end{pmatrix}$$
$$\Sigma_4 = \begin{pmatrix} 1.06581424 \cdot 10^{-3} & 1.87035004 \cdot 10^{-4} \\ 1.87035004 \cdot 10^{-4} & 1.85388871 \cdot 10^{-4} \end{pmatrix}$$
$$\Sigma_5 = \begin{pmatrix} 5.08141189 \cdot 10^{-4} & -1.77860392 \cdot 10^{-4} \\ -1.77860392 \cdot 10^{-4} & 3.37814959 \cdot 10^{-4} \end{pmatrix}$$
$$\Sigma_6 = \begin{pmatrix} 2.50070518 \cdot 10^{-4} & -2.29211809 \cdot 10^{-4} \\ -2.29211809 \cdot 10^{-4} & 2.79921325 \cdot 10^{-4} \end{pmatrix}$$
$$\Sigma_7 = \begin{pmatrix} 7.18838914 \cdot 10^{-4} & 1.96743451 \cdot 10^{-5} \\ 1.96743451 \cdot 10^{-5} & 1.34704139 \cdot 10^{-4} \end{pmatrix}$$

The weight of each distribution in the mixture model is given by:
$w_i = 0.16074789,\ 0.07776874,\ 0.13324705,\ 0.21660087,\ 0.13157819,\ 0.099401,\ 0.18065626$

3.5.2 Reduced data set

The best four two dimensional gaussian distributions are described by the following parameters.

Center:

$$\mu_1 = (0.42788094,\ 0.18825707)$$
$$\mu_2 = (0.27682917,\ 0.14463879)$$
$$\mu_3 = (0.34710199,\ 0.16676328)$$
$$\mu_4 = (0.40635834,\ 0.1259776)$$

Covariance matrices:

$$\Sigma_1 = \begin{pmatrix} 1.45664269 \cdot 10^{-4} & -7.38570045 \cdot 10^{-5} \\ -7.38570045 \cdot 10^{-5} & 1.32940476 \cdot 10^{-4} \end{pmatrix}$$

$$\Sigma_2 = \begin{pmatrix} 4.69738985 \cdot 10^{-4} & 1.04830663 \cdot 10^{-4} \\ 1.04830663 \cdot 10^{-4} & 1.00347615 \cdot 10^{-4} \end{pmatrix}$$

$$\Sigma_3 = \begin{pmatrix} 5.84755957 \cdot 10^{-4} & -2.36759604 \cdot 10^{-5} \\ -2.36759604 \cdot 10^{-5} & 1.02646792 \cdot 10^{-4} \end{pmatrix}$$

$$\Sigma_4 = \begin{pmatrix} 1.95157287 \cdot 10^{-4} & 7.01362581 \cdot 10^{-5} \\ 7.01362581 \cdot 10^{-5} & 6.35548583 \cdot 10^{-5} \end{pmatrix}$$

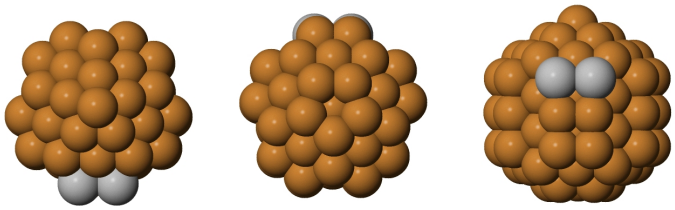
The weight of each distribution in the mixture model is given by:
$w_i = 0.13292409, 0.47049398, 0.28552936, 0.11105257$

4 Some structures from global optimizations

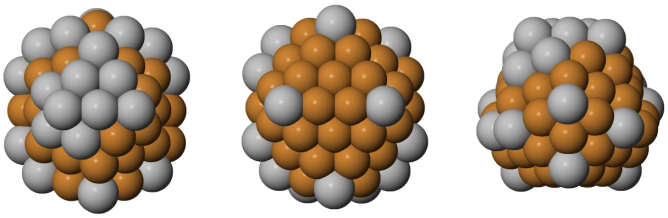
Below we show top (left), bottom (center) and side (right) view of one representative for each class found by K-Means, for both data set. Sometimes the side view is replaced with an inner view.

4.1 N =100

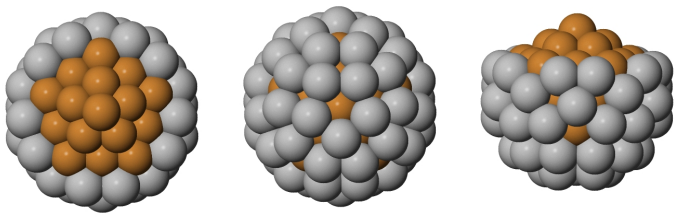
- m =2, n =98



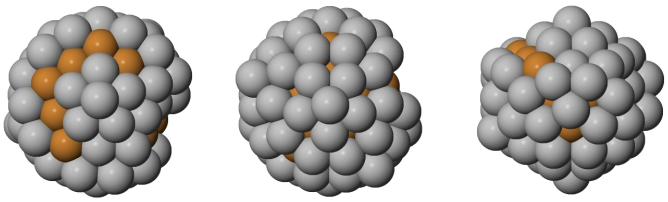
- m =23, n =77



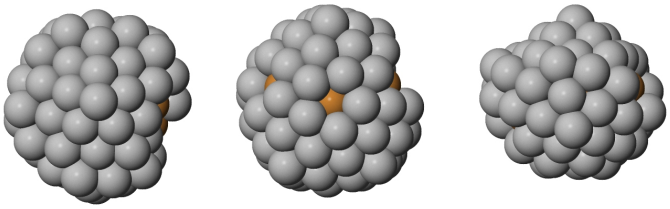
- m =45, n =55



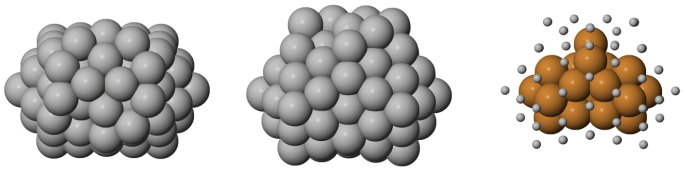
- m =56, n =44



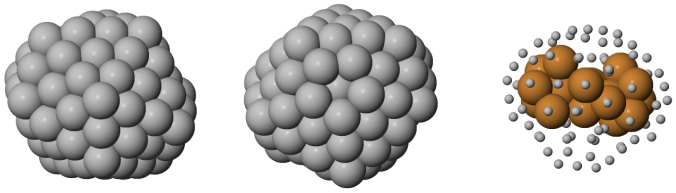
- m =62, n =38



- m =72, n =28

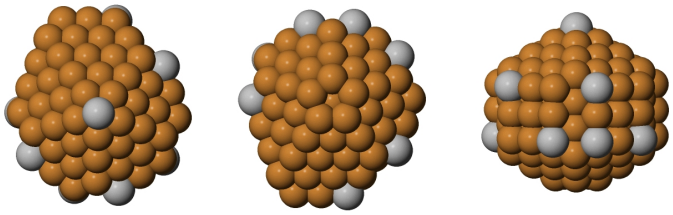


- m =82, n =18

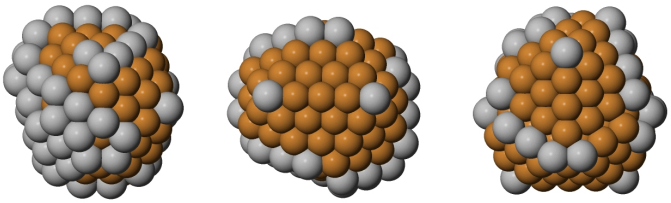


4.2 N =200

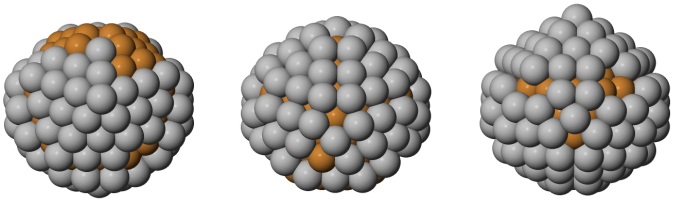
- m =10, n =190



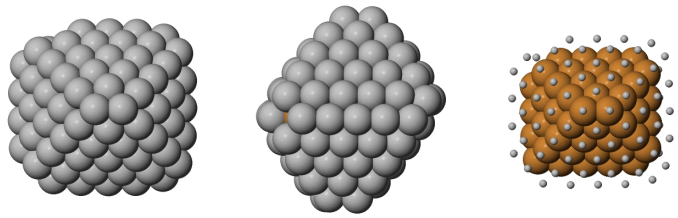
- m =42, n =158



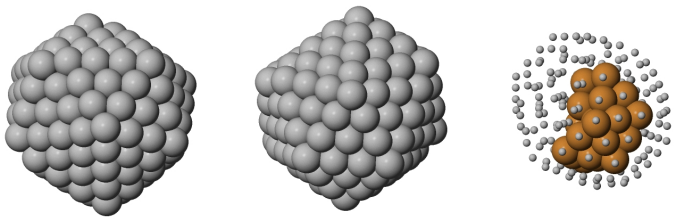
- m =80, n =120



- m =106, n =94



- m =168, n =32



References

[1] J. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1997, **78**, 1396–1396.

[2] P. Giannozzi, S. Baroni, N. Bonini, M. Calandra, R. Car, C. Cavazzoni, D. Ceresoli, G. L. Chiarotti, M. Cococcioni, I. Dabo, A. D. Corso, S. de Gironcoli, S. Fabris, G. Fratesi, R. Gebauer, U. Gerstmann, C. Gougoussis, A. Kokalj, M. Lazzeri, L. Martin-Samos, N. Marzari, F. Mauri, R. Mazzarello, S. Paolini, A. Pasquarello, L. Paulatto, C. Sbraccia, S. Scandolo, G. Sclauzero, A. P. Seitsonen, A. Smogunov, P. Umari and R. M. Wentzcovitch, *J. Phys. Condens. Matter*, 2009, **21**, 395502.