Using a Self Organising Map for Clustering of Atomistic Samples

Aquistapace F.

Facultad de Ciencias Exactas y Naturales, UNCuyo

September 23, 2021

Introduction

This software implements a Self Organising Map (also known as a Kohonen Network) for clustering of atomistic samples through unsupervised learning. It is written in Python 3.8.8 and depends on the external packages NumPy (version 1.20.1) [1] and Pandas (version 1.2.4) [2]. The core of the algorithm is a neural network composed of an input layer and an output layer, where the number of output neurons determines the number of groups the atoms are going to be classified into. This software is based on the work by J. Troncoso on applying a SOM for cluster analysis and lattice defects detection [3].

V. 1

Description:

The first version of this software implements a command-line interface through which the user can select the input file, choose which features of the data to use and set the parameters of the algorithm. This interface also let's the user know the actual status of the program, and the time elapsed when it is finished.

An immediate problem of this version is the normalization method in the pre-processing stage of the analysis. For every data column C, associated with a selected feature, the normalization is given by Eq. 1:

$$C_{norm} = \frac{C}{C_{max}} \tag{1}$$

Where C_{norm} is the normalized data column and C_{max} is the maximum value of the original column. This is not the ideal normalization method, since the range of the new column is now $[C_{min}/C_{max}, 1]$ (where C_{min} is the minimum value of C), instead of being $[0, 1]$. This issue is going to be fixed in the next version of the software.

The elapsed time shown when the algorithm finalises takes into account the training of the SOM, the classification process of the atoms in the sample and the writing of the output file. This last step is extremely inefficient and should be optimized in future updates of the software.

Testing:

BCC Fe Bulk With a Void:

As a simple test, a 25×10^4 atoms BCC Fe bulk with a void was analyzed. The sample was created and relaxed with LAMMPS. Since the sample has no defects, the goal was to use the SOM to identify the void in the center of the simulation box. This was achieved using parameters $\sigma = 1$, $\eta = 0.5$ and $f = 1$, clustering the atoms into 3 groups and using the centro-symmetry parameter and the coordination (via Coordination Analysis with $r_c = 6$), both calculated with LAMMPS, as features.

The results are shown in Fig. 1 and Fig. 3. Although the algorithm classified the atoms into 3 groups, no atom was assigned to group 1. For the previously specified parameters, the SOM didn't produce useful

results when clustering the atoms into only 2 groups. For V.1 of the algorithm, this analysis took around $26s$ to complete. The same sample is presented in Fig. 2 and Fig. 4, with the atoms color coded by the centro-symmetry parameter, for comparison.

HEA Nano-foam Under Compression:

Fig. 5 shows the result of applying the SOM to a compressed HEA nano-foam sample, with the goal of classifying the atoms into 2 groups. The network was trained using parameters $\sigma = 1$, $\eta = 0.5$ and $f = 1$. The features used in this occasion were:

- Coordination, via Coordination Analysis with cutoff radius $r_c = 6$
- Centro-symmetry parameter
- Atomic volume, via Voronoi Analysis

On the other hand, Fig. 6 shows the same sample, with the atoms classified by structure type (FCC, HCP, BCC and Other) using the PTM algorithm. It can be observed that the yellow atoms in Fig. 5 correspond to the HCP, BCC and Other atoms in Fig. 6. While the atoms in the remaining group of the SOM classification correspond to the FCC atoms of the sample. Fig. 7 shows the atoms classified into group 1 (yellow atoms) in a slice of the sample and Fig. 8 shows all non-FCC structure types in that same slice of the sample. In this sense, the SOM is correctly identifying the atoms associated with defects and/or that belong to the surface of the sample. This analysis took $198s$ to complete.

A problem that has been noticed while performing this test is that the order in which the features are specified seems to affect the resulting classification of the atoms. In this case, the algorithm produced the results shown in Fig 5 and Fig. 7 when the features were specified in the order *coordination* $\rightarrow$ *centro-symmetry* $\rightarrow$ *atomic volume*, but couldn't get the same results when the order of the features was *coordination* $\rightarrow$ *atomic volume* $\rightarrow$ *centro-symmetry*. The origin of this issue is unknown, but this could be solved in future updates of the algorithm.

HEA Nano-foam Under Tension:

Fig. 9 shows the result of applying the SOM to a tensioned HEA nano-foam sample, with the goal of classifying the atoms into 6 groups so as to compare the results with a different unsupervised learning method, performed by N. Amigo with the K-means software he designed [4]. The SOM network was trained using parameters $\sigma = 1$, $\eta = 0.5$ and $f = 1$, and the features used were:

- Potential energy (per-atom)
- Centro-symmetry parameter (12 and 18 neighbors)
- Structure type

- Voronoi coordination number
- Radial function coordination

On the other hand, Fig. 10 shows the same sample, with the atoms classified into 6 groups by N. Amigo's K-means clustering method, which also relies on per-atom cuantities. It can be seen that the SOM didn't perform as expected in this occasion, since it couldn't identify the slip planes present in some of the sample's filaments, which were correctly identified by the K-means algorithm.

A second test produced the desired results, as shown in Fig. 11, where the slip planes identified by the K-means method are now correctly identified by the SOM. In this case, the parameters used were the same as in the previous test, but the features used were:

- Atomic volume
- Radial function coordination
- Centro-symmetry parameter (12 and 18 neighbors)

Both tests took around $260s$ to complete.

Sumary

A summary of all the tests performed for V. 1 of the software is presented in Table 1. The same parameters were used on every test:

- $f = 1$
- $\sigma = 1$
- $\eta = 0.5$

References

[1] Harris, C.R., Millman, K.J., van der Walt, S.J. et al. Array programming with NumPy. *Nature 585*, 357–362 (2020). DOI: 0.1038/s41586-020-2649-2.

[2] McKinney, W. Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference, Volume 445* (2010).

[3] Troncoso, J. F. ClasSOMfier: A neural network for cluster analysis and detection of lattice defects. *arXiv e-prints,* (2020).

[4] Amigo, N. Crystalline structure and grain boundary identification in nanocrystalline aluminum using K-means clustering. *Modelling and Simulation in Materials Science and Engineering, 28*(6), 065009, (2020).

Test	**Nº of atoms**	**Features**	**N**	**Time elapsed**	**Results**
BCC Fe with void	$\approx 25 \times 10^4$	Centro-symmetry g_r coordination	3	$26s$	Interpretable
Compressed HEA Nanofoam	$\approx 20 \times 10^5$	g_r coordination Centro-symmetry Atomic volume	2	$198s$	Interpretable
Tensioned HEA Nanofoam	$\approx 20 \times 10^5$	Potential energy Centro-symmetry (12 and 18 neighbors) Structure type Voronoi coordination g_r coordination	6	$260s$	Not interpretable
Tensioned HEA Nanofoam	$\approx 20 \times 10^5$	Atomic volume g_r coordination Centro-symmetry (12 and 18 neighbors)	6	$258s$	Interpretable

Table 1: Size of the sample, features, number of groups (**N**), performance and result of V.1 of the algorithm for every test.

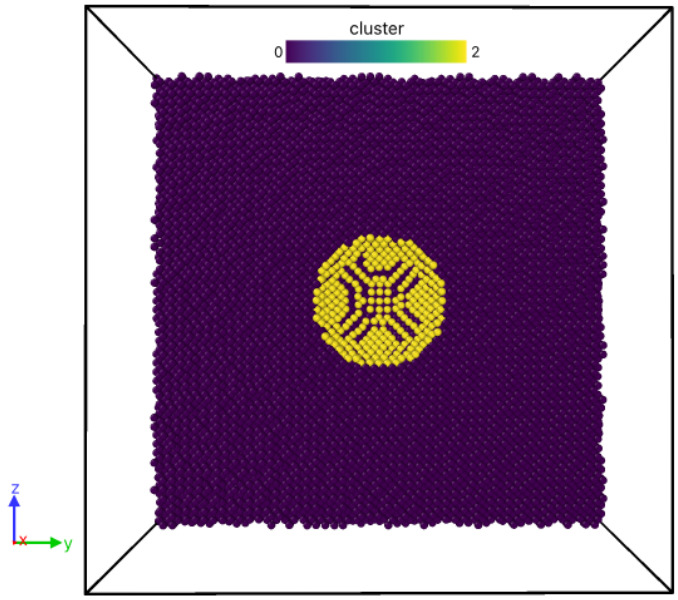

Figure 1: BCC Fe bulk with void. The atoms have been clustered into 3 groups using parameters $\sigma = 1$, $\eta = 0.5$ and $f = 1$. The centro-symmetry and coordination were used as features. A slice of the sample, with groups 0 (purple) and 2 (yellow), is shown.

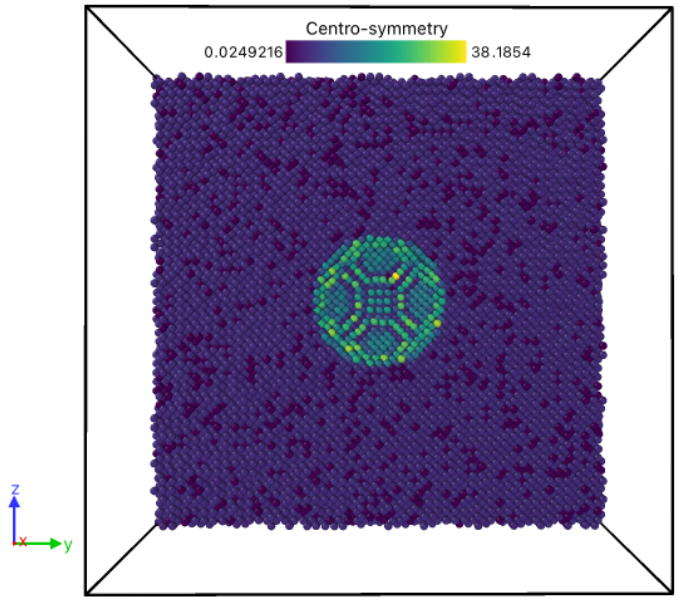

Figure 2: BCC Fe bulk with void. The atoms have been color coded using the centro-symmetry parameter (as calculated by LAMMPS). A slice of the sample is shown.

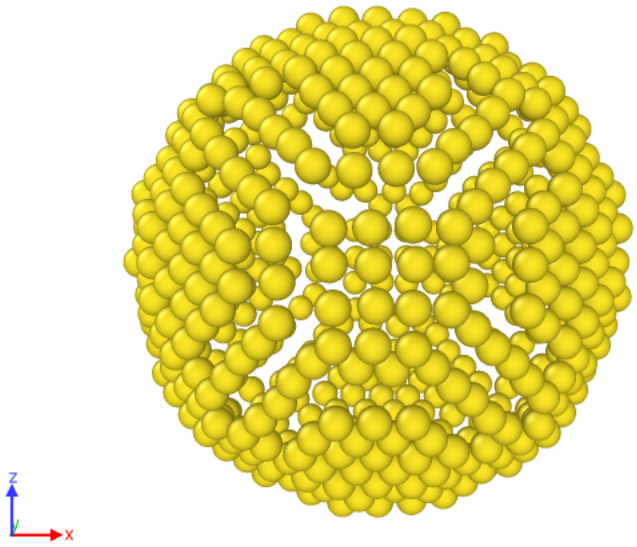

Figure 3: Void of the BCC Fe bulk sample. The atoms belonging to group 2, as classified by the SOM, are shown. The centro-symmetry and coordination were used as features.

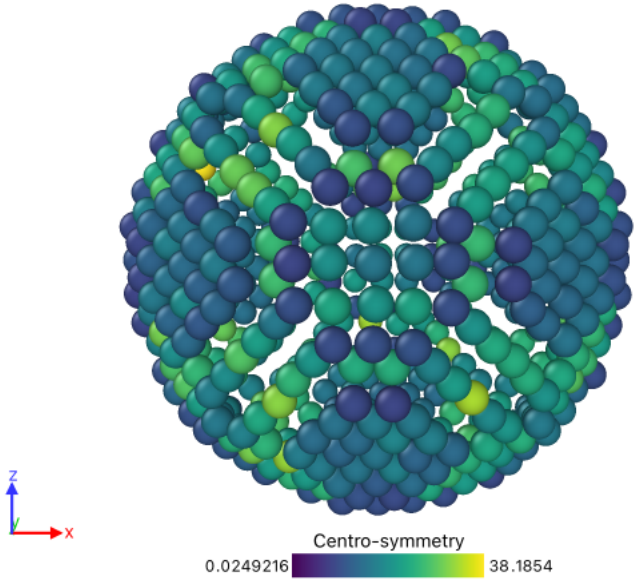

Figure 4: Void of the BCC Fe bulk sample. The atoms have been color coded using the centro-symmetry parameter (as calculated by LAMMPS).

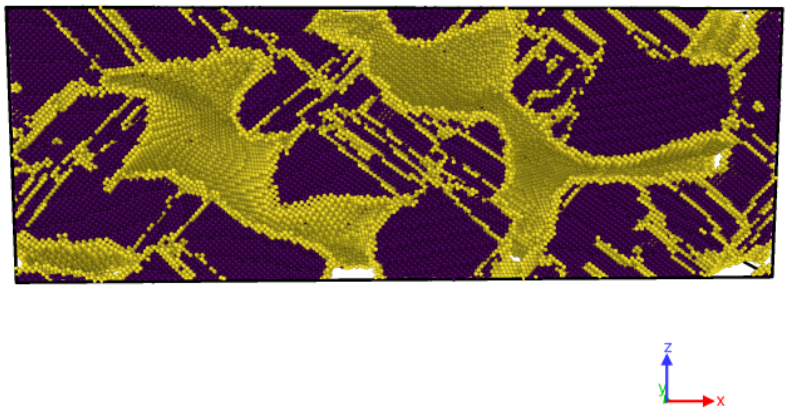

Figure 5: HEA Nano-foam under compression. The atoms have been clustered into 2 groups using parameters $\sigma = 1$, $\eta = 0.5$ and $f = 1$. The centro-symmetry, coordination (via Coordination Analysis with $r_c = 6$) and atomic volume (via Voronoi Analysis) were used as features.

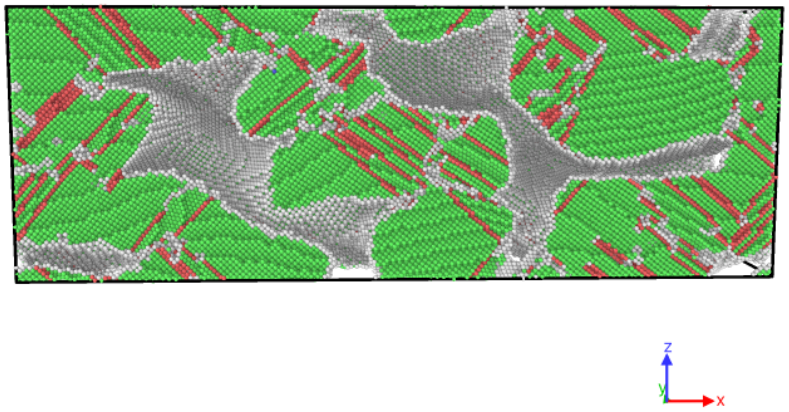

Figure 6: HEA Nano-foam under compression. The atoms have been clustered using the PTM algorithm into 4 categories: FCC (green), HCP (red), BCC (purple) and Other (white).

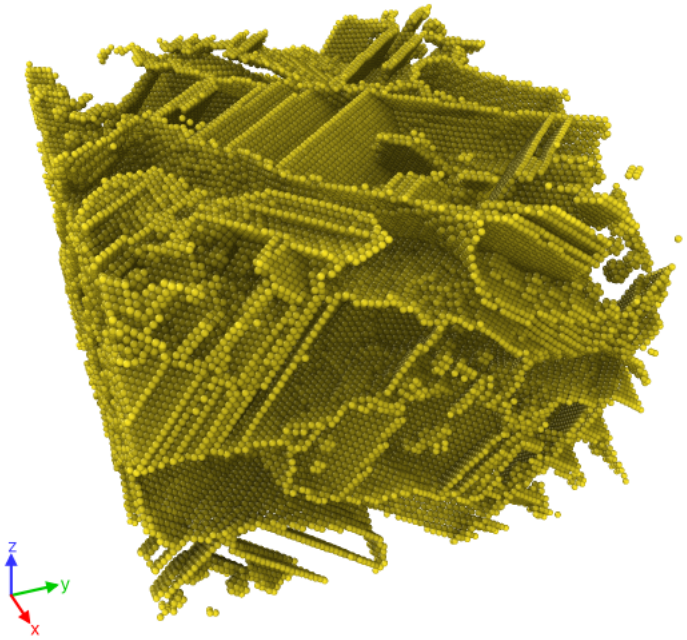

Figure 7: HEA Nano-foam under compression. Atoms classified in group 1 (yellow atoms) are shown in a slice of the sample.

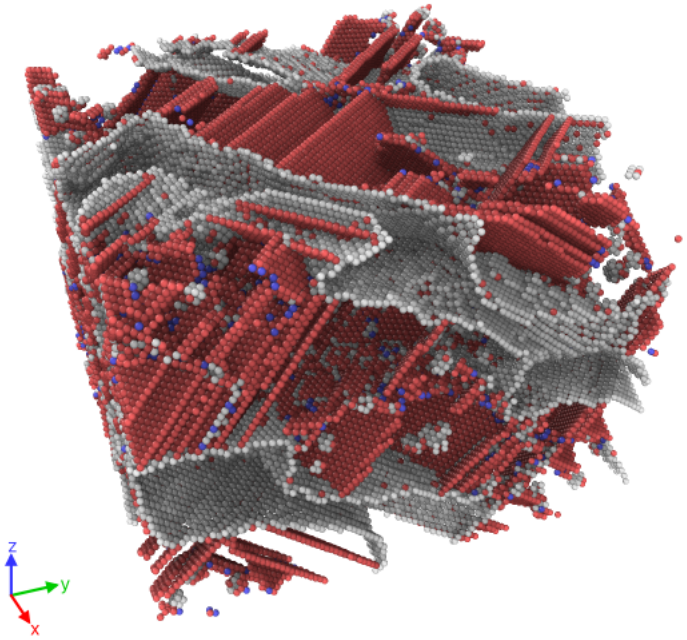

Figure 8: HEA Nano-foam under compression. Structure types HCP (red), BCC (purple) and Other (white) are shown in a slice of the sample.

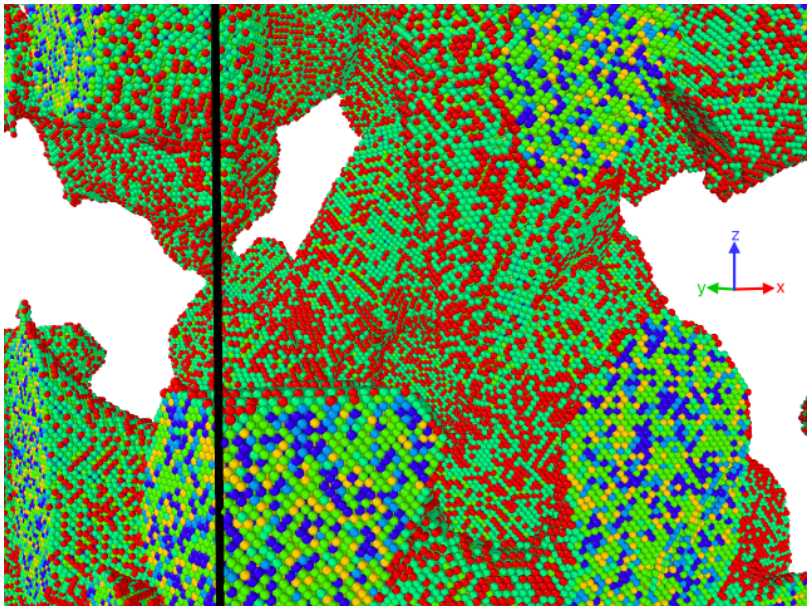

Figure 9: HEA Nano-foam under tension. The atoms have been clustered into 6 groups using parameters $\sigma = 1$, $\eta = 0.5$ and $f = 1$. The potential energy, centro-symmetry parameter (with 12 and 18 neighbors), structure type, Voronoi coordination number and radial function coordination were used as features.

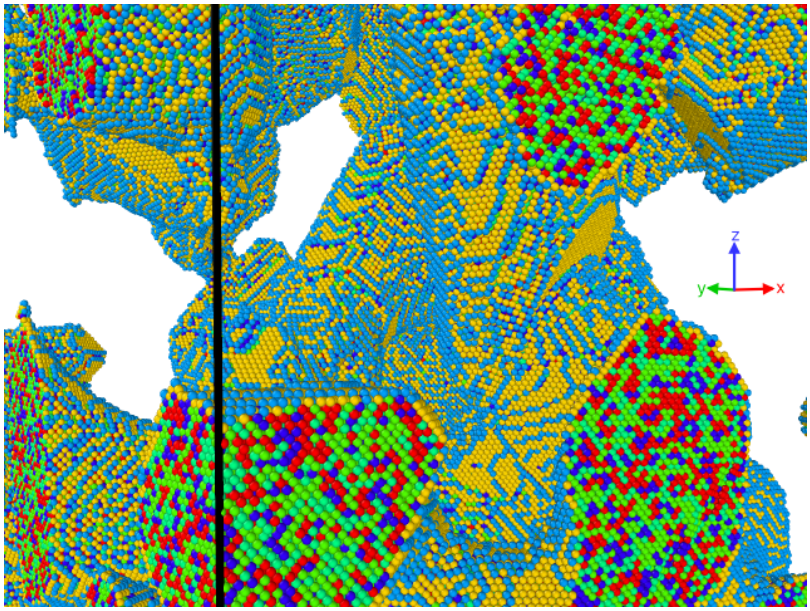

Figure 10: HEA Nano-foam under tension. The atoms have been clustered into 6 groups using N. Amigo's K-means clustering software. This analysis was performed by N. Amigo.

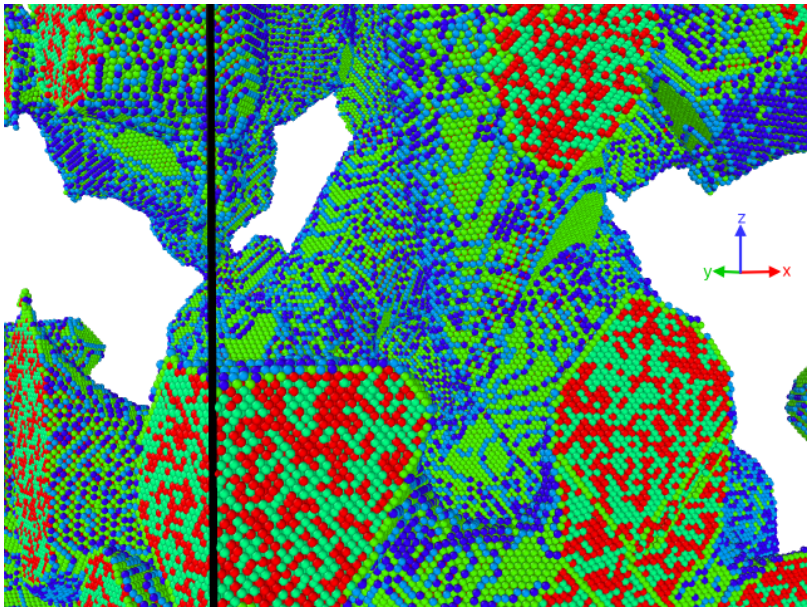

Figure 11: HEA Nano-foam under tension. The atoms have been clustered into 6 groups using parameters $\sigma = 1$, $\eta = 0.5$ and $f = 1$. The atomic volume, radial function coordination and centro-symmetry parameter (with 12 and 18 neighbors) were used as features.