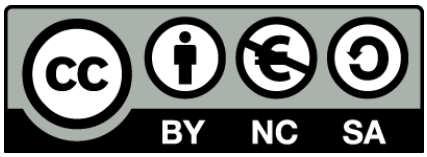
UNIVERSITAT DE BARCELONA

Advanced computational tools for EELS data reduction and clustering, quantitative analysis and 3D reconstructions

Javier Blanco Portals



PhD Thesis

Advanced computational tools for EELS data reduction and clustering, quantitative analysis and 3D reconstructions.

Javier Blanco Portals

Prof. Dr. Francesca Peiró Martínez

Dr. Sònia Estradé Albiol

UNIVERSITAT DE BARCELONA

Advanced computational tools for EELS data reduction and clustering, quantitative analysis and 3D reconstructions.

Memoria presentada para optar al título de doctor por la Universidad de Barcelona.

Programa de doctorado en Nanociencia

Autor: Javier Blanco Portals

Directoras: Prof. Dr. Francesca Peiró Martínez

Dr. Sònia Estradé Albiol

Tutora: Prof. Dr. Francesca Peiró Martínez

UNIVERSITAT DE BARCELONA

Acknowledgements

No one will ever be more surprised than me to see this doctoral thesis come to an end. But here I am, at last, writing the final words of this long manuscript, the result of uncountable hours of work and headaches. It has certainly been a long journey, peppered with many good times and memories, but also with hardship and tough moments. However, standing here, at the end of it, I can say that it certainly was a path worth walking. It has made me grow appreciative of the values of hard working, perseverance and, above all, the importance of being surrounded by awesome people. Precisely this following lines are an attempt to honour all those people who shared this path with me. Without them, this wouldn't have been possible.

First of all, I would like to thank my bosses. Paqui, Sònia, you gave me the opportunity to be a part of this incredible group of people, the LENS. To be honest, I am still amazed by it. Somehow, you were able to see past the image of a guy wearing homemade sleeveless T-shirts, carrying a skateboard everywhere and rocking a hairstyle with bangs. In all seriousness now, I am truly thankful for this opportunity, for the trust you deposited in me throughout these years, and the constant encouragement you gave me in this final race towards the finish line.

To the true LENS, the old guard. Lluís Y., Josep, Alberto, Gemma, Pau and my teacher of arcane TEM knowledge Lluís L. You paved the way, and you were there to get me back in my feet whenever I failed or was lost (more often than I would like to acknowledge). More importantly, you welcomed me into this very special group of people, and for that I will always be thankful.

To MY true LENS, my fellow minions. To Sergi, thank you for encouraging me to join the group and for being an awesome host in 'Maguncia'. To Dani, thank

you for being that guy with an overwhelming positive vibe, for being part of my little verbal and nerf wars, and for boosting my morale with flattering comments about my work. To Cata, from all the people in our group, you are possibly the one to whom I owe more gratitude. You assisted me countless times in all sorts of administrative odysseys and you helped me solve so many problems along the way that it's impossible to list them all here. Also, thank you for organizing and throwing the best and geekiest parties that one can imagine. I am afraid that you set the bar too high for the next generation of LENS minions.

To the rest of the happy bunch from the 'Despatx patera', those who weren't lucky enough to be part of the LENS. Juanlu, Oriol, Julià, Giovanni, Sergio, Sergi, Rafa and Adrià, thank you for enduring the random and wild explosions of activity, and the litany of curse words I would pronounce whenever I succeeded or failed at any given task. Thank you all for being part of this story.

Thanks to Fèlix, Joan, and the rest of the staff members in the TEM-MAT unit at the CCiTUB in Barcelona, for helping me during the long hours I spent there 'testing the resilience' of several instruments and for sharing some good laughs from time to time.

To the people from the 'Departamento de Ciencia de los Materiales e Ingeniería Metalúrgica y Química Inorgánica' at the University of Cádiz, Ana Hungría, 'Japo', Ramón, Susana and everyone else, thank you for making me feel at home down there. Special thanks to Jose Calvino, for taking me in as part of the group during my visits. And very special thanks to Miguel, who shared with me an uncountable number of hours acquiring images and data inside the TITAN room, and who helped me to navigate through the countless alleys in the 'Mercadona' of samples that I often brought with me. And of course, to Esther and the little Emma and Julia, thank you for letting me borrow so much time from Miguel.

And speaking about uncountable hours in the darkness of a TEM room, very special thanks to Emerson Coy. Thank you for being an incredible host and

assisting me during my visits to Poland. You also taught me that 12 straight hours of work followed by 3 litres of beer can be made into a sustainable lifestyle, as long as it is restricted to a few days a year. Thanks also to the rest of the people from the NanoBioMedical Centre at the Adam Mickiewicz University in Poznan for welcoming me into your laboratories.

There is also a very special place here for my friends Pablo and Yass. This PhD is only the last step of a ladder, that began with us enduring all those long nights of study and the mornings of dragging ourselves into the library again. You were my wingmen in our way towards the physics degree, and you still are. Thank you for being there for me during some of the harder and darker parts of those times, I would not have made it without you.

To the rest of friends and roommates who shared their lives with me during all these years, to the 'Bogavantes', to Tato, Oleiros and Pombo, and to everyone else, thanks for so many good times and adventures.

To my friends and relatives from Muros and the Ancoradoiro. You were my escape route, my little window to a different lifestyle, to a calmer place. Very special thanks to Marcos, for bringing me to so many surfing adventures and for giving me some of the wiser but also crazier pieces of advice ever. 'Hakuna Matata, primo'.

To the little 'tortillo', who kept me sane during the second confinement in these strange times.

To my grandparents and to the myriad of uncles, aunts and cousins that I've been lucky enough to have as a family, thank you. Also, thanks to my second family in Euskadi. You've made me love this place.

Very special thanks to my little brother and sister, Sebastián and Carmen. Growing up together you always thought I would end up been a mad scientist. Well, I can say now with confidence that you nailed it. Thank you for being the best siblings I could ever wish.

To my parents. Anything good about me is nothing but a reflection of what I've learned from you. You let me walk my way, you guided and protected me, and you were always there to put me back on track whenever I got lost. My only hope is that one day I may be half as good parent as you are.

To my partner in life, Maritxu, because you endured it all. The long hours of work, the times apart, several moves between cities and some of my occasional 'brilliant' ideas. You gave light to my darker days. For your patience and love, thank you.

Finally, to my wonderful daughter, Maya, who I hope one day will understand all the time I poured into this project, and who I hope that would be proud of what I did. You are the best of me.

And to anyone and everyone else who has accompanied me on this journey, thanks.

It's been a hell of ride.

Table of contents

Chapter 1 Introduction 1

1.1. Electron energy loss spectroscopy.....1

1.2. Electron tomography. 5

1.3. References. 10

Chapter 2 Objectives. 27

Chapter 3 Total variation minimization. Optimizing morphological and analytical electron tomography reconstructions. 33

3.1. Introduction. 33

3.2. Mathematical principles and reconstruction methods. 36

3.2.1. Classic approaches and iterative algebraic algorithms. 36

3.2.2. Compressed sensing (CS) and total variation minimization (TV_{min}). 45

3.3. A qualitative comparison of SIRT and TVM. 51

3.3.1. SIRT vs TVM. Undersampling the angular range. 54

3.3.2. Noise resilience of the TVM method. 62

3.3.3. Resilience of the TVM method to image erosion. Severe undersampling conditions in X-EDS analytical electron tomography. 68

3.4. A case study. 73

3.4.1. Morphological electron tomography. Thick-Thin shell Fe_2O_3 NP... 73

3.4.2. Analytical electron tomography. Iron - Nickel oxide NPs. 80

3.5. Conclusions 88

3.6. References 90

Chapter 4 Machine learning for EELS data analysis. Clustering and dimensionality reduction methods. 97

4.1. Introduction. 97

4.2. Reference sample for the analysis. 100

4.3. Dimensionality reduction methods (DRM). 102

4.3.1. Matrix factorization. 103

4.3.2. Non-linear neighbour-graph approaches 120

4.4. Clustering analysis. 131

4.4.1. General formulation. 131

4.4.2. Distance measurements and metrics. 132

4.4.3. Taxonomical classification of clustering algorithms. 138

4.5. Combining DRM and clustering analysis. 164

4.6. Performance analysis. 173

4.6.1. Clustering and DRM combinations. Quantitative analysis 173

4.6.2. UMAP-HDBSCAN combination. Noise resilience and outlier detection limits. 186

4.7. Conclusions. 197

4.8. References 200

Chapter 5 ELNES analysis combining clustering and NLLS. The WhatEELS software solution. 213

5.1. Introduction. 213

5.2. Elemental quantification in WhatEELS. 217

5.2.1. Inelastic scattering of atomic inner-shell electrons and Bethe's theory. 218

5.2.2. Elemental quantification and the effects of a finite convergence angle. 222

5.3. WhatEELS. Software architecture. 227

5.3.1. Overview. 227

5.3.2. Peripheral modules -1- 230

5.3.3. Core fitting components -2- 240

5.3.4. Others 269

5.4. Iron-Manganese oxide nanoparticles. 271

5.4.1. Clustering Analysis. 271

5.4.2. NLLS model. 273

5.4.3. Reduced chi squared. 275

5.4.4. Centres analysis. 277

5.4.5. White lines analysis. 279

5.4.6. Quantification. 281

5.5. The panel structure. 282

5.6. Conclusions. 295

5.7. References 298

Chapter 6 Analysis of dopant segregation and Ce oxidation state in doped cerium oxide mesoporous materials. 307

6.1. Introduction. 307

6.2. CGO, CPO and CGPO. Qualitative analysis. 312

6.2.1. Pr-doped cerium oxide mesoporous material (CPO). 313

6.2.2. Gd-doped cerium oxide mesoporous material (CGO). 321

6.2.3. Gd-Pr-doped cerium oxide mesoporous material (CGPO). 325

6.3. CGPO. Quantitative analysis and electron tomography. 331

6.3.1. EELS-SI quantitative analysis for the CGPO. 332

6.3.2. ADF and X-EDS electron tomography. 337

6.4. CGO and CGO(HF). Quantitative analysis. 341

6.4.1. Quantitative analysis of the CGO sample. 341

6.4.2. Quantitative analysis of the CGO(HF) sample. 347

6.5. Conclusions. 351

6.6. References. 354

Chapter 7 General conclusions. 359

Appendix A Instrumentation list. 371

Appendix B Scientific curriculum. 375

Appendix C Resumen en castellano 387

Chapter 1

Introduction

This thesis has been primarily dedicated to the exploration and implementation of new computational analysis tools and techniques for the characterisation of nanomaterials and devices via transmission electron microscopy (TEM). In particular, the focus is set on the fields of electron energy loss spectroscopy (EELS) and electron tomography (ET). Despite their long history, both ET and EELS still lie at the cutting edge of the characterization techniques in the field of materials science.

1.1. Electron energy loss spectroscopy.

EELS provides access to the study of a wide range of different properties and physical phenomena of materials at the nanoscale, standing out among the rest of techniques due to its high spatial resolution capabilities[1–3].

This range of physical phenomena at the nanoscale that can be investigated through EEL spectroscopy include such diverse fields of analysis as the measurements of the band-gap energy for a variety of different materials [4–8], or the resolution of the surface plasmonic response and resonant modes of nanoscopic devices for optical applications[9–11]. Furthermore, the recent advances in the instrumentation regarding the brightness of the electron sources[12] and the newly developed monochromators[13] have finally enabled the study of the phonon vibrational modes at the nanoscale through the spectroscopic measurements in the regions nearby to the zero-loss peak[14–17]. These are examples that utilize the so-called low-loss region of the spectra (i.e., the spectral region accounting for the elastic scattering events and the plasmonic response of the materials)[18].

Most commonly, and especially in the context of this PhD, EELS is used for the quantitative and qualitative analysis of elemental distributions at the nanoscale[18,19]. Here, EELS stands out by its sensitivity to the low atomic number elements (e.g., oxygen, nitrogen and carbon) and its capability to retrieve reliable quantification measurements, using the scattering cross-sections extracted from first-principles computations[18,20,21]. It also allows the investigation of the elemental oxidation states and atomic coordination number in a material down to the atomic resolution. This is done through the analysis of the so-called energy-loss near-edge structures (ELNES) of core-loss edges, observed in the core-loss region of the spectra (i.e., the energy loss region of the spectra produced by the excitation of inner-shell electrons of the illuminated material through its interaction with the TEM electron beam) [18,22–25].

The rapid advances in the TEM instrumentation[26–28], including the development of better sensing capabilities for the EELS spectrometers (e.g., direct detection cameras) and better software solutions for faster acquisition modes[29,30], are the driving force behind the ever-growing size and complexity of the acquired spectral datasets. Alongside these newer capabilities influenced by the instrumental advances, the field of ELNES analysis has also experienced a paradigmatic change regarding the acquisition requirements for some specific applications. In a nutshell, the classic demand of high-quality spectra with a large energy resolution and a good signal-to-noise ratio (SNR) has been substituted by the acquisition of larger but noisier spectral datasets (i.e., larger EELS spectrum images, SI). The foundations for this new paradigm are based on the hypothesis that a large statistical sample, even if limited in its quality, provides access to the same level of characterization detail (if not better) than a smaller but crispier spectral sample, with the added advantage of an often-shorter acquisition time. As such, this has become a common practice in the characterization of sensitive materials (i.e., prone to be damaged during the spectral acquisition) or in

experiments focused on the investigation of the oxidation states for samples that are easily reduced by interaction with the electron beam[31–34].

Driven by this growth in the dataset sizes and the challenges of dealing with the noisier datasets acquired with a limited SNR, several analysis techniques from the field of machine learning have been steadily brought into the standard EELS characterization pipelines, a trend that is expected to continue in the future[35].

The algorithms of principal component analysis (PCA)[36–40] and non-negative matrix factorization (NMF)[41–43], which are the most relevant representatives of the family of linear dimensionality reduction methods (DRM) based on the resolution of a matrix factorization minimization problem, are now common procedures for EELS data analysis. These algorithms are often used for the obtention of a qualitative compositional mapping from the EELS SI, or as a fast and simple denoising pre-processing step[44–46].

However, in recent times the focus has been shifted towards the inclusion of newer machine learning methodologies. An example would be the introduction of clustering analysis to resolve problems of EELS SI segmentation in an unsupervised manner[47]. One of the main advantages of these clustering methods is that the spectral representatives for each cluster (i.e., the centroids) are always physically meaningful, as they are computed averaging the spectral signal per cluster (i.e., the non-physical features often observed for the components resolved by PCA or NMF are no longer a standing issue[48]).

Several new methodologies have been recently reported in the literature, promoting the inclusion of clustering as a way to improve the analysis of larger EELS SI. For example, it has been shown that K-means clustering can be employed to extract the spectral endmembers of a dataset (i.e., spectral representatives), which are later used by a multiple-linear least-squares (MLLS) fitting process and improving the accuracy of the obtained results[49]. Also, new approaches combining advanced non-linear dimensionality

reduction algorithms (such as t-SNE[50,51], from the family of graph-based methodologies) and density-based clustering algorithms (OPTICS[52,53]) have made their way into the EELS analysis arena, promoting a methodology with fully data-driven philosophy for the segmentation process[54].

A large portion of the work done during this PhD thesis involves a systematic revision of these new methods (clustering), the already stablished ones (PCA and NMF), and some of the possible combinations of the DRM and clustering techniques. A qualitative and quantitative analysis of their performance, their advantages and their shortcomings, is conducted. Also, a new methodology is introduced, involving the state-of-the-art algorithms for graph-based dimensionality reduction (UMAP[55]) and for the density-based hierarchical clustering analysis (HDBSCAN[56]). This combination was demonstrated to outperform any other possible segmentation technique, providing at the same time a new route towards a fully data-driven methodology for ELNES analysis[57].

Furthermore, the combination of clustering and non-linear least-squares (NLLS) fitting has also been proven as a promising solution to improve the stability of the latter[58]. Thus, it represents a major asset for an in-depth analysis of the ELNES results in experiments involving changes in the oxidation state of the materials, that are often resolved by a NLLS fitting (e.g., localized cation reduction in transition metal oxides and rare earth oxides).

Another focal point for the work done during this PhD is to provide a ready-to-go software solution for all these methodologies. This way, any future user may be able to reproduce and apply these methods to new experiments, without requiring a special knowledge of any programming language. This led to the development of a complete and independent software solution called **WhatEELS**[58]. It was initially devised as an alternative to the already available HyperSpy[59] analysis library (free-access and Python-based) and the proprietary Digital Micrograph software solution from Gatan, as they often failed to provide the required plasticity and computational capabilities to tackle some of the complex problems faced during the development of the

thesis. For example, these available solutions often came short for the quantitative analysis of spectra containing consecutive elemental edges within an extremely small range of energies (CeM_4 and PrM_5) or with directly overlapping features (CeM_{32} and GdM_{54})[60]. It was also done with a future proofing philosophy in mind, coded as a modular platform capable of assimilating new machine learning capabilities in the future.

1.2. Electron tomography.

The field of ET provides the materials scientist with one of the most versatile toolsets for the characterization of materials at the nanoscale, as it allows the reconstruction of 3D volumes from a limited set of 2D projections acquired. ET is fundamentally different from the rest of TEM-related techniques, as they all are confined to a 2D sensing space in which the nature of the acquisition process may sometimes lead to artefactual results, by overlapping features in the projected datasets. The focus on these reconstructions can be directed towards the retrieval of the volumetric morphological structure (morphological ET or, simply, ET) or towards the obtention of the 3D chemical composition (analytical-ET) of the materials under analysis.

Morphological ET uses for the reconstruction process the data acquired by some of the imaging modes in a TEM (i.e., high-angle annular dark field imaging, HAADF, angular dark field imaging, ADF, or bright field imaging, BF). Usually, ET is focused on the reconstruction of volumes at a 'mesoscale'[61] in which one might quantitatively investigate the size distribution and dispersion of agglomerates of NPs[62], resolve the specific facets exposed in the surface of a NP[63] or obtain the structure of potential cavities inside nanostructured materials[64]. Nevertheless, the feasibility of reconstructing the 3D atomic crystalline structure of nanomaterials in an ET experiment has also been proven in the literature, whether by using as set of oriented atomic resolution ADF (or HAADF) projection images (i.e., in a zone axes of the nanomaterial)[65–67], or by a mixed combination of high resolution HAADF image acquisition and 'atom counting' (i.e., atomic modelling from the HAADF

images, using the Z^2 relation of the contrast with the atomic column depth)[68–70].

Regarding the analytical-ET, the reconstructions are computed from the chemical mapping projections extracted from spectral datasets that are acquired by any of the analytical techniques in a TEM (i.e., X-EDS elemental mappings or EELS SI)[71]. In some cases, both signals (X-EDS and EELS) can be co-acquired and reconstructed at the same time[72]. In general, these reconstructions are focused on the obtention of the 3D distribution of elements[73–76]. However, the analytical-ET has also been demonstrated capable of tackling problems with a higher degree of complexity, such as the 3D reconstruction of the distribution of the oxidation states in a material. These experiments rely mostly on the ELNES analysis of the series of experimentally acquired EELS-SI, as it provides an easy access to the identification of cationic oxidation states in several different types of materials (e.g., transition metal and rare earth oxides). The extraction of the projection mappings for the distribution of oxidation states used in the ET reconstruction may be approached in a variety of different ways. One of them would be the computation of the independent component analysis (ICA) decomposition results for the EELS SI, and the subsequent selection of the score mappings corresponding to the signals with the ELNES for the oxidations states of interest, using them as the projections in the ET reconstruction [77]. Another option would be the computation of the MLLS score mappings for the ELNES reference signals of the different oxidation states in the material under analysis (extracted from the SI itself or from theoretical approximations), using them later as the projections in the reconstruction[78,79].

In general, the advances in ET (also applicable to analytical-ET) can be divided into two different subcategories. The first one would be related to the acquisition process itself. This would include both the implementation of new acquisition software solutions and modes, and the advances in the specific TEM instrumentation used for the experimental acquisition of the tomographic series. For example, a large part of the research the field of ET

has been devoted in recent times to the implementation of faster and more efficient acquisition routines that avoid an overexposure of the samples to the electron beam, effectively reducing the induced beam damage without compromising the level of detail recovered (fast-ET)[80,81]. Regarding the pure instrumental advances, the development of the cryo-TEM appears to be in the forefront of the ET at the moment, as it allows the implementation of this characterization technique in samples otherwise too sensitive to the electron beam[82–84]. The maturity reached by this technique and its current relevance within the scientific community is perfectly exemplified by the awarding of 2017 Nobel Prize in Chemistry to Jacques Dubochet, Joachim Frank, and Richard Henderson for their work in the development of the cryo-TEM[85].

Another relevant instrumental advance for the field of ET (both morphological and analytical ET) would be the development of the direct detection cameras and sensors[27,28]. Given their improved detection sensitivity and faster response, these instruments have facilitated the obtention of a larger number of projections in an ET experiment without the previously associated cost increments for the total acquisition time. At the same time, they also allowed a drastic reduction of the electron beam accelerating voltages (down to 60 kV) and currents while still maintaining acceptable levels of SNR.

The combination of advances in both the pure instrumental part (direct detection cameras) and the acquisition modes (continuous image registration during the sample rotation) has been demonstrated to provide an edge in the obtention of high-quality results for some problems where the extreme sensitivity of the sample to the electron beam posed an otherwise insurmountable degree of experimental complexity[86].

Finally, regarding the specific field of X-EDS analytical-ET, the development of windowless detectors[87] and the increment in their surface area sizes, as well as the emergence of the multiple detector configurations in the TEM column (such as the 4-section Super-X detector) coupled with the implementation of software corrections for the inevitable shadowing effects of the sample

holders during the acquisition process[73,88], have also played a fundamental role in the widespread growing usage of the analytical capabilities of the X-EDS ET[71].

Moving on now, the second subcategory of advances in the field of ET would be focused on providing better algorithms for the actual reconstruction process[89–92], while still refining the already available ones[93]. It is not rare to encounter in a variety of applications that some of the older reconstruction algorithms are still used for the experimental analysis. An example would be the recent reconstruction of the 3D structure of the molecular machinery for the SARS-CoV-2 virus from a series of cryo-TEM images[94] using the simultaneous iterative reconstruction technique algorithm (SIRT)[95](i.e., one of the oldest algorithms from the family of algebraic methods for the ET reconstruction). Although potentially successful, these older algorithms require a large number of projections to achieve an accurate reconstruction, and even then the results may still show some of the most persistent reconstruction artefacts (e.g., streaking artefacts[96]). The inclusion of a priori knowledge about the samples in the reconstruction process, as in the case of using the discrete algebraic reconstruction technique (DART)[89], or by using algorithms formulated from the basis of the compressed sensing mathematical theory[91,97,98], have been proven time and again as the superior choice to achieve a high level of quality in the reconstructions from a limited set of projections.

In the context of this PhD thesis, the work is mainly focused on the implementation of advanced algorithms for the ET reconstruction of nanomaterials in Python programming language. The attention is centred on the TVAL3 algorithm[99], a solver for the total variation minimization (TVM) problem with its theoretical foundations in the mathematical field of compressed sensing[100]. The use of this specific algorithm in the field of ET had already been proven advantageous for the reconstruction process many times in the literature[91,92,101,102], as it requires far fewer projections for an accurate reconstruction and shows a higher tolerance to noisy conditions

than the classic approaches (e.g., SIRT). However, although cited in several publications, an actual free access implementation was not available. Therefore, it was decided to carry out a translation of the original scripts (in MATLAB) and its implementation within the Python version of the ASTRA toolbox (a software library used as a control hub for several ET reconstruction algorithms)[103,104]. As a side effect, this work is setting up a solid base for the implementation in the future of an independent software solution dedicated to ET, similarly to the one provided for the ELNES analysis through **WhatEELS**.

Finally, this methodology based on the TVAL3 algorithm is used for the experimental characterisation of the 3D morphology and chemical composition of a wide variety of different nanomaterials[60,105].

1.3. References.

[1] N.D. Browning, D.J. Wallis, P.D. Nellist, S.J. Pennycook, EELS in the STEM: Determination of materials properties on the atomic scale, Micron. 28 (1997) 333–348. https://doi.org/10.1016/S0968-4328(97)00033-4.

[2] C. Colliex, D. Ugarte, Z.L. Wang, M. Gasgnier, V. Paul-Boncour, High spatial resolution analytical electron microscopy studies on the Co/CeO2 system, Surf. Interface Anal. 12 (1988) 3–10. https://doi.org/10.1002/sia.740120104.

[3] P. Cueva, R. Hovden, J.A. Mundy, H.L. Xin, D.A. Muller, Data processing for atomic resolution electron energy loss spectroscopy, Microsc. Microanal. 18 (2012) 667–675. https://doi.org/10.1017/S1431927612000244.

[4] M.C. Cheynet, S. Pokrant, F.D. Tichelaar, J.L. Rouvìre, Crystal structure and band gap determination of Hf O2 thin films, J. Appl. Phys. 101 (2007). https://doi.org/10.1063/1.2697551.

[5] C.S. Granerød, W. Zhan, Ø. Prytz, Automated approaches for band gap mapping in STEM-EELS, Ultramicroscopy. 184 (2018) 39–45. https://doi.org/10.1016/j.ultramic.2017.08.006.

[6] D. Keller, S. Buecheler, P. Reinhard, F. Pianezzi, D. Pohl, A. Surrey, B. Rellinghaus, R. Erni, A.N. Tiwari, Local band gap measurements by VEELS of thin film solar cells, Microsc. Microanal. 20 (2014) 1246–1253. https://doi.org/10.1017/S1431927614000543.

[7] A. Eljarrat, S. Estradé, Z. Gacevic, S. Fernandez-Garrido, E. Calleja, C. Magén, F. Peiró, Optoelectronic properties of InAlN/GaN distributed Bragg reflector heterostructure examined by valence electron energy loss spectroscopy., Microsc. Microanal. 18 (2012) 1143–1154. https://doi.org/10.1017/S1431927612001328.

[8] J. Arbiol, S. Estradé, J.D. Prades, A. Cirera, F. Furtmayr, C. Stark, A. Laufer, M. Stutzmann, M. Eickhoff, M.H. Gass, A.L. Bleloch, F. Peiró, J.R. Morante, Triple-twin domains in Mg doped GaN wurtzite nanowires: Structural and electronic properties of this zinc-blende-like stacking, Nanotechnology. 20 (2009). https://doi.org/10.1088/0957-4484/20/14/145704.

[9] B.S. Archanjo, T.L. Vasconcelos, B.S. Oliveira, C. Song, F.I. Allen, C.A. Achete, P. Ercius, Plasmon 3D Electron Tomography and Local Electric-Field Enhancement of Engineered Plasmonic Nanoantennas, ACS Photonics. 5 (2018) 2834–2842. https://doi.org/10.1021/acsphotonics.8b00125.

[10] G.T. Forcherio, M. Benamara, D.K. Roper, Electron Energy Loss Spectroscopy of Hot Electron Transport between Gold Nanoantennas and Molybdenum Disulfide by Plasmon Excitation, Adv. Opt. Mater. 5 (2017). https://doi.org/10.1002/adom.201600572.

[11] A. Konečná, T. Neuman, J. Aizpurua, R. Hillenbrand, Surface-Enhanced Molecular Electron Energy Loss Spectroscopy, ACS Nano. 12 (2018) 4775–4786. https://doi.org/10.1021/acsnano.8b01481.

[12] N. Brodusch, H. Demers, R. Gauvin, Developments in Field Emission Gun Technologies and Advanced Detection Systems, in: F. Emiss. Scanning Electron Microsc., Springer, 2018: pp. 5–12. https://doi.org/10.1007/978-981-10-4433-5_2.

[13] O.L. Krivanek, T.C. Lovejoy, N. Dellby, R.W. Carpenter, Monochromated STEM with a 30 meV-wide, atom-sized electron probe, J. Electron Microsc. (Tokyo). 62 (2013) 3–21. https://doi.org/10.1093/jmicro/dfs089.

[14] O.L. Krivanek, T.C. Lovejoy, N. Dellby, T. Aoki, R.W. Carpenter, P. Rez, E. Soignard, J. Zhu, P.E. Batson, M.J. Lagos, R.F. Egerton, P.A. Crozier, Vibrational spectroscopy in the electron microscope, Nature. 514

(2014) 209–212. https://doi.org/10.1038/nature13870.

[15] F.S. Hage, G. Radtke, D.M. Kepaptsoglou, M. Lazzeri, Q.M. Ramasse, Single-atom vibrational spectroscopy in the scanning transmission electron microscope, Science (80-.). 367 (2020) 1124–1127. https://doi.org/10.1126/science.aaz5312.

[16] F.S. Hage, D.M. Kepaptsoglou, Q.M. Ramasse, L.J. Allen, Phonon Spectroscopy at Atomic Resolution, Phys. Rev. Lett. 122 (2019) 16103. https://doi.org/10.1103/PhysRevLett.122.016103.

[17] M.J. Lagos, A. Trügler, U. Hohenester, P.E. Batson, Mapping vibrational surface and bulk modes in a single nanocube, Nature. 543 (2017) 529–532. https://doi.org/10.1038/nature21699.

[18] R.F.F. Egerton, Electron Energy-Loss Spectroscopy in the Electron Microscope, Springer Science, Boston, MA, 2011. https://doi.org/10.1007/978-1-4419-9583-4.

[19] D.B. Williams, C.B. Carter, The Transmission Electron Microscope, in: Transm. Electron Microsc., Springer US, Boston, MA, 1996: pp. 3–17. https://doi.org/10.1007/978-1-4757-2519-3_1.

[20] F. Hofer, B. Luo, Towards a practical method for EELS quantification, Ultramicroscopy. 38 (1991) 159–167. https://doi.org/https://doi.org/10.1016/0304-3991(91)90117-O.

[21] F. Hofer, Determination of inner-shell cross-sections for EELS-quantification, Microsc. Microanal. Microstruct. 2 (1991) 215–230. https://doi.org/10.1051/mmm:0199100202-3021500.

[22] M. Haruta, H. Higuchi, T. Nemoto, H. Kurata, Local quantification of coordination number for perovskite-related oxides using atomic resolution EELS maps, Appl. Phys. Lett. 113 (2018). https://doi.org/10.1063/1.5041063.

[23] L.A.J. Garvie, P.R. Buseck, Determination of Ce4+/Ce3+ in electron-

beam-damaged CeO2 by electron energy-loss spectroscopy, J. Phys. Chem. Solids. 60 (1999) 1943–1947. https://doi.org/10.1016/S0022-3697(99)00218-8.

[24] H. Tan, J. Verbeeck, A. Abakumov, G. Van Tendeloo, Oxidation state and chemical shift investigation in transition metal oxides by EELS, Ultramicroscopy. 116 (2012) 24–33. https://doi.org/10.1016/j.ultramic.2012.03.002.

[25] P. Torruella, A. Ruiz-Caridad, M. Walls, A.G. Roca, A. López-Ortega, J. Blanco-Portals, L. López-Conesa, J. Nogués, F. Peiró, S. Estradé, Atomic-Scale Determination of Cation Inversion in Spinel-Based Oxide Nanoparticles, Nano Lett. 18 (2018) 5854–5861. https://doi.org/10.1021/acs.nanolett.8b02524.

[26] M.W. Tate, P. Purohit, D. Chamberlain, K.X. Nguyen, R. Hovden, C.S. Chang, P. Deb, E. Turgut, J.T. Heron, D.G. Schlom, D.C. Ralph, G.D. Fuchs, K.S. Shanks, H.T. Philipp, D.A. Muller, S.M. Gruner, High Dynamic Range Pixel Array Detector for Scanning Transmission Electron Microscopy, Microsc. Microanal. 22 (2016) 237–249. https://doi.org/10.1017/S1431927615015664.

[27] C.R. Booth, P.E. Mooney, B.C. Lee, M. Lent, A.J. Gubbens, K2: A Super-Resolution Electron Counting Direct Detection Camera for Cryo-EM, Microsc. Microanal. 18 (2012) 78–79. https://doi.org/10.1017/S1431927612002243.

[28] M. Sun, C.M. Azumaya, E. Tse, D.P. Bulkley, M.B. Harrington, G. Gilbert, A. Frost, D. Southworth, K.A. Verba, Y. Cheng, D.A. Agard, Practical considerations for using K3 cameras in CDS mode for high-resolution and high-throughput single particle cryo-EM, J. Struct. Biol. 213 (2021) 107745. https://doi.org/10.1016/j.jsb.2021.107745.

[29] J.L. Hart, A.C. Lang, A.C. Leff, P. Longo, C. Trevor, R.D. Twesten, M.L. Taheri, Direct Detection Electron Energy-Loss Spectroscopy: A Method

to Push the Limits of Resolution and Sensitivity, Sci. Rep. 7 (2017) 1–14. https://doi.org/10.1038/s41598-017-07709-4.

[30] P. Longo, J.L. Hart, A.C. Lang, R.D. Twesten, M.L. Taheri, Advantages of Direct Detection and Electron Counting for Electron Energy Loss Spectroscopy Data Acquisition and the Quest of Extremely High-Energy Edges Using Eels, Microsc. Microanal. 23 (2017) 60–61. https://doi.org/10.1017/S1431927617000988.

[31] Z.J.W.A. Leijten, A.D.A. Keizer, G. De With, H. Friedrich, Quantitative Analysis of Electron Beam Damage in Organic Thin Films, J. Phys. Chem. C. 121 (2017) 10552–10561. https://doi.org/10.1021/acs.jpcc.7b01749.

[32] R.F. Egerton, Mechanisms of radiation damage in beam-sensitive specimens, for TEM accelerating voltages between 10 and 300 kV, Microsc. Res. Tech. 75 (2012) 1550–1556. https://doi.org/10.1002/JEMT.22099.

[33] D. Zhang, Y. Zhu, L. Liu, X. Ying, C.-E. Hsiung, R. Sougrat, K. Li, Y. Han, Atomic-resolution transmission electron microscopy of electron beam–sensitive crystalline materials, Science (80-.). 359 (2018) 675–679. https://doi.org/10.1126/SCIENCE.AAO0865.

[34] J.P. Winterstein, C.B. Carter, Electron-beam damage and point defects near grain boundaries in cerium oxide, J. Eur. Ceram. Soc. 34 (2014) 3007–3018. https://doi.org/10.1016/j.jeurceramsoc.2014.02.017.

[35] S.R. Spurgeon, C. Ophus, L. Jones, A. Petford-Long, S. V. Kalinin, M.J. Olszta, R.E. Dunin-Borkowski, N. Salmon, K. Hattar, W.-C.D. Yang, R. Sharma, Y. Du, A. Chiaramonti, H. Zheng, E.C. Buck, L. Kovarik, R.L. Penn, D. Li, X. Zhang, M. Murayama, M.L. Taheri, Towards data-driven next-generation transmission electron microscopy, Nat. Mater. 2020 203. 20 (2020) 274–279. https://doi.org/10.1038/S41563-020-00833-Z.

[36] I. Jolliffe, Principal Component Analysis, Encycl. Stat. Behav. Sci. (2005). https://doi.org/10.1002/0470013192.BSA501.

[37] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 374 (2016). https://doi.org/10.1098/RSTA.2015.0202.

[38] F. de la Peña, M.H. Berger, J.F. Hochepied, F. Dynys, O. Stephan, M. Walls, Mapping titanium and tin oxide phases using EELS: An application of independent component analysis, Ultramicroscopy. 111 (2011) 169–176. https://doi.org/10.1016/j.ultramic.2010.10.001.

[39] S. Estradé, L. Yedra, A. López-Ortega, M. Estrader, G. Salazar-Alvarez, M.D. Baró, J. Nogués, F. Peiró, Distinguishing the core from the shell in MnO x/MnO y and FeO x/MnO x core/shell nanoparticles through quantitative electron energy loss spectroscopy (EELS) analysis, Micron. 43 (2012) 30–36. https://doi.org/10.1016/j.micron.2011.04.002.

[40] G. Haberfehlner, S.F. Hoefler, T. Rath, G. Trimmel, G. Kothleitner, F. Hofer, Benefits of direct electron detection and PCA for EELS investigation of organic photovoltaics materials, Micron. 140 (2021) 102981. https://doi.org/10.1016/J.MICRON.2020.102981.

[41] V.P. Pauca, J. Piper, R.J. Plemmons, Nonnegative matrix factorization for spectral data analysis, Linear Algebra Appl. 416 (2006) 29–47. https://doi.org/10.1016/J.LAA.2005.06.025.

[42] Y.X. Wang, Y.J. Zhang, Nonnegative matrix factorization: A comprehensive review, IEEE Trans. Knowl. Data Eng. 25 (2013) 1336–1353. https://doi.org/10.1109/TKDE.2012.51.

[43] M. Shiga, K. Tatsumi, S. Muto, K. Tsuda, Y. Yamamoto, T. Mori, T. Tanji, Sparse modeling of EELS and EDX spectral imaging data by nonnegative matrix factorization, Ultramicroscopy. 170 (2016) 43–59. https://doi.org/10.1016/J.ULTRAMIC.2016.08.006.

[44] P. Trebbia, N. Bonnet, EELS elemental mapping with unconventional

methods I. Theoretical basis: Image analysis with multivariate statistics and entropy concepts, Ultramicroscopy. 34 (1990) 165–178. https://doi.org/10.1016/0304-3991(90)90070-3.

[45] M. Bosman, V.J. Keast, J.L. García-Muñoz, A.J. D'Alfonso, S.D. Findlay, L.J. Allen, Two-dimensional mapping of chemical information at atomic resolution, Phys. Rev. Lett. 99 (2007) 1–4. https://doi.org/10.1103/PhysRevLett.99.086102.

[46] K.J. Dudeck, M. Couillard, S. Lazar, C. Dwyer, G.A. Botton, Quantitative statistical analysis, optimization and noise reduction of atomic resolved electron energy loss spectrum images, Micron. 43 (2012) 57–67. https://doi.org/10.1016/J.MICRON.2011.07.008.

[47] P. Torruella, M. Estrader, A. López-Ortega, M.D. Baró, M. Varela, F. Peiró, S. Estradé, Clustering analysis strategies for electron energy loss spectroscopy (EELS), Ultramicroscopy. 185 (2018) 42–48. https://doi.org/10.1016/J.ULTRAMIC.2017.11.010.

[48] P. Potapov, Why Principal Component Analysis of STEM spectrum-images results in "abstract", uninterpretable loadings?, Ultramicroscopy. 160 (2016) 197–212. https://doi.org/10.1016/j.ultramic.2015.10.020.

[49] R.-F. Cai, M.-T. Chang, S.-C. Lo, C.-C. Chen, Novel spectral unmixing approach for electron energy-loss spectroscopy, New J. Phys. 22 (2020) 033029. https://doi.org/10.1088/1367-2630/AB7A89.

[50] L. der Maaten, G. Hinton, Visualizing data using t-SNE., J. Mach. Learn. Res. 9 (2008).

[51] F.H.M. Oliveira, A.R.P. MacHado, A.O. Andrade, On the Use of t -Distributed Stochastic Neighbor Embedding for Data Visualization and Classification of Individuals with Parkinson's Disease, Comput. Math. Methods Med. 2018 (2018). https://doi.org/10.1155/2018/8019232.

[52] AnkerstMihael, B. M., KriegelHans-Peter, SanderJörg, OPTICS, ACM

SIGMOD Rec. 28 (1999) 49–60. https://doi.org/10.1145/304181.304187.

[53] E. Schubert, M. Gertz, Improving the cluster structure extracted from OPTICS plots, CEUR Workshop Proc. 2191 (2018) 318–329.

[54] J. Ryu, H. Kim, R.M. Kim, S. Kim, J. Jo, S. Lee, K.T. Nam, Y.-C. Joo, G.-C. Yi, J. Lee, M. Kim, Dimensionality reduction and unsupervised clustering for EELS-SI, Ultramicroscopy. (2021) 113314. https://doi.org/10.1016/J.ULTRAMIC.2021.113314.

[55] L. Mcinnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 2020.

[56] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering, J. Open Source Softw. 2 (2017) 205. https://doi.org/10.21105/joss.00205.

[57] J. Blanco-portals, F. Peiró, S. Estradé, Strategies for EELS Data Analysis . Introducing UMAP and HDBSCAN for Dimensionality Reduction and Clustering, Microsc. Microanal. (2021) 1–14. https://doi.org/10.1017/S1431927621013696.

[58] J. Blanco-Portals, P. Torruella, F. Baiutti, S. Anelli, M. Torrel, A. Tarancón, F. Peiró, S. Estradé, WhatEELS . A new python-based interactive software solution for ELNES analysis combining clustering and NLLS, Ultramicroscopy. 232 (2021) 113403. https://doi.org/10.1016/j.ultramic.2021.113403.

[59] E.P. Francisco de la Peña Vidar Tonaas Fauske, Pierre Burdet, Petras Jokubauskas, Magnus Nord, … Andreas Garmannslund, hyperspy/hyperspy: HyperSpy v1.5.2, (2019). https://doi.org/10.5281/zenodo.3396791.

[60] F. Baiutti, J. Blanco-Portals, S. Anelli, P. Torruella, M. López-Haro, J.J. Calvino, S. Estradé, M. Torrell, F. Peiró, A. Tarancón, J.B. Portals, S. Anelli, P. Torruella, J.J. Calvino, Tailoring the Transport Properties of

Mesoporous Doped Cerium Oxide for Energy Applications, J. Phys. Chem. C. 125 (2021) 16451–16463. https://doi.org/10.1021/acs.jpcc.1c04861.

[61] A.B. Hungría, J.J. Calvino, J.C.H. Garrido, HAADF - STEM Electron Tomography in Catalysis Research, Top. Catal. 62 (2019) 808–821. https://doi.org/10.1007/s11244-019-01200-2.

[62] R. Leary, Z. Saghi, M. Armbruster, G. Wowsnick, R. Schlogl, J.M. Thomas, P.A. Midgley, Quantitative high-angle annular dark-field scanning transmission electron microscope (HAADF-STEM) tomography and high-resolution electron microscopy of unsupported intermetallic GaPd2 catalysts, J. Phys. Chem. C. 116 (2012) 13343–13352. https://doi.org/10.1021/jp212456z.

[63] J.C. Hernandez, A.B. Hungria, A. Perez-Omil, S. Trasobares, S. Bernal, P.A. Midgley, A. Alavi, J.J. Calvino, Structural Surface Investigations of Cerium - Zirconium Mixed Oxide Nanocrystals with Enhanced Reducibility, J. Phys. Chem. C. 111 (2007) 9001–9004. https://doi.org/10.1021/jp072466a.

[64] D. Eder, A.H. Windle, P.A. Midgley, A.B. Hungria, Visualization of the three-dimensional microstructure of TiO 2 nanotubes by electron tomography, Catal. Today. 143 (2009) 225–229. https://doi.org/10.1016/j.cattod.2008.09.014.

[65] B. Goris, S. Bals, W. Van Den Broek, E. Carbó-argibay, S. Gómez-graña, L.M. Liz-marzán, G. Van Tendeloo, Atomic-scale determination of surface facets in gold nanorods, Nat. Mater. 11 (2012) 930–935. https://doi.org/10.1038/nmat3462.

[66] M.C. Scott, C. Chen, M. Mecklenburg, C. Zhu, R. Xu, P. Ercius, U. Dahmen, B.C. Regan, J. Miao, Electron tomography at 2.4-angstrom resolution., Nature. 483 (2012) 444–447. https://doi.org/10.1038/nature10934.

[67] K. Lepinay, F. Lorut, R. Pantel, T. Epicier, Chemical 3D tomography of

28 nm high K metal gate transistor : STEM XEDS experimental method and results, Micron. 47 (2013) 43–49. https://doi.org/10.1016/j.micron.2013.01.004.

[68] S. Van Aert, K.J. Batenburg, M.D. Rossell, R. Erni, G. Van Tendeloo, Three-dimensional atomic imaging of crystalline nanoparticles, Nature. 470 (2011) 374–377. https://doi.org/10.1038/nature09741.

[69] S. Bals, S. Van Aert, G. Van Tendeloo, High resolution electron tomography, Curr. Opin. Solid State Mater. Sci. 17 (2013) 107–114. https://doi.org/10.1016/j.cossms.2013.03.001.

[70] A. De Backer, G.T. Martinez, K.E. Macarthur, L. Jones, A. Béché, P.D. Nellist, S. Van Aert, Dose limited reliability of quantitative annular dark fi eld scanning transmission electron microscopy for nano-particle atom-counting, Ultramicroscopy. 151 (2015) 56–61. https://doi.org/10.1016/j.ultramic.2014.11.028.

[71] S.M. Collins, P.A. Midgley, Progress and opportunities in EELS and EDS tomography, Ultramicroscopy. 180 (2017) 133–141. https://doi.org/10.1016/j.ultramic.2017.01.003.

[72] G. Haberfehlner, A. Orthacker, M. Albu, J. Li, G. Kothleitner, Nanoscale voxel spectroscopy by simultaneous EELS and EDS tomography, Nanoscale. 6 (2014) 14563–14569. https://doi.org/10.1039/c4nr04553j.

[73] T.J.A. Slater, A. Janssen, P.H.C. Camargo, M.G. Burke, N.J. Zaluzec, S.J. Haigh, STEM-EDX tomography of bimetallic nanoparticles : A methodological investigation, Ultramicroscopy. 162 (2016) 61–73. https://doi.org/10.1016/j.ultramic.2015.10.007.

[74] A. Genc, L. Kovarik, M. Gu, H. Cheng, P. Plachinda, L. Pullan, B. Freitag, C. Wang, XEDS STEM tomography for 3D chemical characterization of nanoscale particles, Ultramicroscopy. 131 (2013) 24–32. https://doi.org/10.1016/j.ultramic.2013.03.023.

[75] R. Grothausmann, G. Zehl, I. Manke, S. Fiechter, P. Bogdanoff, I. Dorbandt, A. Kupsch, A. Lange, M.P. Hentschel, G. Schumacher, J. Banhart, Quantitative Structural Assessment of Heterogeneous Catalysts by Electron Tomography, J. Am. Chem. Soc. 133 (2011) 18161–18171. https://doi.org/10.1021/ja2032508.

[76] L. Yedra, A. Eljarrat, R. Arenal, E. Pellicer, M. Estrader, J. Sort, M.D. Baró, S. Estradé, F. Peiró, Electron energy-loss spectroscopic tomography of Fe x Co (3− x) O 4 impregnated Co 3 O 4 mesoporous particles: unraveling the chemical information in three dimensions, Analyst. 141 (2016) 4968–4972. https://doi.org/10.1039/c6an00562d.

[77] P. Torruella, R. Arenal, F. de la Peña, Z. Saghi, L. Yedra, A. Eljarrat, L. López-Conesa, M. Estrader, A. López-Ortega, G. Salazar-Alvarez, J. Nogués, C. Ducati, P.A. Midgley, F. Peiró, S. Estradé, 3D Visualization of the Iron Oxidation State in FeO/Fe 3 O 4 Core–Shell Nanocubes from Electron Energy Loss Tomography, Nano Lett. 16 (2016) 5068–5073. https://doi.org/10.1021/acs.nanolett.6b01922.

[78] B. Goris, S. Turner, S. Bals, G. Van Tendeloo, Three-Dimensional Valency Mapping in Ceria Nanocrystals, ACS Nano. 8 (2014) 10878–10884. https://doi.org/10.1021/nn5047053.

[79] B. Goris, M. Meledina, S. Turner, Z. Zhong, K.J. Batenburg, S. Bals, Three dimensional mapping of Fe dopants in ceria nanocrystals using direct spectroscopic electron tomography, Ultramicroscopy. 171 (2016) 55–62. https://doi.org/10.1016/j.ultramic.2016.08.017.

[80] W. Albrecht, S. Bals, Fast electron tomography for nanomaterials, J. Phys. Chem. C. 124 (2020) 27276–27286. https://doi.org/10.1021/acs.jpcc.0c08939.

[81] F. Eisenstein, R. Danev, M. Pilhofer, Improved applicability and robustness of fast cryo-electron tomography data acquisition, J. Struct. Biol. 208 (2019) 107–114. https://doi.org/10.1016/j.jsb.2019.08.006.

[82] S. Sheibani, K. Basu, A. Farnudi, A. Ashkarran, M. Ichikawa, J.F. Presley, K.H. Bui, M.R. Ejtehadi, H. Vali, M. Mahmoudi, Nanoscale characterization of the biomolecular corona by cryo-electron microscopy, cryo-electron tomography, and image simulation, Nat. Commun. 12 (2021) 1–9. https://doi.org/10.1038/s41467-020-20884-9.

[83] R.I. Koning, A.J. Koster, T.H. Sharp, Advances in cryo-electron tomography for biology and medicine, Ann. Anat. 217 (2018) 82–96. https://doi.org/10.1016/j.aanat.2018.02.004.

[84] J. Plitzko, W.P. Baumeister, Cryo-electron tomography, in: Springer Handb. Microsc., Springer, 2019: pp. 189–228. https://doi.org/https://doi.org/10.1007/978-3-030-00069-1_4.

[85] P.S. Shen, The 2017 Nobel Prize in Chemistry: cryo-EM comes of age, Anal. Bioanal. Chem. 410 (2018) 2053–2057. https://doi.org/https://doi.org/10.1007/s00216-018-0899-8.

[86] V. Migunov, H. Ryll, X. Zhuge, M. Simson, L. Struder, K.J. Batenburg, L. Houben, R.E. Dunin-Borkowski, Rapid low dose electron tomography using a direct electron detection camera, Sci. Rep. 5 (2015) 1–5. https://doi.org/10.1038/srep14516.

[87] P. Schlossmacher, D.O. Klenov, B. Freitag, H.S. von Harrach, Enhanced Detection Sensitivity with a New Windowless XEDS System for AEM Based on Silicon Drift Detector Technology, Micros. Today. 18 (2010) 14–20. https://doi.org/10.1017/s1551929510000404.

[88] T.J.A. Slater, P.H.C. Camargo, M.G. Burke, N.J. Zaluzec, S.J. Haigh, Understanding the limitations of the Super-X energy dispersive x-ray spectrometer as a function of specimen tilt angle for tomographic data acquisition in the S/TEM, J. Phys. Conf. Ser. 522 (2014) 12025. https://doi.org/10.1088/1742-6596/522/1/012025.

[89] K.J. Batenburg, S. Bals, J. Sijbers, C. Kübel, P.A. Midgley, J.C. Hernandez,

U. Kaiser, E.R. Encina, E.A. Coronado, G. Van Tendeloo, 3D imaging of nanomaterials by discrete tomography, Ultramicroscopy. 109 (2009) 730–740. https://doi.org/10.1016/J.ULTRAMIC.2009.01.009.

[90] T. Roelandts, K.J. Batenburg, E. Biermans, C. Kübel, S. Bals, J. Sijbers, Accurate segmentation of dense nanoparticles by partially discrete electron tomography, Ultramicroscopy. 114 (2012) 96–105. https://doi.org/10.1016/J.ULTRAMIC.2011.12.003.

[91] R. Leary, Z. Saghi, P.A. Midgley, D.J. Holland, Compressed sensing electron tomography, Ultramicroscopy. 131 (2013) 70–91. https://doi.org/10.1016/J.ULTRAMIC.2013.03.019.

[92] M. López-Haro, M. Tinoco, S. Fernández-Garcia, X. Chen, A.B. Hungria, M.Á. Cauqui, J.J. Calvino, A Macroscopically Relevant 3D-Metrology Approach for Nanocatalysis Research, Part. Part. Syst. Charact. 35 (2018) 1700343. https://doi.org/10.1002/ppsc.201700343.

[93] J.M. Muñoz-ocaña, A. Bouziane, F. Sakina, R.T. Baker, A.B. Hungría, J.J. Calvino, A.M. Rodríguez-chía, M. López-haro, Optimization of STEM-HAADF Electron Tomography Reconstructions by Parameter Selection in Compressed Sensing Total Variation Minimization-Based Algorithms, Part. Part. Syst. Charact. 37 (2020) 2000070. https://doi.org/10.1002/ppsc.202000070.

[94] S. Klein, M. Cortese, S.L. Winter, M. Wachsmuth-Melm, C.J. Neufeldt, B. Cerikan, M.L. Stanifer, S. Boulant, R. Bartenschlager, P. Chlanda, SARS-CoV-2 structure and replication characterized by in situ cryo-electron tomography, Nat. Commun. 11 (2020) 1–10. https://doi.org/10.1038/s41467-020-19619-7.

[95] P. Gilbert, Iterative methods for three dimensional reconstruction of an object from projections, J. Theor. Biol. 36 (1972) 105–117. https://doi.org/10.1016/0022-5193(72)90180-4.

[96] K. Sentosun, M.N. Sanz Ortiz, K.J. Batenburg, L.M. Liz-Marzán, S. Bals,

Combination of HAADF-STEM and ADF-STEM Tomography for Core-Shell Hybrid Materials, Part. Part. Syst. Charact. 32 (2015) 1063–1067. https://doi.org/10.1002/ppsc.201500097.

[97] Z. Saghi, G. Divitini, B. Winter, R. Leary, E. Spiecker, C. Ducati, P.A. Midgley, Compressed sensing electron tomography of needle-shaped biological specimens – Potential for improved reconstruction fidelity with reduced dose, Ultramicroscopy. 160 (2016) 230–238. https://doi.org/10.1016/J.ULTRAMIC.2015.10.021.

[98] P. Torruella, J. Blanco-Portals, L. Yedra, L. López-Conesa, J.M. Rebled, F. Peiró, S. Estradé, Electron Tomography, Springer Ser. Mater. Sci. 308 (2021) 257–283. https://doi.org/10.1007/978-3-030-60473-8_11.

[99] C. Li, W. Yin, H. Jiang, Y. Zhang, An efficient augmented Lagrangian method with applications to total variation minimization, Comput. Optim. Appl. 56 (2013) 507–530. https://doi.org/10.1007/s10589-013-9576-1.

[100] B. Bougher, Introduction to compressed sensing, Lead. Edge. (2015) 1256–1258. https://doi.org/http://dx.doi.org/10.1190/tle34101256.1.

[101] Z. Saghi, D.J. Holland, R. Leary, A. Falqui, G. Bertoni, A.J. Sederman, L.F. Gladden, P.A. Midgley, Three-Dimensional Morphology of Iron Oxide Nanoparticles with Reactive Concave Surfaces. A Compressed Sensing-Electron Tomography (CS-ET) Approach, Nano Lett. 11 (2011) 4666–4673. https://doi.org/10.1021/nl202253a.

[102] J.M. Thomas, R. Leary, P.A. Midgley, D.J. Holland, A new approach to the investigation of nanoparticles: Electron tomography with compressed sensing, J. Colloid Interface Sci. 392 (2013) 7–14. https://doi.org/10.1016/J.JCIS.2012.09.068.

[103] W. van Aarle, W.J. Palenstijn, J. De Beenhouwer, T. Altantzis, S. Bals, K.J. Batenburg, J. Sijbers, The ASTRA Toolbox: A platform for advanced

algorithm development in electron tomography, Ultramicroscopy. 157 (2015) 35–47. https://doi.org/10.1016/j.ultramic.2015.05.002.

[104] F. Bleichrodt, T. van Leeuwen, W.J. Palenstijn, W. van Aarle, J. Sijbers, K.J. Batenburg, Easy implementation of advanced tomography algorithms using the ASTRA toolbox with Spot operators, Numer. Algorithms 2015 713. 71 (2015) 673–697. https://doi.org/10.1007/S11075-015-0016-4.

[105] M. Dalmases, P. Torruella, J. Blanco-Portals, A. Vidal, M. Lopez-Haro, J.J. Calvino, S. Estradé, F. Peiró, A. Figuerola, Gradual Transformation of Ag2S to Au2S Nanoparticles by Sequential Cation Exchange Reactions: Binary, Ternary, and Hybrid Compositions, Chem. Mater. 30 (2018) 6893–6902. https://doi.org/10.1021/acs.chemmater.8b03208.

Chapter 2

Objectives.

The objectives of this PhD thesis were, first and foremost, centred around the development of new analysis strategies for the characterization of nanomaterials via electron energy loss spectroscopy (EELS) and analytical electron tomography (ET).

Much like the structure of the manuscript, and of the overall work done during the PhD thesis, the specific formulation of the objectives can be divided in three main blocks: (1) the implementation and use of compressed sensing algorithms for ET reconstructions, (2) the exploration of advanced clustering and non-linear dimensionality reduction techniques for the analysis of EELS datasets, (3) and the implementation of new software solutions for energy loss near edge structures (ELNES) analysis. (4) A fourth objective is the application of these developments to resolving the localized changes in the Ce oxidation state and Pr-Gd dopant segregation towards grain boundaries (GB) and surfaces (GS) in mesoporous CeO_2.

Every single one of these objectives is formulated individually from a purely theoretical point of view in the following lines. However, they all responded to real experimental necessities that evolved along with the progression of the PhD thesis (i.e., the difficulties encountered along the way gave shape to some of the objectives set for the completion of the PhD thesis).

(1) Electron tomography. Implementation and use of a compressed sensing algorithm (TVM) (Chapter 3. Chapter 6, section 6.3).

The first objective of this PhD thesis was the implementation of the TVAL3 algorithm in Python. This solver of the total variation

minimization (TVM) problem has been repeatedly depicted in the literature as the central piece of a reconstruction method capable of achieving unparalleled high-quality volumetric and analytical reconstructions at the nanoscale, dealing in many cases with a very limited number of projections and low signal-to-noise-ratio (SNR) datasets.

The initial focus was set in the translation of this specific algorithm into a free distributed programming language (Python), given that all its previous implementations as an ET reconstruction method in the literature implied the use of the proprietary MATLAB programming language.

Another major objective was to test its reported capabilities, comparing them with the understood iterative reconstruction methods (i.e., SIRT). The end goal was to carry out a series of 3D reconstructions (both morphological and analytical) in real experimental samples with a severe constraint in the number of allowed projections acquired.

(2) Electron energy loss spectroscopy. New and advanced clustering and non-linear dimensionality reduction techniques (Chapter 4. Chapter 6, sections 6.2, 6.3 and 6.4).

The major second objective of this PhD thesis was the exploration of new routes for EELS data analysis via the use of advanced clustering and dimensionality reduction algorithms. Two of the most promising candidates for these fields of analysis are, respectively, the hierarchical density-based spatial clustering of applications with noise (HDBSCAN) and the Uniform Manifold Approximation and Projection for Dimension Reduction (UMAP).

A successful incorporation of these algorithms for EELS analysis became a major focus of the work done towards the end of the PhD thesis, when an efficient Python implementation that showed an

outstanding performance dealing with high dimensional datasets became available for these state-of-the-art algorithms.

In general, this objective follows the clear trend in data analysis of incorporating machine learning algorithms to the mainstream characterization processes. This tendency is heavily influenced by the increment in size and complexity of the acquired datasets, that continue their upwards progression pushed by the new developments of the instrumental capabilities. Also, part of the interest in these types of techniques for EELS analysis rests on their general ability to produce classification results in an unsupervised manner. In many instances, this capability might help with the reduction of possible interpretation biases, which is a valuable addition in the path towards a fully data driven analysis.

Furthermore, the field of EELS is not alien to the inclusion of machine learning algorithms that deal with the detailed analysis of large spectrum images (SI). Some examples would be the use of PCA and NMF as dimensionality reduction techniques for the discovery of the elemental distributions in EELS SI, or the use of the hierarchical agglomerative clustering or K-means algorithms for an SI segmentation. With the introduction of UMAP and HDBSCAN, a systematic revision of the performance of the older methodologies and a comparison of the results attained by some of the possible combinations of algorithms for the task of spectral classification also became a major objective.

(3) Electron energy loss spectroscopy. Development of new software solutions for ELNES analysis (Chapter 5. Chapter 6.3 and 6.4).

The third major objective for this PhD thesis was the development of a Python based software solution for the ELNES analysis in transition metal and rare earth oxides EELS SI. In general, and up to this point, the majority of analysis procedures in the literature relied on the already

available software solutions, namely HyperSpy (free access and programmed in Python) and Digital Micrograph (proprietary software from Gatan). However, in recent times they fell behind in the inclusion of advanced clustering analysis techniques for EELS characterization. Furthermore, they showed early on in the path of this PhD the inability in many cases to produce satisfactory results when dealing with an in-depth analysis of ELNES thought a non-linear least squares (NLLS) fitting.

At the same time, and as part of the objective of exploring new routes for the ELNES analysis, the combination of clustering and non-linear least squares (NLLS) fitting was proposed. This combination was thought of as a possible solution to the common pitfalls that the latter (NLLS) presents when dealing with large datasets with an intricate combination of different components to describe the whole richness of the ELNES for datasets with several different elements under analysis.

The objective was then set on the programming of an integral software tool that included all these major advances in clustering analysis and allowed an easy transition of the segmentation results to the NLLS fitting process (which would ultimately become the **WhatEELS** software solution).

(4) The analysis of the set of Gd-Pr doped CeO_2 mesoporous materials (Chapter 6).

This experimental analysis constitutes an individual objective by its own. The complexity of the characterization required (the elemental quantification of nearby and overlapping edges for the Ce, Pr and Gd elements and the oxidation state analysis of the Ce in these samples prone to be reduced under the electron beam) pushed the development of the analysis techniques introduced in the other objectives. The analysis of these samples effectively constitutes the main driving force

behind a large portion of the development of techniques in this PhD thesis.

Chapter 3

Total variation minimization. Optimizing morphological and analytical electron tomography reconstructions.

3.1. Introduction.

Electron Tomography (ET) in the Transmission Electron Microscope (TEM) refers to the problem of reconstructing a 3D volume from a set of 2D projection images experimentally acquired for a finite angular range. Historically, the TEM tomography was first applied to the field of biology[1], but the operational and instrumental advances in the last couple of decades, as well as the formulation of new reconstruction algorithms, have introduced ET into the realm of materials and physical sciences. Nowadays, it plays a central role in many applications centred around the study and fabrication of nanostructured materials, such as the field of nanomagnetism in metal oxide nanoparticles[2].

Initially, the ET analysis was restricted to the morphological study of structures. It was quickly understood that the nature of the reconstruction problem and the general formulation of the algorithms imposed a series of restrictions over the valid types of signals experimentally acquired. These signals had to fulfil the so-called *Projection Requirement*, which states that the contrast in the set of projection images must change monotonically with a given (and unspecified) property of the sample (e.g. thickness or Z number)[3]. In a nutshell, the absorption function of the electron beam for a given sample and imaging mode ($\mu(x, y, z)$, in cartesian coordinates) must be proportional

to the point-density function of such sample ($\rho(x, y, z)$)[4] for the imaging mode to constitute a valid candidate for ET. Hence, the high resolution TEM (HRTEM) images, formed by a phase contrast process[5], were automatically discarded. This fact left the bright field (BF) and dark field (DF) diffraction contrast imaging modes as the only initial candidates for ET reconstructions. Furthermore, in the case of crystalline samples, the scanning-TEM high-angle annular dark field (STEM-HAADF) images were preferred, as the inherent incoherent nature of the HAADF signal always depends monotonically on the thickness and the Z number[5].

As the fields of Electron Energy Loss Spectroscopy (EELS)[6] and Energy-Dispersive X-Ray Spectroscopy (X-EDS)[5] flourish, alongside the improvements on the scanning speeds and the reduction of the exposure times required by the more advanced an sensitive detectors, the generalized use of the combination of these analytical techniques and ET is being spread to an ever-growing number of applications. The spectrum images acquired in both spectroscopic analysis techniques cited above fulfil the projection requirement and, thus, are valid candidates for the ET reconstruction. The birth to this analytical-ET field of study allowed the access to the characterization of the complete spectrum volume for any given sample[7]. This fourth dimension of analysis included by the analytical-ET opens new perspectives for the characterization of nanomaterials, as it facilitates some very powerful characterization routes, such as the 3D identification of the cationic oxidation states in nanoparticles[2,8]. It is also a good complementary analysis technique for the morphological ET in cases were the Z-contrast differences between areas of varying elemental composition is not enough to sperate them (e.g., in mixtures of different transition metal oxides, where the cations are close in the periodic table of elements).

This chapter summarizes the work done in the context of this PhD thesis to adopt part of the computational advances in the reconstruction process for the field of electron tomography. The focus was set on the algorithms based on the minimization of the total variation (TVM) quantity, which is formulated on top

of the compress sensing (CS) theoretical framework. The actual solver (algorithm) for the minimization problem is called TVAL3[9], and it constitutes one of the most popular approaches for the minimization problems tackled in an ET reconstruction experiment[10–12]. The bulk of the actual work was devoted to the development of a new code in Python language, starting from the original version in MATLAB, to allow its free distribution in a non-proprietary language platform in the future.

These new algorithms based on the compressed sensing (CS) theory present several advantages over their historical predecessors. For once, they allow a dramatic decrease in the number of acquired image projections without losing quality in reconstruction. This would relax the acquisition constraints in long exposure experiments (analytical mappings), resolving the excess of electron dose problem, one of the major drawbacks in analytical electron tomography experiments. To allow a deeper understanding of the actual advantages brought by these CS reconstruction algorithms, a minor review of the historical developments in the field is included, alongside some comparative examples of use over a known synthetic dataset and a real case scenario combining morphological and analytical ET reconstructions.

The chapter is structured as follows:

- ***Mathematical principles and reconstruction methods.***

  A general historical overview is given for the most relevant algorithms employed in tomography reconstructions. This section is ended with the introduction of the theoretical background for the TVM methodology, emphasizing its potential advantages over the most commonly spread algorithms.

- ***A qualitative comparison of SIRT and TVM .***

  A synthetic dataset is devised to test the translated TVAL3 algorithm to Python, as part of the TVM methodology for electron tomography reconstructions. This phantom and the set of experiments included are

created with the intention to test the most commonly observed artefacts in ET reconstructions, and compare the capabilities of the default 'go-to' iterative algebraic algorithms with the TVM in a controlled manner.

- ***A case of study.***

  A set of hollow iron oxide and nickel-iron oxide NPs are characterized making use of the TVM methodology for ET and analytical ET.

3.2. Mathematical principles and reconstruction methods.

3.2.1. Classic approaches and iterative algebraic algorithms.

The first methodologies proposed to retrieve a 3D reconstruction of an object from a set of 2D projections in a TEM (i.e., an electron tomography experiment, ET) were based on Fourier transform theory[4,13]. As such, they were known as **Fourier space methods** or algorithms. These methods took advantage of the well-known Fourier transform mathematical foundations and the properties of the Fourier (reciprocal) space described by them. Specifically, these methodologies were based on the so-called ***central section theorem***, which says: the projection of an object (i.e., a projected density function) at any given angle is a central section through the Fourier transform of that object. Hence, the Fourier space of an object could be sampled experimentally by calculating the Fourier transform of a large number of projection images at different angles in the TEM.

Given the discrete nature of an experimental procedure (i.e., only a finite and limited number of projections can be recorded during an experiment), the reconstruction algorithms always had to deal with some degree of missing information. In the case of the Fourier methods, this missing information would be recovered via interpolation in the reciprocal space, where the final solution is achieved by solving a least squares minimization problem. After reconstructing the object of interest in the Fourier space, the actual density

function in the direct (real) space would be recovered by an inverse Fourier transform.

Although theoretically and historically relevant, the algorithms based in these types of methodologies were quickly displaced in practical applications by the ones that reconstructed the object directly in the image space, as they showed better performance in under-sampled experiments. In fact, nowadays it is difficult to find any software solution that uses algorithms based on the Fourier space properties for electron tomography.

The methodologies using the real image space projections for the tomographic reconstruction are commonly denominated as **direct space methods**. As in the case of the **Fourier space methods**, an underlying mathematical foundation was already in place to be used: the Radon transform formulation [4,14,15].

For instance, let us consider that the object to be reconstructed can be described by a density function $\rho(x, y, z)$. In a tomography experiment a set of projections σ_{φ} would be acquired by rotating the sample an angle (φ) about a common given axis (z), recording the images as in a conventional TEM experiment. However, mathematically (and conceptually) it is easier to formulate the equivalent problem of rotating the detection plane an angle ϕ and, thus, enforcing a solidary rotation of the illumination direction (i.e., the electron beam). Now, a new set of rotating axes x', y' and z'=z is used to describe the projection space (see **Figure 3.1**). The direction of illumination would always be fixed in the y' axis, and the detection planes would always be perpendicular to y'. Furthermore, each one of the projections of the original

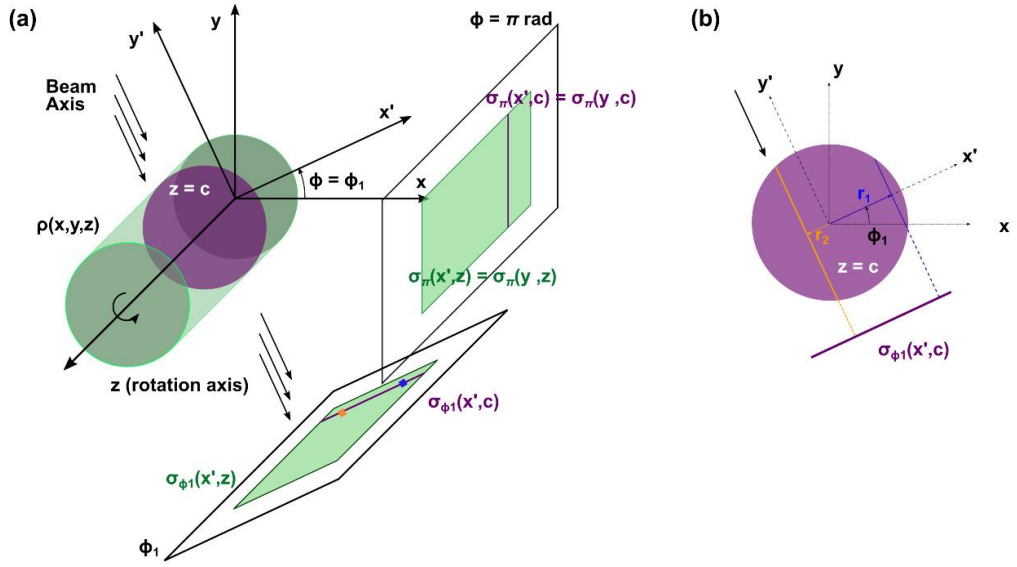


Figure 3.1. Scheme of an electron tomography experiment and the Radon transform formulation. (a) 3D schematic. (b) Slice at z = c through the volume ρ(x,y,z) showcased in (a). The different line paths of integration in (b) are marked with different colours.

$\rho(x, y, z)$ density function is described by a $\sigma_\phi(x', z)$ projection density function. Each of these projections at an angle ϕ would actually be a sample of the Radon transform space.

To prove this relation, one can consider a slice through the volume $\rho(x, y, z)$ at z = c (i.e., perpendicular to the rotation axis, $\rho(x, y, z = c)$). The projection of this slice $\sigma_{\phi 1}(x', c)$ at an angle ϕ_1 would result in a line over the total projected image $\sigma_{\phi 1}$(x',z), which formally would be expressed as

$$\sigma_{\phi 1}(x', c) = \int_{-\infty}^{+\infty} \rho(x, y, c)\, dy' \qquad (1)$$

Thus, each one of the recorded intensities at every point in this projection line $\sigma_{\phi 1}(x', c)$ would be the result of a path integral following an 'electron ray' in the y' direction (illumination) traversing the slice of the sample at z = c.

The relation of this formulation with the line-integral described in the Radon transform theory[14] is straight forward. Indeed, the Radon transform for a 2D function $f(x, y)$ is formulated as

$$\check{f}(r,\theta) = \mathcal{R}_f = \int_L f(x,y)ds \tag{2}$$

where $\check{f}$ is a common notation for the transform of $f(x,y)$ in the literature[4].

Introducing now polar (cylindrical) coordinates into the problem (r,ϕ,z) and taking into account the extra dimension in our 3D problem, all the integration path lines L (parallel to the y' axis direction) for a projection at an angle ϕ of the slice z = c can be described by a separation radius of r (perpendicular to the y' axis) and the actual rotation angle ϕ (see **Figure 3.1**(b)). Then, the total set of projection images for the original $\rho(x,y,z)$ volume and for a continuously varying angle ϕ would describe the Radon transform of the slice(c)

$$\check{\rho}(r,\phi,c) = \iint_{-\infty}^{\infty} \rho(x,y,z)\,\delta(r - x\cos\phi - y\sin\phi)\,dxdy \tag{3}$$

As $\check{\rho}$ can be calculated from the projection images in a TEM tomography experiment, the original $\rho(x,y,z)$ can be recovered for each one of the slices perpendicular to the rotation axis by calculating the inverse Radon transform of $\check{\rho}(r,\phi,c)$. Finally, the total volume $\rho(x,y,z)$ is reconstructed by stacking all the z = c slices together.

However, in a real experiment the number of projections, the resolution along the z axis (for the slices) and the detector array, are finite and described in a discrete way. Hence, different algorithms will take different approaches to mitigate the effects of this missing information.

Among all these different algorithms available, the so-called iterative algebraic methods were the ones mostly employed for tomography experiments in a wide range of different applications. Part of their success arises from the superior quality of the reconstructions obtained with respect to the Fourier methods[16], which was in part provided by the ease to integrate a priori known information into the reconstruction process. For example, these algorithms include the positivity constraint for the density of the object.

One of such algorithms is the so-called algebraic reconstruction technique (ART) algorithm[17]. It is formulated on the same basis of the Radon transform already described, but for a discrete space. Each one of the slices of the volume perpendicular to the rotation axis is now regarded as a grid of *n* x *n* points (*i* , *j*). In the reconstruction process, each (*i* , *j*) position of the slice will have an optical density $\rho_{i,j}$. The integration paths are now described by a k-set of illumination rays, fully characterized by the pair of coefficients (k, ϕ). The total sum of the optical densities within a given k-ray at an angle is $R_{k,\phi} = \sum \rho_{i,j}$. However, the experimental values for intensities of each projection ray are $P_{k,\phi}$ (which are measured from the projection images themselves). Therefore, obtaining the set of $\rho_{i,j}$ (n x n) from $P_{k,\phi}$, minimizing the discrepancy with $R_{k,\phi}$ is the actual problem resolved by ART.

To do so, an iterative scheme is proposed for the algorithm that can be updated by one of two different paths: multiplicative ART (mult.) and additive ART (add.):

$$ART\ Mult.: \quad {\rho_{i,j}}^{q+1} = \frac{\left(P_{k,\phi}\right)}{\left({R_{k,\phi}}^{q}\right)} {\rho_{i,j}}^{q} \tag{4.1}$$

$$ART\ Add.: \ {\rho_{i,j}}^{q+1} = max\left[{\rho_{i,j}}^{q} + \frac{\left(P_{k,\phi} - {R_{k,\phi}}^{q}\right)}{N_{k,\phi}}, 0\right] \tag{4.2}$$

where the initial guess for the optical density $\rho_{i,j}^{0} = T/n^2$ considers a homogeneous distribution of the total intensity *(T)* for each (slice). In the additive method, $N_{k,\phi}$ is the number of grid points *(i, j)* traversed by any given ray (k, ϕ) (a normalization parameter of sorts). Notice how both updating schemes intrinsically positively define $\rho_{i,j} \geq 0$. Thus, ART introduces the fact that the signal is positively constrained ($\rho(x, y, z) \geq 0$) into the calculation itself.

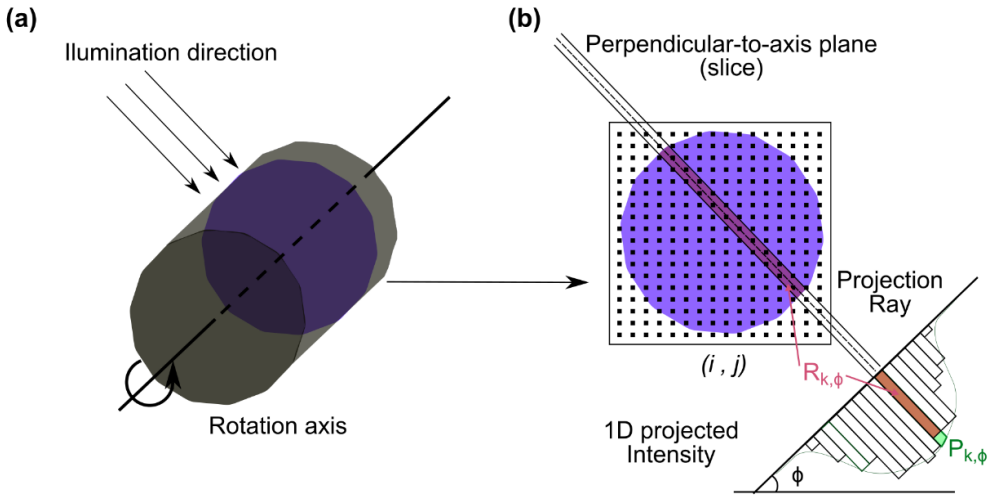


Figure 3.2. General scheme for the iterative algebraic methods ART and SIRT (a) 3D representation of the rotating object under analysis, the rotation axis and the illumination direction. (b) 2D slice through the object and 1D representation of the projected intensity curve of such slice.

A schematic representation of the general formulation for the ART algorithm is shown in **Figure 3.2**. The relation of ART with the theory described for the Radon transform is straightforward (compare both **Figure 3.1** and **Figure 3.2**).

The Simultaneous Iterative Reconstruction Technique (SIRT)[16] arises from the same mathematical framework as ART (see **Figure 3.2**). The main difference lies in the use of the information in the algorithm. ART gets an estimation of the $\rho_{i,j}$ at each iteration for each one of the projections involved (i.e., for each angle ϕ), using exclusively the information from that same projection. However, SIRT uses all the information from all the available projections at the same time at each iteration. Again, two update schemes (multiplicative and additive) were proposed for the SIRT algorithm

$$SIRT\ Mult. \qquad {\rho_{i,j}}^{q+1} = \frac{\sum(P_{k,\phi})}{\sum(L_{k,\phi})}\frac{\sum(N_{k,\phi})}{\sum\left(R_{k,\phi}^{q}\right)}{\rho_{i,j}}^{q} \tag{5.1}$$

$$SIRT\ Add. \qquad {\rho_{i,j}}^{q+1} = max\left[{\rho_{i,j}}^{q} + \frac{\sum(P_{k,\phi})}{\sum(L_{k,\phi})} - \frac{\sum({R_{k,\phi}}^{q})}{\sum(N_{k,\phi})}, 0\right] \tag{5.2}$$

where the regularization term $L_{k,\phi}$ is the length of the ray (k, ϕ).

One of the major advantages of the SIRT algorithm was its improved stability in noisy conditions, especially compared to the ART counterpart. The initial assessments of the performance of SIRT[13] indicated a slower convergence than the ART algorithm, but the quality of the results was overall better.

This elementwise updating scheme formulation introduced for the SIRT algorithm (as well as for the ART method) is both historically relevant and useful as a 'first contact' description of the inner workings of these types of methods. It gives as a solid base from which one can acquire an intuition of how the specific different algorithms incorporate the theoretical basis of the Radon transform for the electron tomography field of study.

However, colloquially speaking it has fallen out of date, as the modern electron tomography reconstruction techniques are almost ubiquitously presented via a matrix formulation. Following the same approach as before to present a simple basic example, let us consider that a 3D $\rho(u, v, z)$ volume can be deconstructed again as a set of parallel slices at z = s, which are perpendicular to a singular rotation axis (i.e., axis z, at different slicing position s). Now, each one of the slices through this volume $\rho(u, v, z = s)$ would be represented by a flattened vector $\vec{\boldsymbol{x}} \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{1}}$, where $n = w \times h$ would be the width and height dimensions of the slice through the volume (see **Figure 3.3**). The grid of coordinates is reshaped by a concatenation of the succession of rows in the slice.

The 2D image projections acquired experimentally for P-different angles $\phi_{p=\{1,\dots,P\}}$ can therefore be deconstructed as a set of projected lines of density. All these lines for each one of the slices $\vec{\boldsymbol{x}}$ considered would be encoded in a 'projection' vector $\vec{\boldsymbol{b}} \in \mathbb{R}^{\boldsymbol{m}\times\boldsymbol{1}}$, where the actual flattened dimensionality is $m = p \times k$. The finite nature of the detector (a discrete and finite array of pixels) is considered via the index k, and the finite number of acquired projections is taken into account by index p. As depicted by the schematic representation in

Figure 3.3(b), the actual notion of illumination ray is intertwined with the discrete nature of the detectors themselves (index k).

In this scheme, $\boldsymbol{A} \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{m}}$ is called the projection matrix. Each one of the elements of this projection matrix, $\{\boldsymbol{a_{i,j}} \mid i=(1,\dots,n), j=(1,..,m)\}$, accounts for the weight of any given pixel x_i to the projected ray at b_j (i.e., to a specific position in the projected line (k) for a specific given angle (p)). Therefore, the complete system can be described as

$$\boldsymbol{A}\vec{\boldsymbol{x}} = \vec{\boldsymbol{b}} \tag{6}$$

The tomography reconstruction problem becomes, thus, the extraction of $\vec{\boldsymbol{x}}$ from the measured $\vec{\boldsymbol{b}}$, without knowing a priori the values of $\boldsymbol{A}$. To further complicate the matter at hands, typically n >> m (i.e., the linear system described is, thus, undetermined and present an infinite number of possible solutions) and noise pollution of the signal $\vec{\boldsymbol{b}}$ is experimentally unavoidable. At the end, the tomography reconstruction usually evolves to a minimization problem in the implementation of most algorithms.

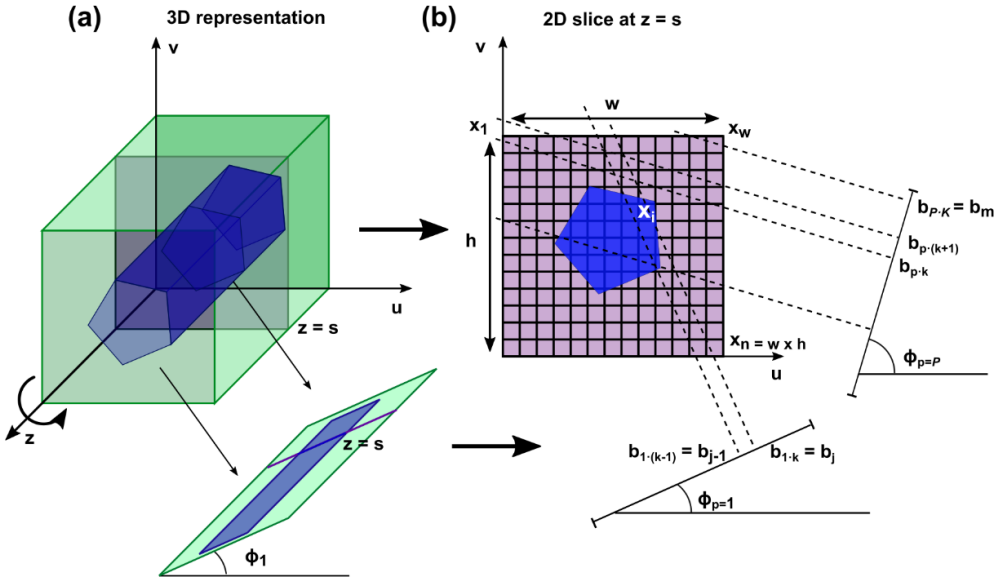


Figure 3.3. Schematic view for an electron tomography experiment expressed in a matrix for: (a) 3D representation of the volume space and deconstruction into a set of slices at z = s. (b) 2D representation of the tomography experiment reduced to the 2D slice at consideration.

For instance, in this matrix formulation the SIRT algorithm updating scheme would be expressed as follows (additive path)

$$\vec{\boldsymbol{x}}^{q+1} = \vec{\boldsymbol{x}}^{q} + \boldsymbol{C}\boldsymbol{A}^{T}\boldsymbol{R}\left(\vec{\boldsymbol{b}} - \boldsymbol{A}\vec{\boldsymbol{x}}^{q}\right) \tag{7}$$

where $\boldsymbol{C} \in \mathbb{R}^{\boldsymbol{n}\times\boldsymbol{n}}$ and $\boldsymbol{R} \in \mathbb{R}^{\boldsymbol{m}\times\boldsymbol{m}}$ are the inverted column and row sum diagonal matrices, that are calculated from

$$\begin{aligned} c_{ii} &= \frac{1}{\sum_j a_{i,j}} \\ r_{jj} &= \frac{1}{\sum_i a_{i,j}} \end{aligned} \tag{8}$$

This is the actual formulation used in the implementation of the SIRT algorithm for the free and Python-based ASTRA toolbox[18], that has been extensively used in every tomography experiment involved with this PhD work.

In the case of noiseless data (which is never the case for real experiments), the algorithm converges to a weighted least squared solution $\vec{\boldsymbol{x}}^{*}$

$$\vec{\boldsymbol{x}}^{*} = \arg\min_{\vec{\boldsymbol{x}}} \left\| A\vec{\boldsymbol{x}} - \vec{\boldsymbol{b}} \right\|_{\mathbf{R}}^{2} \tag{9}$$

where the operator $\|\vec{\boldsymbol{y}}\|_{\mathbf{R}}^{2} = \vec{\boldsymbol{y}}^{T}\boldsymbol{R}\,\vec{\boldsymbol{y}}$. Hence, the actual problem solved by the iterative scheme of SIRT is a minimization one.

Finally, although still relevant and commonly used in many applications in the tomography field (mainly due to its easy implementation and relatively fast convergence times compared to more advanced algorithms), the SIRT algorithm presents several shortcomings when it comes to tackle challenging experiments with a low signal-to-noise ratio or a small number of projections.

For instance, in many cases a balance has to be set between the angle step selected for the set of projections recorded (i.e., the total number of projections recorded) and the maximum total time of recording allowed (i.e., electron dose) for the experiment. This is an important trade of resolution for sample stability in materials with a high sensitivity to the electron beam and,

thus, prone to suffer structural changes with higher electron doses. At the end, the resolution of the reconstructed volume is compromised, as the algorithm struggles to converge to the optimum solution of eq.(9) due to the lack of information. These effects have to be considered especially in the case of analytical electron tomography experiments, where acquiring a full set of signal projection images usually require of large amount of time and electron dose absorbed by the sample.

Furthermore, the pole piece inside the TEM column poses a real experimental physical constraint, making practically impossible to cover the whole range of projection angles $\pm 90^{o}$ with conventional sample holders (i.e., it is impossible to sample the entirety of the Radon transform space). The under-sampling in the illumination direction caused by this experimental constraint is responsible for the most predominant artefact in a TEM-tomography reconstruction: *the missing wedge*[3]. For short, the missing wedge degrades the resolution of the reconstructions in an anisotropic way, causing an elongation of the volume in the illumination direction. There are several approaches to deal with the effect of the missing wedge. Experimentally, the preparation of samples in a needle-shape[19,20] is one possible solution (to cover the $\pm 90^{o}$tilting range), but in principle excludes the study of nanoparticles. Double tilt experiments[21] are another viable option, reducing the non-sampled projection space delimited by the missing wedge to a non-sampled space delimited by a *missing cone,* at the cost of an increased difficulty of image acquisition (i.e., the experiment itself).

More on these experimental limitations of the electron tomography reconstructions provided by SIRT will be covered later, in a comparison with the more modern total variation minimization techniques.

3.2.2. Compressed sensing (CS) and total variation minimization (TV_{min}).

In recent times, the mathematical field of *Compressed Sensing (CS)* [22] has emerged as a powerful tool for image processing, acquisition and compression,

as it states that a sparse signal can be recovered from a limited number of measurements below the classic Nyquist-Shannon limit criterium[23]. The image compression algorithms (JPEG and JPEG-2000)[24] are among the most prominent examples of the techniques arising from this field of study. Also, this initial description already hints some potential advantages of using this mathematical foundations in the field of electron tomography reconstructions [25,26]. Specifically, it could provide a viable solution when a lower number of projections (measurements) are considered, as long as the object itself can be expressed in a sparse manner in some basis other than the direct imaging space.

A complete derivation of the principles behind the CS theoretical framework requires a fair knowledge of the concepts of sparse representation, compressible signal and sensing processes (measurement). However, this work will be solely focused on highlighting the main core concepts and how those apply to the problem at issue: enhanced electron tomography reconstructions.

In a standard signal acquisition process (e.g., a conventional image acquisition in TEM, or a series of projection images in an electron tomography experiment), an initial signal $\vec{\boldsymbol{y}}$ with n components will be measured against the so-called *sensing waveform* (i.e., a functional basis Φ) giving a recorded signal $\vec{\boldsymbol{p}}$ with *m* components. This is:

$$\vec{\boldsymbol{p}} = \Phi\, \vec{\boldsymbol{y}} \tag{10}$$

In general, it is fair to assume that the process may suffer from undersampling (i.e., $m \ll n$) and, thus, the equation system is undetermined. For example, the exact denoised TEM image may present 'burnt' areas due to defects in the detectors, or the discrete number of projections acquired in a tomography experiment fails to sample the entirety of the projection Radon transform space. CS theory shows that even in those cases, a unique optimum solution can be calculated for this problem. Two major restrictions are put in place:

i. $\vec{\boldsymbol{y}}$ is required to be sparse in a certain basis Ψ.

ii. The basis for the sparse representation Ψ and for the sensing waveform (direct space) Φ must be incoherent.

Signal $\vec{\boldsymbol{y}}$ is considered *sparse* in a certain domain whenever all the information can be entirely expressed through a small subset of s coefficients $\vec{\boldsymbol{c}} = \{c_1, \dots, c_s \mid c_k \neq 0\}$. For example, consider an image with stark contrast differences in localised areas. Such image could be easily represented solely by the boundaries between such differentiated areas (i.e., the values for the contrast gradients). Mathematically, the sparse transform is expressed as:

$$\vec{\boldsymbol{c}} = \Psi\, \vec{\boldsymbol{y}} \tag{11}$$

and the signal is said to be sparse only if $s \ll n$, being then $\vec{\boldsymbol{c}}$ the actual sparse representation of $\vec{\boldsymbol{y}}$. This is a fairly rigid mathematical constraint, so in practical applications CS allows the relaxation of *sparsity* to *compressibility*. In this relaxed scenario, the transform Ψ is still forced to retrieve a certain small number of significant coefficients k > s. However, the surplus amount of $k - s$ coefficients different than 0 poses a lower significance for the information recovery than the s-number of $\{c_s\}$ coefficients, and $\vec{\boldsymbol{y}}$ is still sufficiently represented by them (i.e., s < k << n). Therefore, these extra $k - s$ small coefficients can be filtered away (or set to 0) with minimal information loss. In that case, $\vec{\boldsymbol{y}}$ is said to be a *compressible signal.*

The incoherence requirement for basis Ψ and Φ means that none of them (their base vectors) is allowed to be sparsely represented. This ensures mathematically that each p_i (sparse space) contains information about many of the coefficients y_j (direct space) and, thus, it guarantees that undersampling artefacts in the direct (Φ) domain are distributed as noise throughout the signal in the sparse domain (Ψ).

Going back to the specific problem at hand, conventionally, an electron tomography reconstruction struggles with the problem of fully recovering a 3D ($\vec{\boldsymbol{x}}$) object from a limited amount of information in the form of a set of 2D image projections ($\vec{\boldsymbol{b}}$) in the direct domain (sensing basis) measured (Φ), in which the projection process is encoded in the matrix form $\boldsymbol{A}$ (i.e., the problem

resolved by the conventional electron tomography algorithms described by eq.(6)). From this concept, it follows naturally that the lower the number of projections, the lower the measured information. Consequently, as the number of projections is reduced in the minimization problems proposed by different algorithms (e.g., SIRT in eq.(9)), the likelihood of achieving a suboptimum 'solution' far from an actual mathematically optimal one ($\vec{\boldsymbol{x}}^*$) is increased. It may be even the case that no optimum solution is available (i.e., divergent behaviour of the algorithm), and the retrieved solution is directly extracted after reaching a maximum number of iterations (a 'safety-mode' stopping criteria included in most iterative algorithms, as SIRT, to avoid getting trapped in infinite calculation loops). Hence, the artefactual problems (as those explained at the end of the SIRT algorithm description) would emerge from the insufficient sensing of the projection space in the language introduced for the CS theory framework.

Fortunately, this same CS theory seems to provide a way out, whenever one can find a sparse representation for the object under study. The most successful electron tomography reconstruction algorithms based on CS theory propose the use of the so-called total variation (TV) quantity in the minimization problems (the sparse domain Ψ becomes the space of the total variation, or gradient space). This quantity can be generally formulated as follows

$$TV(\boldsymbol{y}) = \sum_i \|D_i \boldsymbol{y}\|_p \tag{12}$$

where $\boldsymbol{y} \in \mathbb{R}^n$ (could be a 2D image such that $n = h \cdot w$, or a 3D volume such that $n = h \cdot w \cdot d$). The total variation quantity is deemed anisotropic-TV whenever one considers $p = 1$ (l_1-norm), and isotropic-TV if $p = 2$ (l_2-norm). From now on, $\|\vec{\boldsymbol{a}}\|$ will refer exclusively to $\|\vec{\boldsymbol{a}}\|_2$, as the solver implemented for electron tomography reconstruction is done for the isotropic-TV.

The problem for a minimization method based on the TV is that its inherent non-linearity and non-differentiability properties complicate the computation.

One of the iterative solvers proposed to approach this CS TV-minimization problem is the so-called ***TV*** *minimization by* ***a****ugmented* ***L****agrangian and* ***al****ternating direction* ***al****gorithms*, ***TVAL3*** for short[9,27].

To understand how this solver operates, let us consider again the same basic problem described by eq.(6), in which the electron tomography reconstruction tries to retrieve $\vec{\boldsymbol{x}}$ from $\vec{\boldsymbol{b}}$ with an a priori unknown $\boldsymbol{A}$. Now, the quantity to be minimized would be the isotropic-TV (i.e., on the sparse representation space)

$$\min_{\mathbf{x}} TV(\boldsymbol{x}) \triangleq \sum_i \|D_i \boldsymbol{x}\|, \qquad s.t.\ A\boldsymbol{x} = \boldsymbol{b} \tag{13}$$

TVAL3 pulls a couple of smart decisions to tackle the inherent issues encountered when resolving these types of minimization problems through iterative updating schemes. For example, it performs a variable splitting to ensure an easy differentiation of one part of the problem

$$\min_{\mathbf{x}} TV(\boldsymbol{x}) \triangleq \sum_i \|\boldsymbol{y}_i\|, \qquad s.t.\ A\boldsymbol{x} = \boldsymbol{b}\ and\ \boldsymbol{y}_i = D_i\boldsymbol{x} \tag{14}$$

where a new variable $\boldsymbol{y}_i = D_i\boldsymbol{x}$ is introduced. Or, it uses the theory of the augmented Lagrangian method[28] to mathematically ensure the existence of an optimal solution (i.e., local convergence to a solution $\boldsymbol{x}^*$). Then, instead of directly minimizing the total variation quantity, the algorithm is based upon the minimization of the augmented Lagrangian function of the problem

$$\mathcal{L}_{\boldsymbol{A}}(\boldsymbol{y}_i, \boldsymbol{x}, v_i, \lambda; \beta_i, \mu) =$$
$$\sum_i \left(\|\boldsymbol{y}_i\| - v_i^T (D_i\boldsymbol{x} - \boldsymbol{y}_i) + \frac{\beta_i}{2} \|D_i\boldsymbol{x} - \boldsymbol{y}_i\|^2 \right) - \lambda(A\boldsymbol{x} - \boldsymbol{b}) + \frac{\mu^2}{2} \|A\boldsymbol{x} - \boldsymbol{b}\|^2, \tag{15}$$

where v_i, λ are the Lagrangian coefficients for the variables $\boldsymbol{y}_i, \boldsymbol{x}$ and β_i, μ are the penalty parameters for the augmented Lagrangian minimization. The problem becomes then the following for each iterative step in the algorithm

$$\min_{\boldsymbol{y}_i,\boldsymbol{x}} \boldsymbol{\mathcal{L}_A}(\boldsymbol{y}_i,\boldsymbol{x}) = \min_{\boldsymbol{y}_i,\boldsymbol{x}} \left(\begin{array}{c} \sum_i \left(\|\boldsymbol{y}_i\| - v_i^T(D_i\boldsymbol{x} - \boldsymbol{y}_i) + \frac{\beta_i}{2}\|D_i\boldsymbol{x} - \boldsymbol{y}_i\|^2 \right) - \\ \lambda(A\boldsymbol{x} - \boldsymbol{b}) + \frac{\mu^2}{2}\|A\boldsymbol{x} - \boldsymbol{b}\|^2 \end{array} \right), \quad (16)$$

The actual iterative update scheme is well described in both the original publication accompanying TVAL3[9] and the PhD text where the mathematical principles are detailed and the convergence of the algorithm is demonstrated[27].

In the current implementation of the algorithm, both penalty parameters are kept constant, and given by the user as hyperparameters when initiating the reconstruction. These hyperparameters have an important weight on the aspect of final reconstruction achieved, and only in recent times a methodology to estimate the optimal values has been explored specifically for the field of electron tomography[11]. Regarding the scope of the work done in this PhD thesis, only the penalty parameter for the reconstruction error (μ) has ever been modified, leaving a fixed value for $\beta = 2^7$ (as instructed by default).

The effects of μ and the potential advantages of modifying it for specific reconstruction cases will be further explored bellow, by studying a heavily under-sampled phantom volume. Nevertheless, for the sake of completion, notice that μ weights the effects of the reconstruction error $A\boldsymbol{x} - \boldsymbol{b}$. The higher the value of μ, the higher the penalty over deviations from the optimal solution and, thus, the higher the level of detail that can be recovered from the reconstruction. However, increasing the value of μ may impede the convergence to such optimal solution in many cases, as the problem becomes an analogue to the actual solution for the SIRT algorithm and, thus, the information limit is raised as the sparsity is demoted. On the other hand, low values of μ tend to provide a smoothed reconstruction, much more resilient in heavy undersampling cases.

The original code for the TVAL3 solution is freely distributed to be used in MATLAB. However, one of the highlights marked in the roadmap for the work done in this PhD was the translation of the algorithm to the Python programming language, which ensures the possibility of a free distribution of this methodology and its incorporation into the standard set of analysis tools for TEM in our research group (LENS) which, for the most part, are programmed in Python (MATLAB requires a proprietary licence to be used). Moreover, the original algorithm was restricted to a 2D total variation computation, defining the base vectors of the illumination plane as the principal directions for the derivatives (i.e., perpendicular to the direction of illumination). As a consequence, the first results that can be found in the literature incorporating TVAL3 for electron tomography experiments showed a sort of stacking artifact (similar to a mismatch expected between slices of a 3D volume reconstructed stacked together slightly out of alignment, in the directions perpendicular to the illumination axis)[10]. To address this issue, the Python translated version includes an extra step in the algorithm to consider the TV minimization in all 3 spatial dimensions. This is not the first time that something like this has been effectively implemented [12], but to our knowledge it is the first time that it has been implemented in Python. The algorithm is later operated by the framework provided by the ASTRA-toolbox[18], through the spot-operators functionalities[29].

3.3. A qualitative comparison of SIRT and TVM.

Before the introduction of some of the experimental electron tomography results produced in the context of the present thesis, it is important to understand the potential improvements in the quality of the reconstructions that the inclusion of this algorithm (TVAL3, based on the minimization of the total variation) brought to our group capabilities (from now on, the CS tomography reconstruction methodology based upon TVAL3 is referred as TVM).

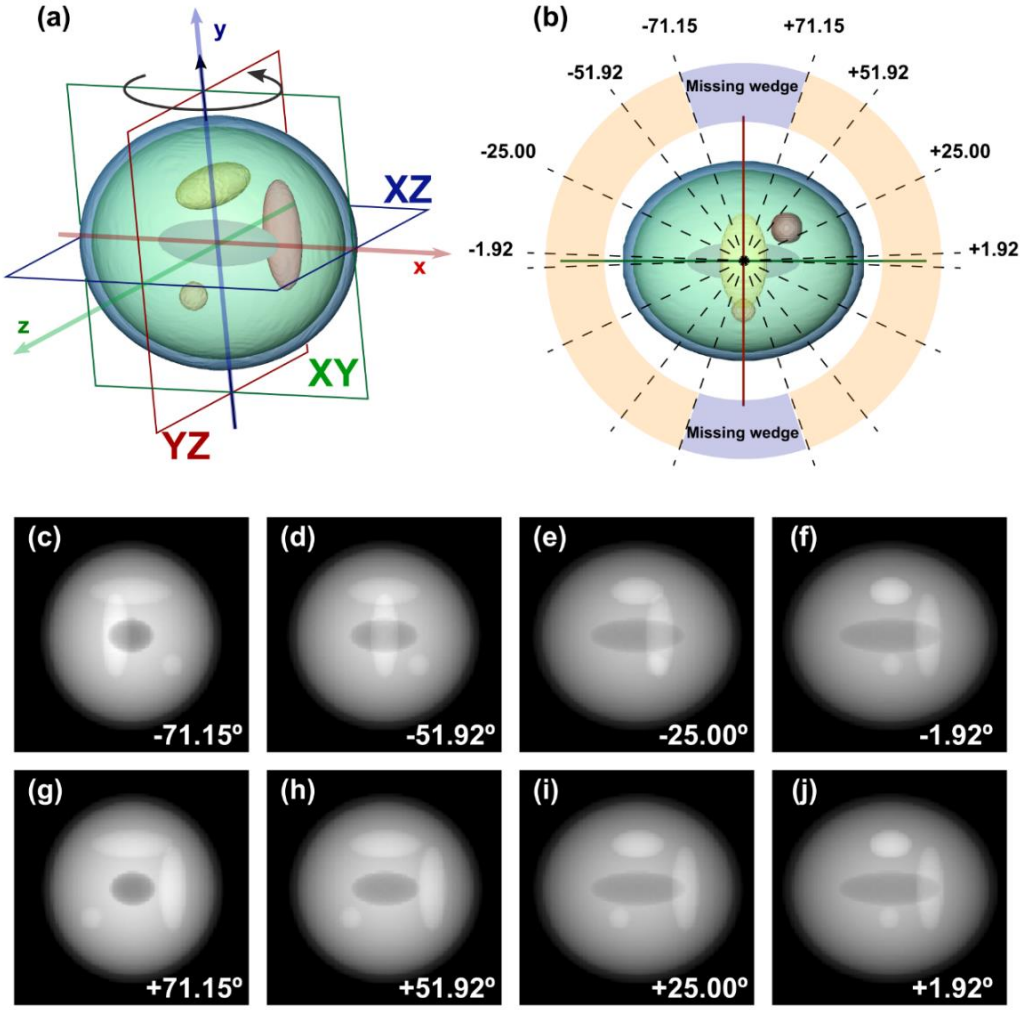


Figure 3.4. (a) Render of the phantom volume created for the electron tomography experiments to compare SIRT and TVM. The colours are given at random, and the central object is actually a void in the structure. (b) Top-view (along the rotation axis for the particle) of the volume rendering in (a). A set of viewing planes (planes of the projections) are indicated for a group of 8 angles. (c-j) Partial subset of projections used in some of the reconstruction experiments for the planes marked in (b).

To do so, a series of controlled experiments of electron tomography were carried out using a phantom volume created from scratch, to be able to compare the results of TVM and SIRT under several synthetic different experimental conditions. The phantom volume, as well as a subset of projections mimicking the potential acquired projection images in a TEM experiment are shown in **Figure 3.4**. This phantom object is devised to

represent as core-shell nanoparticle, with a basic ellipsoid shape (see the isometric perspective of the volume rendering in **Figure 3.4**(a)). Two elongated smaller nanoparticles are embedded within the core in the **z** (yellow NP) and **y** (red NP) directions. A smaller and spherical NP (orange) is also embedded within the core. Finally, a central void elongated in the **x** direction is carved from the core volume (see **Figure 3.4**(a)). This last feature, the central void, will play a key role in the comparisons between algorithms and qualitative evaluation of performance, as it turns out that precisely those type of 'negative density features' are among the most challenging structures to be reconstructed by the iterative algorithms. It is also a commonly observed feature in studies of materials for catalysis. In fact, the necessity of a quantitative analysis solution for the study of the 3D pore size distribution in materials with catalysis applications, in conjunction with the insufficient quality of the conventional tomography reconstructions methods, have been a major drive towards the inclusion and development of advanced algorithms in this field of TEM electron tomography analysis[11,30,31].

The colours given to the volume rendering in **Figure 3.4** (a) are arbitrary, and are only valid as a visual guide for the sub-volumes included in the phantom object, although they will be maintained throughout the rest of the figures in this section for consistency reasons. The rotation axis for this synthetic nanoparticle is also indicated in the image, along the denominated **y** axis. A small subset of hypothetical projection planes is marked in **Figure 3.4** (b) overlaid with the top view of the volume rendering from **Figure 3.4** (a) (i.e., visualizing the particle in the **-y** direction). Once again, we are making use of the equivalence relation between a rotation of the particle and the rotation of the viewing and projection perspectives. In **Figure 3.4**(b), the projection planes are marked as in the case of a perspective rotation for the projection images acquired in a hypothetical experimental tomography data acquisition.

As in many other electron tomography experiments, the angular range contains both positive and negative entries. The absolute reference plane at

$\phi = 0^o$ (arbitrarily selected in this case) would correspond to the XY orthoslice plane marked in **Figure 3.4** (a).

Figure 3.4 (c-j) present some of the actual projections of the volume for the planes indicated in **Figure 3.4** (b). These projections were calculated using the functions included in the ASTRA toolbox to project the volume in a set of specified angles each time.

The phantom volume itself is a cubic array with a size of (128x128x128) voxels. Each voxel from the volume is initially configured with a fixed intensity value ranging from 0 (background and central void) to 4 (the red ellipsoidal NP oriented in the **y** direction). A value of 3 is given to the elongated NP oriented in the **z** direction (yellow ellipsoidal NP) and 3.5 to the smaller sphere embedded in the core region (orange coloured spherical NP). The shell of the principal NP is set with an intensity value of 1 (the blue shell in **Figure 3.4** (a)), and the main core intensity value is set as 2. After this initial setting, the whole phantom is rescaled to a range of values between 0 and 255, and those are the actual intensities projected by ASTRA to generate the artificial set of projection images for the reconstructions (examples given in **Figure 3.4** (b)).

Having described the phantom volume, in the following sections a synthetic comparative of experiments is carried out to examine the effects of undersampling in both the angular range and projection space and to evaluate the resilience of the TVM method to noise.

3.3.1.SIRT vs TVM. Undersampling the angular range.

The quality of the reconstructions obtained from conventional iterative algebraic reconstruction methodologies (i.e., from the SIRT algorithm) is tightly bounded to the number of projections used in the reconstruction process. As discussed in the theoretical introduction of this chapter, below a certain information limit the algorithm proposed by SIRT would not be able to converge. Furthermore, the effects of the so-called missing wedge (an undersampling of the direct space of projections in a very specific angular range) are usually unavoidable in the reconstructions made by SIRT. As a

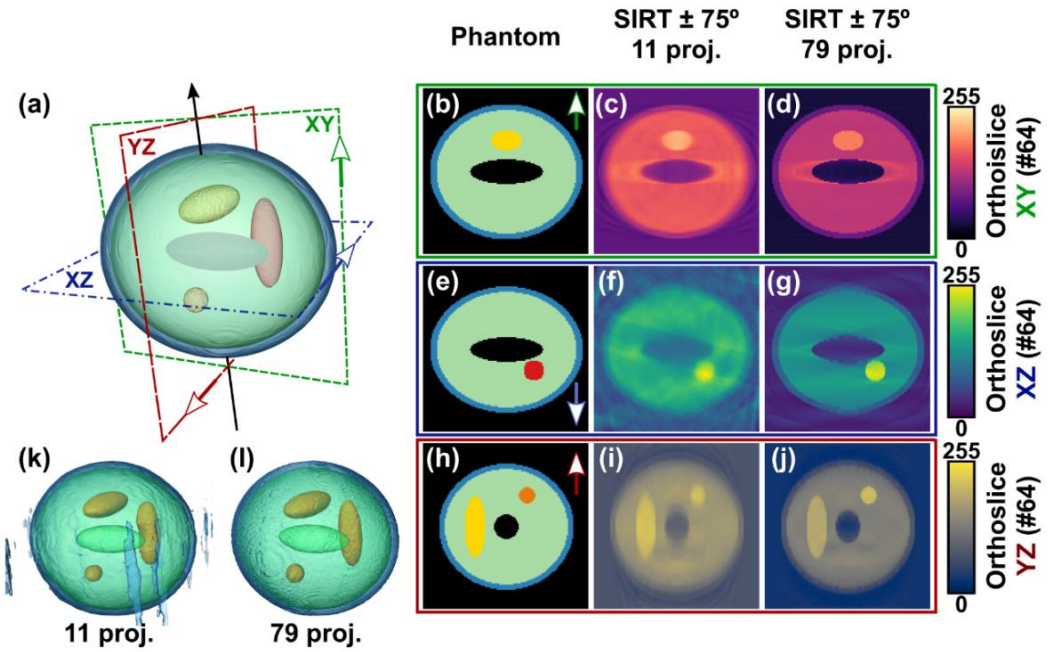


Figure 3.5. (a) Perspective of a volume rendering for the phantom from **Figure 3.4**. (b-d) Orthoslices for the XY plane through the original phantom, 11 projections and 79 projections SIRT reconstruction volumes respectively. (e-g) Orthoslices for the XZ plane through the original phantom, 11 projections and 79 projections SIRT reconstruction volumes respectively. (h-j) Orthoslices for the XZ plane through the original phantom, 11 projections and 79 projections SIRT reconstruction volumes respectively. (k-l) Volume rendering for the 11projections and 79 projections reconstructions respectively. These representations are an ensemble of two different objects: A volume rendering for the core-shell and central void of the phantom NP, and isosurfaces renders for the 3 embedded NPs.

result, the reconstructed volumes tend to present a deformation aligned with the bisectrix axis defined by the missing wedge. An example of these types of results from SIRT reconstructions of the phantom volume is shown in **Figure 3.5**.

The rotation axis for the phantom particle is the y axis as indicated in the perspective view of the volume rendering in **Figure 3.5** (a). The hypothetical projection 0º of the rotation coincides with the XY plane highlighted in green, whereas the missing wedge would be bisected by the YZ plane highlighted in red. The coloured arrows in the perspective view are also shown over the slicing planes, as a visual aid for the recognition of directions. The

reconstruction results presented in this figure correspond to two different experiments with a different set of angles. The first one corresponds to an experiment set with a number of 11 equidistant projections in the angular range ±75º (i.e., an angular separation of approximately 13.64º). The second one was set with a number of 79 projections for the exact same angular range (i.e., 1.9º of separation between projections). The resulting render for the 11 projections case is showcased in **Figure 3.5** (k), and the 3 orthoslices (cuts of the volume in the x, y and z directions) corresponding to the planes XY, XZ and YZ are showcased in **Figure 3.5** (c), (f) and (i) respectively. For the 79 projections case, the render is shown in **Figure 3.5** (l), and the orthoslices for the planes XY, XZ and YZ are shown in **Figure 3.5** (d), (g) and (j) respectively. All these orthoslices correspond to the central slices through the 128x128x128 reconstructed volume.

In both experiments an elongation of the phantom structure can be observed in the perpendicular direction to the XY plane (i.e., the principal direction of the missing wedge). The deformation of the volume reconstructed is directly observable in the XZ and YZ orthoslices, specially for the void structure in the middle. Also, the so-called streaking artefacts are present in both cases, although much more severely in the 11 projections case (up to the point of even appearing as structural elongated volumes in the margins of the rendering at **Figure 3.5** (k)).

Before moving into the TVM results, a word of caution must be issued for the volume render visualizations. Presenting the reconstruction information though these types of visualization methods (volume renderings and isosurfaces) can be often misleading. The end user has always the possibility of playing both with the perspective and the voxel intensity levels to present the information in the most convenient way (i.e., confirmation bias might be an issue regarding the analysis of the reconstructions through the renders). For that reason, in this work the renders are always accompanied by orthoslices. Moreover, in the particular case of the renders in **Figure 3.5** (k) and (l), the intensity ranges used for the visualization were identical, to allow

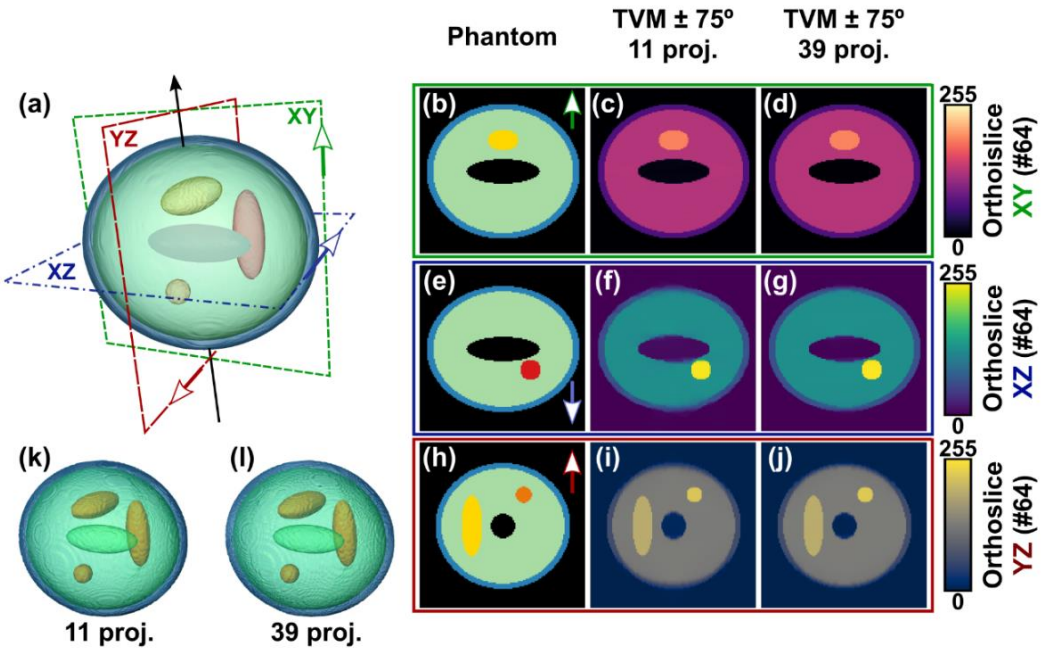


Figure 3.6. (a) Perspective of a volume rendering for the phantom from **Figure 3.4**. (b-d) Orthoslices for the XY plane through the original phantom, 11 projections and 79 projections TVM reconstruction volumes respectively. (e-g) Orthoslices for the XZ plane through the original phantom, 11 projections and 79 projections TVM reconstruction volumes respectively. (h-j) Orthoslices for the XZ plane through the original phantom, 11 projections and 79 projections TVM reconstruction volumes respectively. (k-l) Volume rendering for the 11 projections and 79 projections reconstructions respectively. These representations are an ensemble of two different objects: A volume rendering for the core-shell and central void of the phantom NP, and isosurfaces renders for the 3 embedded NPs.

for a direct qualitative comparison of the reconstruction results and overall quality. This criteria of using the same information limits in all the renders shown in the same figure will be maintained throughout the whole text of this PhD if the contrary is not explicitly indicated.

Almost the same 'experimental' conditions were used to reconstruct the volumes with the TVM methodology. Although the angular range for the projections is kept as ±75º, in this case the limiting number of projections used for the reconstruction was set as 39 (instead of the 79 number of projections for the SIRT case). As the projections were completely noise-free, the penalty

parameter for the reconstruction error was set with a value of $\mu = 2^7$ (considered an intermediate-large value for the TVAL3 algorithm).

The resulting reconstructions are shown in **Figure 3.6**, with an identical layout to **Figure 3.5**. The TVM method retrieves an accurate reconstruction a number of projections as for low as 11. Also, in both cases of 11 and 39 projections the results are very close, and of a higher quality than the case of 79 projection for the SIRT algorithm. No major negative effects from the missing wedge can be now seen in the reconstructions, with the exception of a slight anisotropic thinning of the shell in the direction defined by the intersection of the planes XZ (blue) and YZ (red). This thinning effect is easily observable precisely in the orthoslices XZ (**Figure 3.6** (e-g)) and YZ (**Figure 3.6** (h-j)). Also, a small smearing effect over the intensity in the same general direction as the thinning one is observable. It is slightly more severe for the 11 projections case than for 39 projections one from a qualitative comparison between the images of **Figure 3.6** (f-g) and of **Figure 3.6** (i-j).

The streaking artefacts are not visible in any of the orthoslices analysed for the TVM reconstructions, and the inner void structure is reproduced by the reconstructions with an apparent high fidelity (i.e., closer to the shape in the synthetic reference phantom, and with an average intensity value very close to the background).

In order to close this comparative experiment, a series of profile lines taken from orthoslices through the reconstructed volumes are analysed for each one of the algorithms and experimental conditions described in **Figure 3.5** and **Figure 3.6**. The analysis is subdivided according to the slicing plane considered.

Beginning with the slicing direction that was overall less affected by the missing wedge, the results for the profiles in the XY plane are shown in **Figure 3.7**. Both profiling directions (horizontal **Figure 3.7**(f) and vertical **Figure 3.7**(g)) for the TVM reconstructions (**Figure 3.7**(b-c)) show an almost perfect fit with the original intensity signal of the phantom (solid line). The only

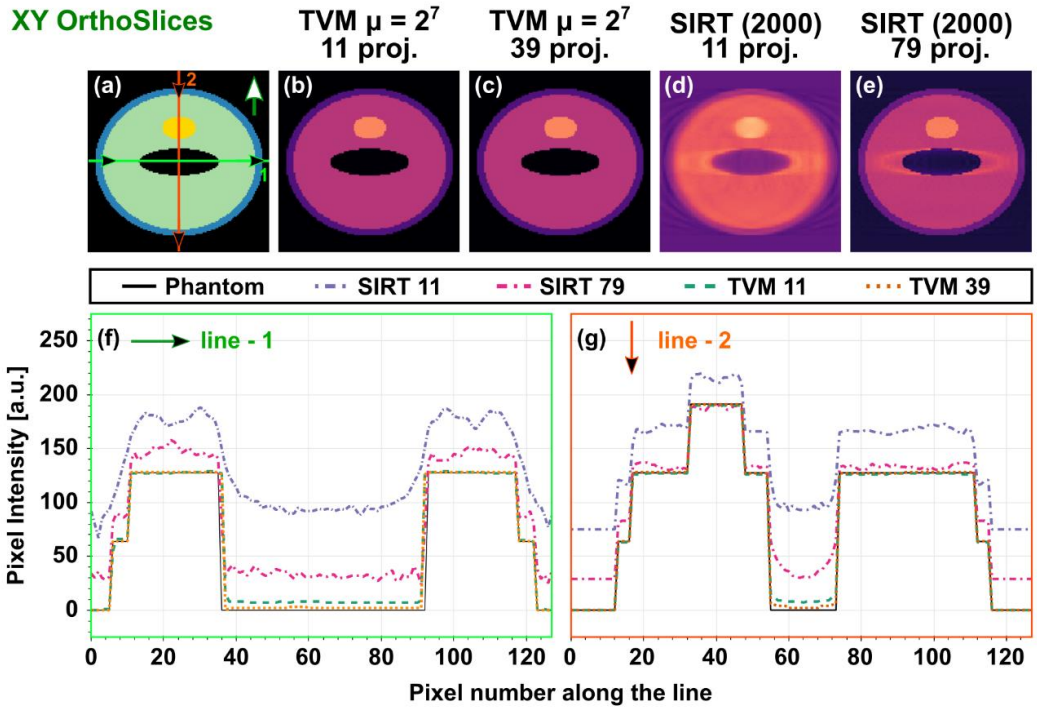


Figure 3.7. XY orthoslices for the reference phantom volume (a), for the TVM reconstructions using 11 projections (b) and 39 projections (c), and for the SIRT reconstructions using 11 (d) and 79 (e) projections. (f) Intensity profiles for the horizontal path highlighted in green (1) in (a). Intensity profiles for the vertical path highlighted in orange (2) in (a).

difference in both cases arise in the slight overestimation of the intensity in the reconstructions for the central void region (i.e., the central part in both profiles), where the 39 projections case is closer to the phantom than the 11 projections one. Both the 11 and 79 projections cases of the SIRT reconstructions (**Figure 3.7**(d-e)) show a clear separation form the phantom intensities, regardless of the profiling direction. As expected, the 79 projections case is closer to the original intensity profiles than the 11 projections case. It is also observable a higher separation for the horizontal profiling direction, caused by the anisotropic smearing artefact surrounding the central void and clearly oriented in the same direction.

The profiles done over the XZ slices are shown in **Figure 3.8**. Once again, the profiles for the TVM reconstructions are almost identical to the phantom profiles. The slight overestimation of the intensity in the void region is still

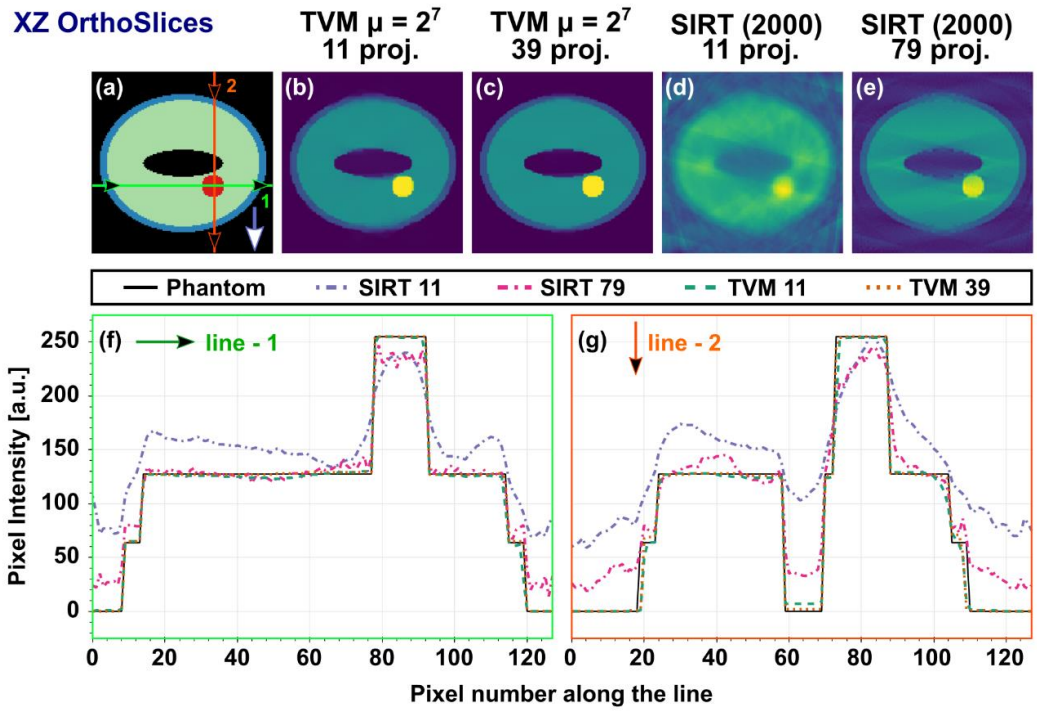


Figure 3.8. XZ orthoslices for the reference phantom volume (a), for the TVM reconstructions using 11 projections (b) and 39 projections (c), and for the SIRT reconstructions using 11 (d) and 79 (e) projections. (f) Intensity profiles for the horizontal path highlighted in green (1) in (a). Intensity profiles for the vertical path highlighted in orange (2) in (a).

happening here (vertical profile, **Figure 3.8** (g)), and still more prominent for the case with a lower number of initial projections. However, a quick comparison between the profiles in the horizontal (**Figure 3.8** (f)) and vertical (**Figure 3.8** (g)) directions for the TVM signals shows the marginal but still existing influence of the missing wedge on the reconstructions. Notice how, in the vertical profiling direction, a slight shaving of the squared corners defined by the intensity steps is observable. This effect is more accentuated for the 11 projections case, equivalently to the higher smoothing effect over the boundaries described for the TVM slices in **Figure 3.6**.

The profiles for the SIRT reconstructions tell a completely different story. The horizonal profiling **Figure 3.8** (f) for the 79 projections case (**Figure 3.8** (e)) shows a fairly good agreement with the phantom profile. However, the vertical profile **Figure 3.8** (f) accuses the streaking artefacts in the reconstruction

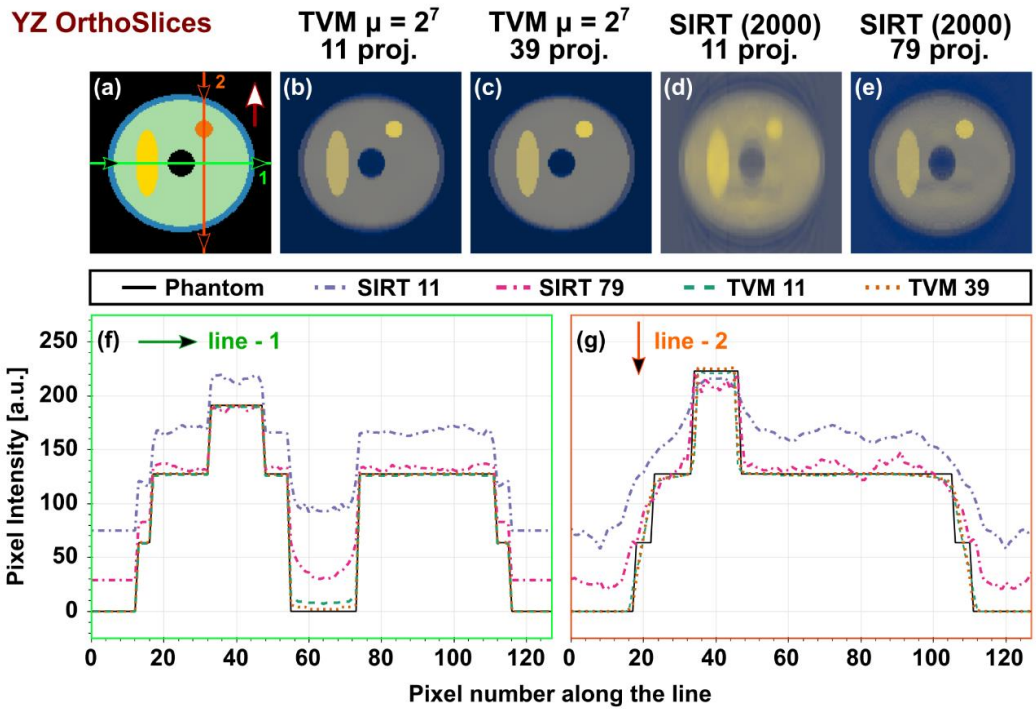


Figure 3.9. YZ orthoslices for the reference phantom volume (a), for the TVM reconstructions using 11 projections (b) and 39 projections (c), and for the SIRT reconstructions using 11 (d) and 79 (e) projections. (f) Intensity profiles for the horizontal path highlighted in green (1) in (a). Intensity profiles for the vertical path highlighted in orange (2) in (a).

caused by the lack of information in the missing wedge. The profiles for the 11 projections case (**Figure 3.8** (d)) are far from the phantom profile in both the horizontal (**Figure 3.8** (f)) and perpendicular (**Figure 3.8** (g)) directions. In the latter, the reconstruction is so corrupted by artefacts (streaking lines) that the intensity traversing the void region almost reaches the same level of the previous plateau.

The same overall behaviour can be observed for the line profiles taken over the YZ slices showcased in **Figure 3.9**, with a slightly better behaviour in the 79 projections case of the SIRT reconstruction (**Figure 3.9** (e)) with respect to the original profiles of the phantom image (**Figure 3.9** (a)).

One thing that only becomes apparent when one compares all 3 principal slicing directions (**Figure 3.7** XY , **Figure 3.8** XZ and **Figure 3.9** YZ) is that the TVM does a much better job maintaining the overall contrast structures than

the SIRT algorithm, especially for the severe undersampling case of 11 projections. In that sense, the XY profiles for the 11 projection SIRT reconstruction (**Figure 3.7** (d)) show a clear overestimation of the intensity crossing the inner particle region in both directions with regard to the phantom reference intensity profile. However, in the XZ profiles in both directions (**Figure 3.8** (f)(g)) the 11 projections SIRT reconstruction signals (**Figure 3.8** (d)) underestimate the maximum intensity traversing the NP with respect to the one measured for the reference phantom (**Figure 3.8** (a)). Finally, in the YZ case the horizontal profiling line (**Figure 3.9** (f)) and the vertical one (**Figure 3.9** (g)) cross two different NPs. Notice how the horizontal case overestimates the intensity for the 11 projections SIRT reconstruction (**Figure 3.9** (d)), whereas the vertical line case underestimates it. In both profiles, the maximum intensity values are close. In a sense, these results indicate that in cases of severe undersampling the SIRT algorithm will automatically tend to flatten the overall signal, and will saturate the maximum intensity values in the localized regions with higher contrast values (in this case, all the embedded NPs).

To summarize the results of this subsection in a brief statement, the TVM outperforms the SIRT algorithm in all cases by a long margin, and even in the extreme case of including such a small number of projections as 11 is capable of retrieving a very accurate result in good agreement with the original phantom.

3.3.2. Noise resilience of the TVM method.

From this point onward in this PhD work, only the TVM methodology will be employed for electron tomography reconstructions (morphological and analytical electron tomography reconstructions). In the previous section, the reconstructions were carried out utilizing exclusively pristine projections (i.e., without noise). However, in real experiments one has to account for the effects of noise pollution in every single projection acquired. To test the resilience of the modified TVAL3 algorithm to experimental noisy conditions, 2 different

levels of a mixture of Gaussian and Poissonian noise were added to the phantom volume before carrying out the reconstructions. These noise levels are characterized by an average signal-to-noise ratio (SNR) estimation calculated from

$$SNR = 10\log_{10}\left(\frac{\sum_i (x_i)^2}{\sum_k (x_k - x_k^*)^2}\right) \tag{17}$$

where $\vec{\boldsymbol{x}} = \{x_i\}_{i=1,\dots,n}$ would be the intensities of the pristine phantom volume and $\vec{\boldsymbol{x}}^* = \{x_i^*\}_{i=1,\dots,n}$ would be the intensities of the noise polluted phantom volume. The higher the SNR value, the lower the quantity of noise added to the projections.

In order to be able to compare them, the same set of hypothetical experimental conditions (31 equidistant projections with an angular step of 5º in the ±75º angular range) were used in every reconstruction. For the lower noise level (NLvl.1, SNR 23.67dB) the TVM was set with a penalty parameter $\mu = 2^7$. The higher noise level (NLvl.2, SNR 9.68dB) was reconstructed with $\mu = 2^7$ and $\mu = 2^4$. The results of the reconstructions are shown in **Figure 3.10**. To visualize the two different levels of noise, **Figure 3.10** (a) shows a comparative of the projections at an angle of 0º. The volume reconstructions for the 2 different noise levels are shown in **Figure 3.10** (b), alongside a render of the phantom volume. The volume renderings for the reconstructions are sliced by the XY plane, highlighted in green in the phantom volume. Cutting at half these renders provides a clear image of the embedded NPs, represented by isosurfaces (i.e., surfaces of at equal voxel intensity value).

Also notice that in the case of the higher noise level (Lvl.2, SNR of 9.68dB) two stripes appear in the background in both the $\mu = 2^7$ and $\mu = 2^4$ reconstructions. Although the voxel intensity of the 3D renders was capped with a lower boundary to avoid the representation of the background noise, these artefacts persist. They are the result of the effects of the finite size of the projections and, thus, the unavoidable mismatch between the geometries of the data cylinder defined by them (oriented along the rotation axis) and the

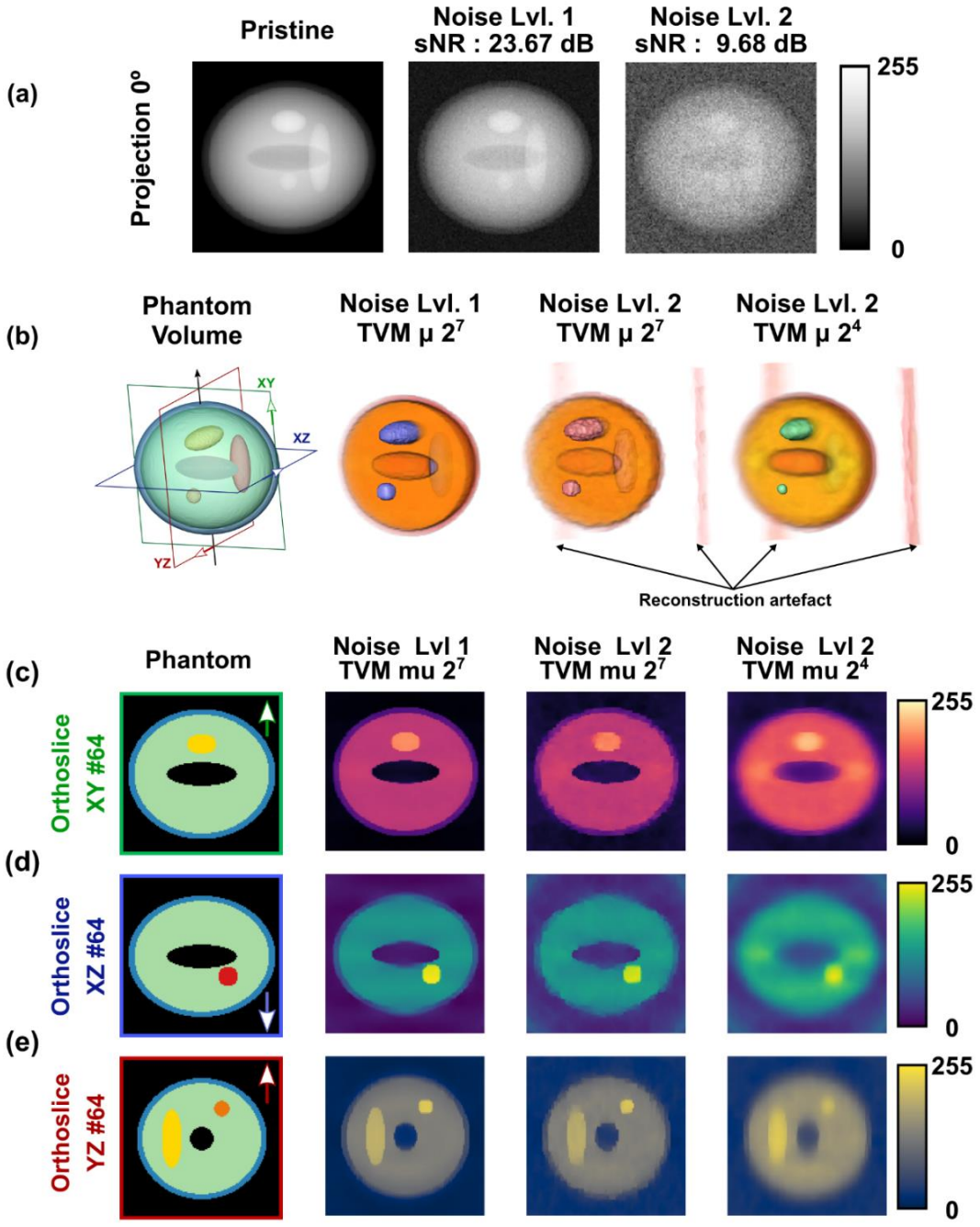


Figure 3.10. (a) Projections at the 0º angle for the phantom volume from **Figure 3.4** showing different levels of noise (pristine-original phantom, SNR 23.67dB and SNR 9.68dB). (b) Volume renderings and isosurfaces for the original phantom volume and the 3 reconstructions done with different levels of noise and penalty (μ) parameters. The orthoslices at the middle of the voxel cube reconstructed are shown for the XY plane in (c), for the XZ plane in (d) and the YZ plane in (e).

intended data cube volume computed by the algorithm. In noiseless conditions, the border pixels were given an intensity value of 0, and thus the gradient computed by the algorithm remained flat and a voxel intensity of 0 was retrieved in the corners of the cube volume reconstructed. When the background signal is far from 0 (noisy conditions), a cylindrical envelope is created by the algorithm surrounding the volume reconstruction at the centre of the cube, and in the direction of the rotation axis. In the case of the lower noise level (Lvl.1, SNR of 23.67dB and TVM reconstruction with $\mu = 2^7$), the lower boundary capping was sufficient to extract the artefactual intensity at the corners. These artefactual stripes can also be observed in the XZ slices through the volume for all the reconstructions (**Figure 3.10** (d), planes perpendicular to the rotation axis), although more prominently for the cases of noise Lvl.2. This type of enveloping artefact can be very often observed in experimental reconstructions. Precisely for this reason, it is always convenient to place the region of interest as centred as possible in the projections during the acquisition process. Otherwise, the risk of retrieving a deformed reconstruction towards the limits of the cube volume (cause by the gradient computations of the algorithm) is high, especially for noisy datasets.

The slices through the volume for the planes XY, XZ and YZ (i.e., perpendicular to the 3 directions of the frame of reference or cube-volume) are shown in **Figure 3.10** (c)(d) and (e) respectively. For the lower noise level (NLvl.1), the reconstruction with the penalty parameter $\mu = 2^7$ show an excellent agreement with the phantom slices, although a hint of the possible effects of the missing wedge is visible for the XZ and YZ planes. The ellipsoidal shape is slightly deformed at the top, and at the same time the shell region is thinned. The results for the higher noise level (NLvl.2) are not as good though. They are still in good agreement with the phantom volume, as all the regions (inner void, NPs and core-shell structure) are represented in the reconstructions. The higher penalty parameter case ($\mu = 2^7$) shows a higher granularity effect, whereas the lower penalty case ($\mu = 2^4$) provides a smoothed reconstruction (given the higher weight given to the gradient computations in the TVAL3

algorithm). However, the enforcement of a higher weight for the gradient computation in the TVAL3 algorithm ($\mu = 2^4$) flattens the overall voxel intensity distribution retrieved and causes a loss of definition in the image (i.e., losing contrast definition between differentiated structures).

Once again, the semiquantitative analysis through the intensity profiles extracted from lines drawn over the volume slices is carried out to confirm the results of the qualitative analysis shown in **Figure 3.10**, and seal the concepts explored in this section. These profiling results are shown in **Figure 3.11**. The curves for both noise levels reconstructions with a penalty parameter of $\mu = 2^7$ follow the same step pattern of the reference phantom curve for all the horizontal lines proposed. However, the vertical profiles affected by the missing wedge (XZ and YZ) show a separation from the reference curves, as the steps are blended in an almost continuum gradient curve. Also, as the noise level is incremented, the background signal intensity of the reconstructions is increased and, thus, the contrast resolution is lowered (see the orthoslices from **Figure 3.10**). This effect is translated into intensity profiles as an increment of the base level of intensity.

Regarding the profiles for the higher level of noise and the penalty parameter $\mu = 2^4$, they present a higher degree of separation from the reference phantom signals. At the same time, the contrast is reduced and the intensities for the NPs tend to saturate to the same values (as the cases of the SIRT reconstructions for the severe angular undersampling datasets, **Figure 3.7**, **Figure 3.8** and **Figure 3.9**).

In any case, from the combined results from **Figure 3.10** and **Figure 3.11** it is safe to assume that the TVM methodology is capable of retrieving high quality 3D volume reconstructions even in cases of severe noise pollution. The trade-off between smoothness and lateral resolution (through the penalty parameter μ) should always be considered when facing extreme cases of low SNR.

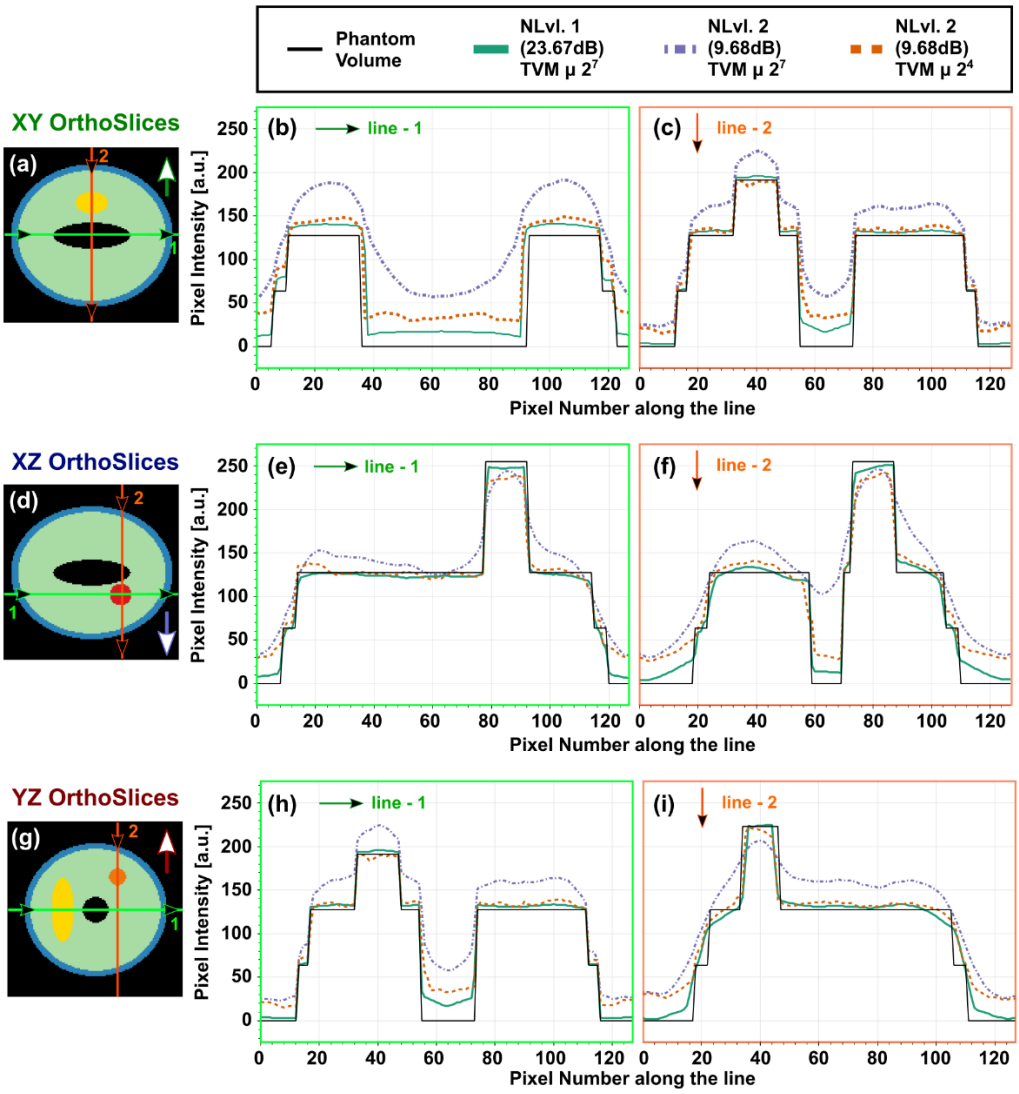


Figure 3.11. (a) XY slice through the phantom volume. The intensity profiles for the lines drawn in (a) and for all the XY slices shown in **Figure 3.10**(c) are shown in (b) for the horizontal and (c) for the vertical directions. (d) XY slice through the phantom volume. The intensity profiles for the lines drawn in (d) and for all the XZ slices shown in **Figure 3.10**(d) are shown in (e) for the horizontal and (g) for the vertical directions. (g) YZ slice through the phantom volume. The intensity profiles for the lines drawn in (g) and for all the YZ slices shown in **Figure 3.10**(g) are shown in (h) for the horizontal and (i) for the vertical directions.

3.3.3. Resilience of the TVM method to image erosion. Severe undersampling conditions in X-EDS analytical electron tomography.

The last experiment on the volume phantom presented in **Figure 3.4** is centred around the effects of image erosion on the reconstructions. The term 'image erosion' refers to a complete loss of signal (i.e., a pixel intensity equal to 0) in arbitrary regions of the projection images used for the 3D reconstructions. This is a common scenario in X-EDS analytical electron tomography experiments.

This erosion defect is particularly persistent in cases of low pixel/time registration or low dose setups for sensitive samples (i.e., samples easily degraded under the electron beam). The projection images extracted from the X-EDS count mappings are usually incomplete, and contain several areas without any X-ray count. Hence, these reconstructions would be faced with a dual undersampling problem (i.e., discrete number of projections and eroded images).

To test the capability of the TVM methodology to resolve these types of problems, the original phantom volume is divided in 3 subsets (or volumes): the shell, the embedded nanoparticles and the central core. The latter contains the central void and extra cavities in the positions formerly occupied by the NPs. These 3 volumes where projected in an angular range of ±75º with an angular step of 5º (31 projections). The resulting projections where polluted with noise (low level of noise, at a SNR of 20dB), and further modified by erosion at two different percentages (30% and 70%). This way, these synthetic experiments try to mimic the conditions of an analytical electron tomography reconstruction with incomplete signals in the X-EDS mappings. The results for the low level of erosion are shown in **Figure 3.12**. In this case, 30% of the initially non-zero valued pixels were set to 0 at random in each projection.

The level of detail in the eroded projections (**Figure 3.12** (a)) is still high enough to distinguish partially the structures within the phantom. The

reconstructions were done for two different penalty parameters (μ) to compare the effects of the smoothing factor of the TVAL3 algorithm. The volume representations for the reconstructions are shown in **Figure 3.12** (b).

In both the $\mu = 2^3$ and $\mu = 2^5$ cases, the core and shell reconstructions are presented as volume renderings with a slightly saturated colour-scale and sliced by YZ central plane, and the NPs reconstructions are presented by complete isosurfaces. The $\mu = 2^3$ case (blue-greens and pink NPs) presents an overall smoother finish, but the shell reconstruction shows a clear anisotropic behaviour in the main direction defined by the missing wedge, as it appears thinner (with a higher transparency level).Meanwhile, the $\mu = 2^5$ presents a reconstruction with a higher level of granularity, but the missing wedge effects are minimized.

The orthoslices corresponding to the XY, XZ and YZ central planes are shown in **Figure 3.12** (c), (d) and (e) respectively. A higher level of anisotropic deformation and distribution of intensities (associated to the missing wedge) is confirmed for the $\mu = 2^3$ reconstructions by analysing the XZ and YZ slices. In both cases ($\mu = 2^5$ and $\mu = 2^3$) the central void and cavities associated to the NPs are approximately resolved, although the central void in the 'core subset' presents a clear deformation with respect to the phantom slice (especially in the XZ and YZ slicing planes). The $\mu = 2^5$ presents an overall lower contrast level with respect to the $\mu = 2^3$ case, with lower values of the maximum intensity levels for the slices in all 3 directions. This effect appears to work in an opposite way to the contrast degradation of the previous sections, where a lower value of μ resulted in reconstructions with an overall lower dynamic range for the intensity values. The reason is the limitation imposed by μ on the gradient computations in combination with a large number of void regions in between non-zero datapoints in every projection. As the intensities of several non-boundary and non-background pixels were set to 0 (i.e., image erosion), the algorithm had to compute that missing information. One may think about the penalty parameter as an effective radius for the gradient computations. In previous sections the noise introduced was

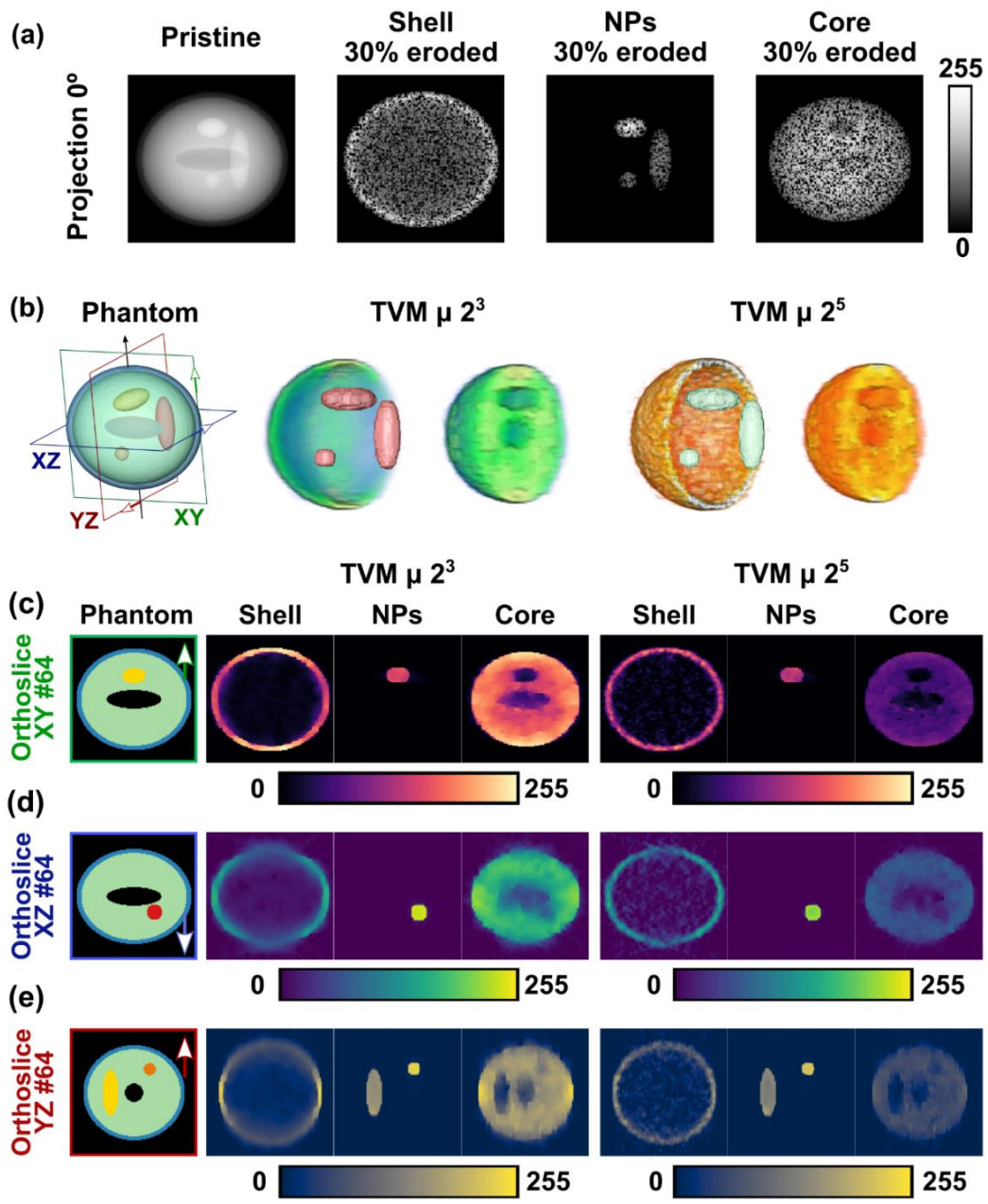


Figure 3.12. Panel for the partial reconstruction of volumes at a 30% level of erosion (a) Projections at an angle of 0º for the pristine phantom, the shell, the nanoparticles and the main core. (b) 3D representations of the phantom volume, and the reconstructions for $\mu=2^3$ and $\mu=2^5$. The orthoslices for the phantom volume, the 3 separated components of the $\mu=2^3$ reconstruction and the $\mu=2^5$ reconstruction are shown in (c) for the XY plane, (d) for the XZ plane and (e) for the YZ plane.

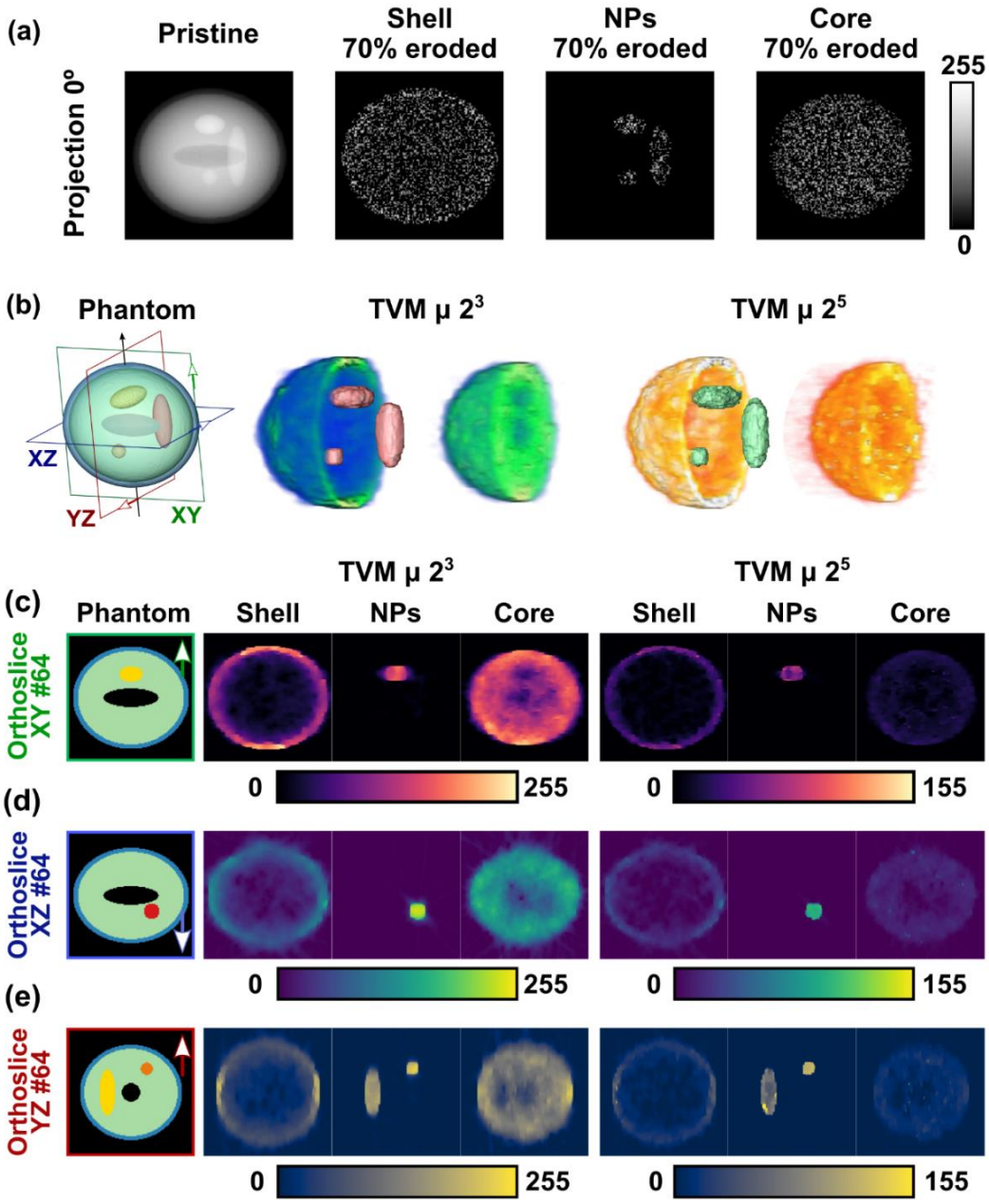


Figure 3.13. Panel for the partial reconstruction of volumes at a 70% level of erosion (a) Projections at an angle of 0º for the pristine phantom, the shell, the nanoparticles and the main core. (b) 3D representations of the phantom volume, and the reconstructions for $\mu=2^3$ and $\mu=2^5$. The orthoslices for the phantom volume, the 3 separated components of the $\mu=2^3$ reconstruction and the $\mu=2^5$ reconstruction are shown in (c) for the XY plane, (d) for the XZ plane and (e) for the YZ plane.

additive in nature, and the effects of setting a higher radius for the gradient computation was an excessive blurring effect over the details of the reconstruction.

To test the limits of the TVM methodology dealing with projection image erosion, a second set of reconstructions using the same experimental setup and penalty parameters were carried out, forcing a 70% average undersampling of the phantom dataset. The results are shown in **Figure 3.13**. Both $\mu = 2^3$ and $\mu = 2^5$ are capable of reproducing the major details on the phantom volume, such as the volumes associated to the NPs and a fairly well constructed shell (**Figure 3.13** (b)). However, the negative density spaces (i.e., the voids and NPs cavities in the core part) are retrieved with a lower fidelity level, especially in the case of $\mu = 2^5$. It is important to highlight that the colour-scale of the volume rendering for the core and shell reconstructions in the $\mu = 2^5$ case (**Figure 3.13** (b), on the right in red and orange) was artificially saturated to be able to distinguish the details retrieved

.In fact, notice how the colour-scale for the orthoslices is also limited to a maximum level of 155, and even so the contrast in the images is still poor in comparison to the $\mu = 2^3$ case.

In any case, even for such unfavourable conditions the TVM methodology seems to be able to achieve results of a high enough quality to serve, at least, as a qualitative method to analyse the distribution of structures in the volume. Also, as exemplified by this erosion experiment, it is always recommendable to pair the analytical X-EDS reconstructions with the morphological grey-scale ones from the co-acquired images (i.e., HAADF electron tomography reconstructions). This is of especial interest for samples with void spaces inside larger volumes, that can be completely

misrepresented by the analytical signal if the quality of the mapping images is low.

3.4. A case study.

To test the capabilities of the TVM methodology on a real case scenario, a series of electron tomography experiments were carried out for a set of spherical-shaped and hollow nanoparticles (i.e., NPs purely form as an spherical shell surrounded a voided core) with three different composition: (1) A sample of hollow iron oxide NPs (γ-F_2O_3) with a thin shell configuration, (2) a sample of hollow iron oxide NPs (γ-F_2O_3) with a thick shell configuration and (3) a sample of hollow nickel-iron oxide NPs which was supposed to present a dual-shell structure. The samples were synthesised by Dr. Davide Peddis and co-workers at the Institute of Structure of Matter of the National Research Council (CNR-ISM).

This section is subdivided in two blocks containing two sets of experiments using the TVM method for the electron tomography reconstructions as the backbone for the analysis. The first one is centred around a semi-quantitative morphological analysis of the thin-thick hollow nanoparticles. The objective of this experiment was to approximate the shell thickness of the samples from the 3D morphological tomography reconstructions, in order to verify the existence of an actual statistical difference in thickness. The second experiment is centred around the analytical separation of elements. The objective was to test if the mixed-composition iron-nickel oxide NPs presented an actual dual shell structure. To do so, a X-EDS analytical electron tomography was carried out, complementing a conventional analysis of EELS spectrum images.

3.4.1. Morphological electron tomography. Thick-Thin shell Fe_2O_3 NP.

The maghemite hollow-shell NPs samples where synthesized in a colloidal suspension, which helped with the long-term stability and prevented the

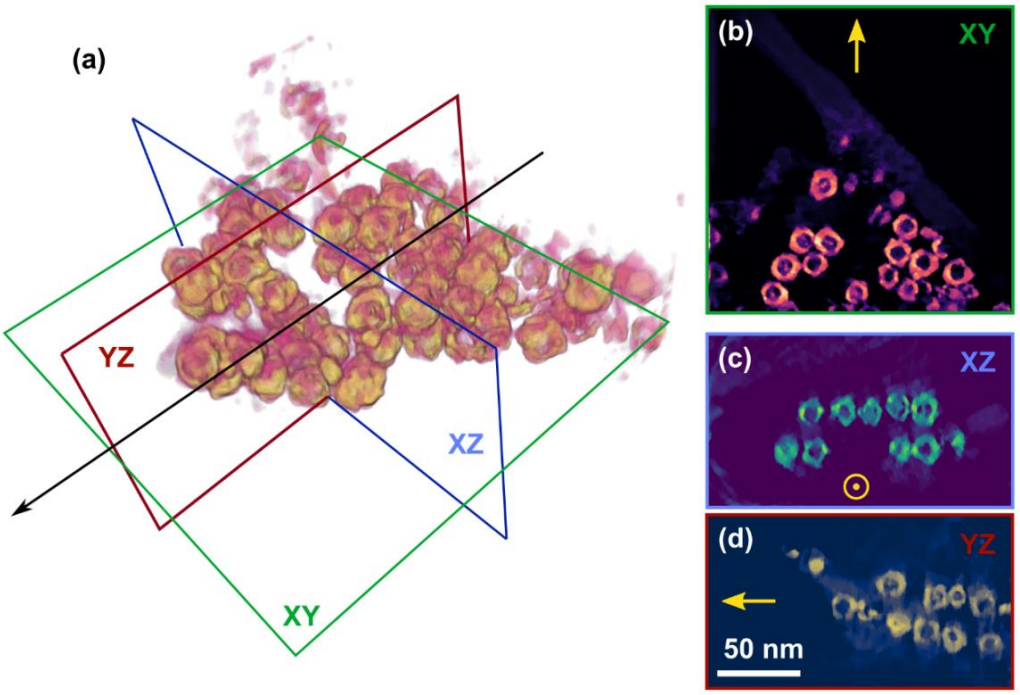


Figure 3.14. (a) Volume rendering of the TVM ET reconstruction of a large bundle of thin-shell γ-Fe_2O_3 NPs. The slices through the volume highlighted in the perspective view (a) are shown in (b-d) for the XY, XZ and YZ planes respectively. The rotation axis is indicated with an arrow in (a), (b) and (d), and a dot ('exiting the plane') in (c).

amalgamation of NPs in an eventual structural collapse. Unfortunately, NPs with these rests of colloidal shell tend to suffer from severe sample contamination under the electron beam. The molecular envelope is evaporated by the continuous energy input and later deposited on top of the sample, attracted by the negative charge of stream of electrons. This is translated into shorter acquisition times, limited number of projections and lower beam intensity (i.e., lower SNR) allowed for a tomography acquisition. Therefore, the TVM methodology was the only realistic choice to retrieve reconstructions with a high enough quality to perform a quantitative analysis.

The acquisition of the angular set of HAADF images for the thin-shell hollow γ-Fe_2O_3 NPs sample was done at 200keV in a TITAN Themis, with a beam current of 110pA. The sample was rotated from -65º to 60º, with an angular step of 5º.

The registration time was 10s per angular projection, for a total set of 27 images of 512x512 resolution. After the alignment of the stack of images, the reconstruction was done setting a penalty parameter $\mu = 2^7$.

The resulting volume reconstructed is shown in **Figure 3.14** (a), and the set of orthoslices through the volume are shown in **Figure 3.14** (b), (c) and (d) for the XY, XZ and YZ slicing planes highlighted in **Figure 3.14** (a). The arrows indicate the rotation axis for the sample during the HAADF stack acquisition. Despite the large number of closely packed NPs in the field of view, the quality of the reconstruction was high enough to extract individual NPs in some localized regions of interest through an automatized volume segmentation

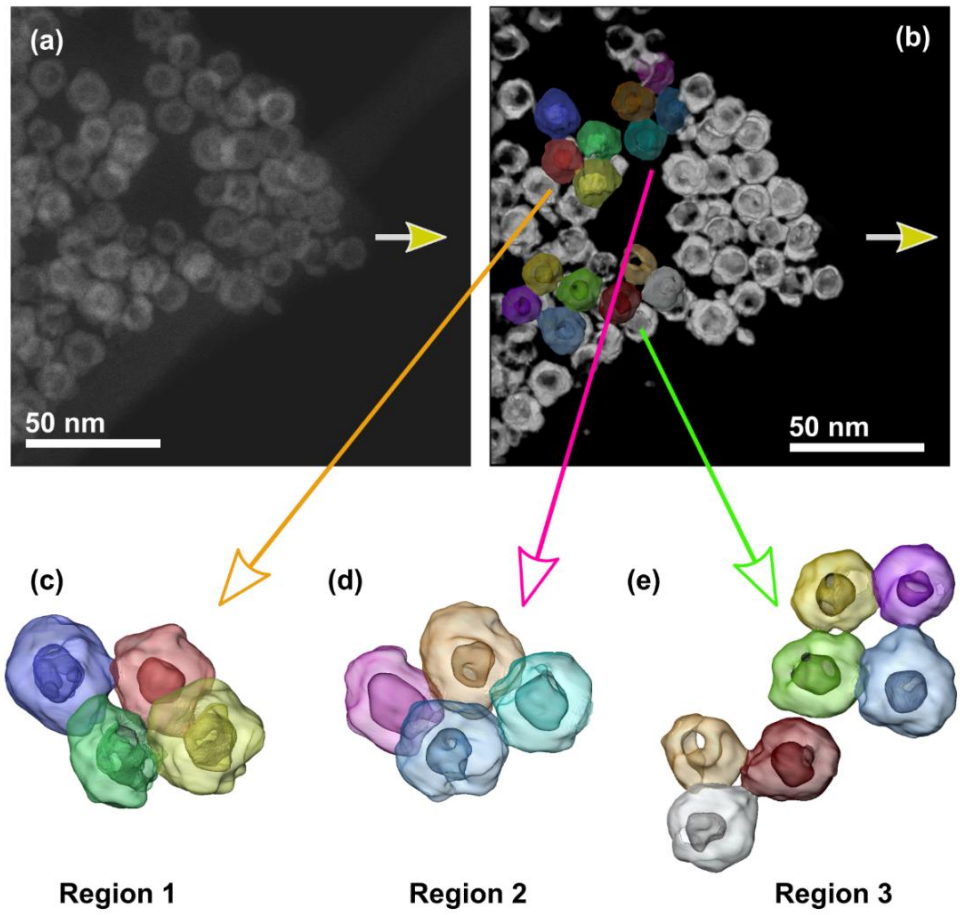


Figure 3.15. (a) HAADF projection at 0º for the hollow thin-shell γ-Fe_2O_3 NPs sample. (b) Perspective view of the volume rendering for the reconstruction in the same direction defined by the normal of the 0º projection plane. The isosurfaces of colour overlayed with the render are presented separately and zoomed in (c),(d) and (e).

technique. This segmentation was carried out by the watershed algorithm included in the Avizo software solution. It was applied to 3 handpicked regions presenting a good separation between particles, identified by making use of orthoslices similar from those presented in **Figure 3.14** (b), (c) and (d). The final segmentation results are shown in **Figure 3.15**.

The good agreement between the HAADF projection at 0º presented in **Figure 3.15** (a) and the perspective view of the volume rendering in **Figure 3.15** (b), highlights once again the high degree of accuracy in the reconstruction. The arrows in both panels indicate the rotation axis for the sample during the experiments.

The regions of interest 1 (**Figure 3.15** (c)), 2 (**Figure 3.15** (d)) and 3 (**Figure 3.15** (e)) present a total of 15 NPs separated. The isosurfaces shown in these panels reproduce the hollow shell structure of the γ-Fe_2O_3 NPs, and at the same time put forward the existence of holes in the shells connecting the internal void in the 'core' with the exterior. The isosurfaces also confirmed the

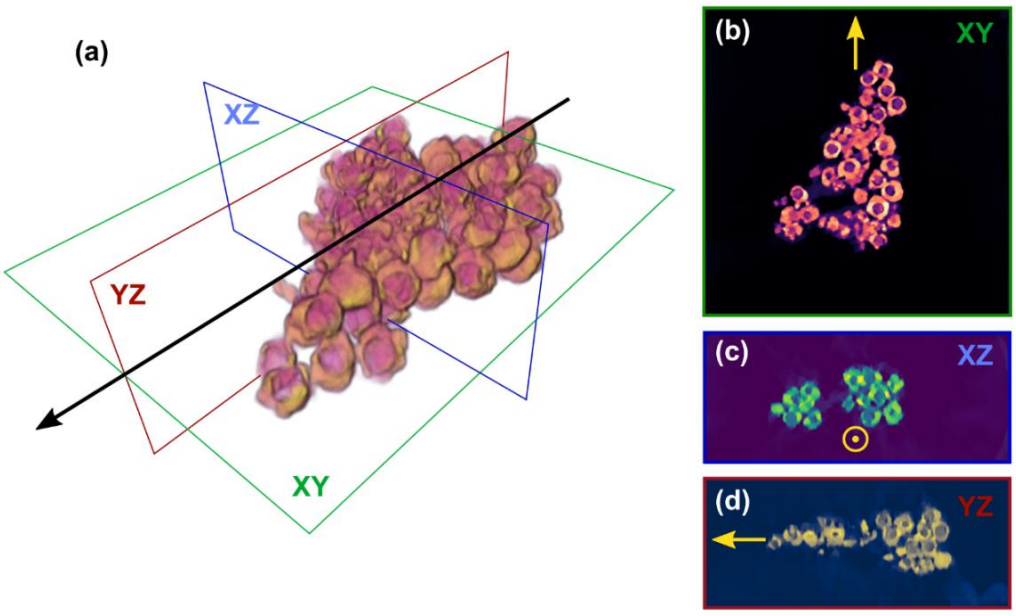


Figure 3.16. (a) Volume rendering of the TVM ET reconstruction of a large bundle of thick-shell γ-Fe_2O_3 NPs. The slices thought the volume highlighted in the perspective view (a) are shown in (b-d) for the XY, XZ and YZ planes respectively. The rotation axis is indicated with an arrow in (a), (b) and (d), and a dot ('exiting the plane') in (c).

existence of an actual inner structure (i.e., the cavity is not perfectly spherical) inside the hollow shell, as already hinted in the HAADF projections of the sample for most of the NPs. Overall, the NPs segmented appear to present a homogeneous morphology and size distribution.

The acquisition of the of 512x512 HAADF images for the thick-shell hollow γ-Fe_2O_3 NPs sample was done with the same experimental conditions of the previous thin-shell sample experiment. However, this time the angular range was ±65º, with a step of 5º between projections. The reconstruction was done

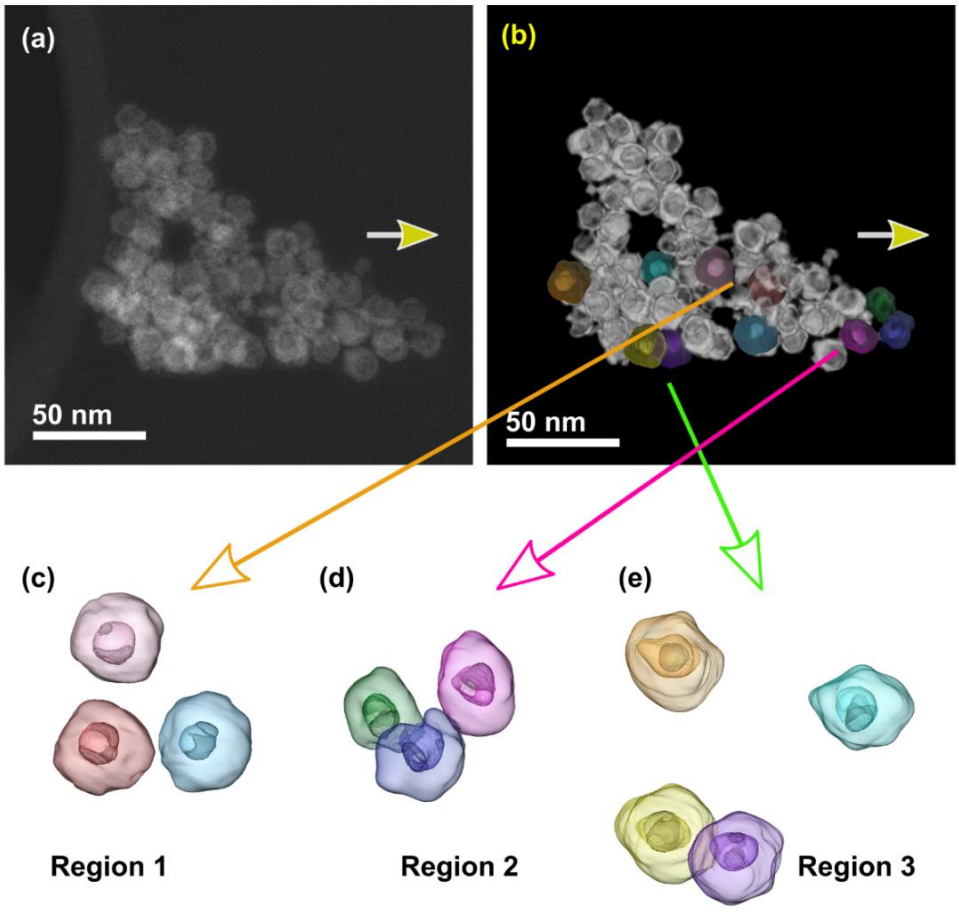


Figure 3.17. (a) HAADF projection at 0º for the hollow thick-shell γ-Fe_2O_3 NPs sample. (b) Perspective view of the volume rendering for the reconstruction in the same direction defined by the normal of the 0º projection plane. The isosurfaces of colour overlayed with the render are presented by separate and zoomed in (c), (d) and (e).

setting a penalty parameter $\mu = 2^7$ again, after the stack alignment and preparation.

The resulting volume is shown in **Figure 3.16** (a), alongside the orthoslices through the volume of the XY (**Figure 3.16** (b)), XZ (**Figure 3.16** (c)) and ZY (**Figure 3.16** (d)) planes. As with the thin shell case, several areas were selected for an automatic segmentation of particles using the watershed algorithm by observing these orthoslices. The results are shown in **Figure 3.17**.

Once again, a good agreement can be observed between the original HAADF projection at 0º (**Figure 3.17** (a)) and the perspective view of the reconstructed volume (**Figure 3.17** (b)). The isosurfaces for the segmented NPs in the areas of interest are shown in **Figure 3.17** (c), (d) and (e). Those same isosurfaces representations are overlayed with the volume rendering in **Figure 3.17** (b) as a reference of their position in the overall dataset. As in the

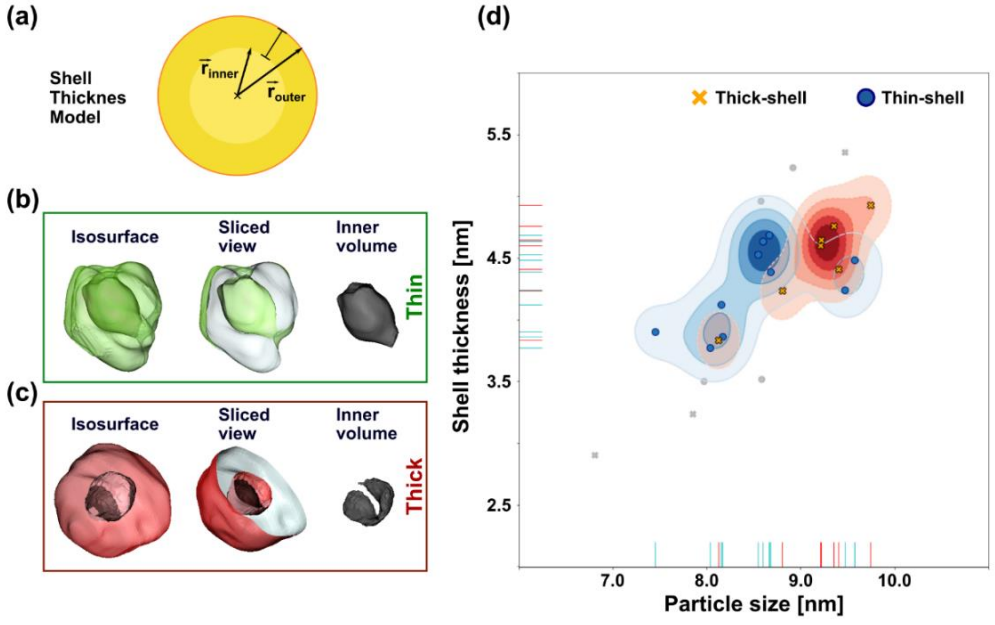


Figure 3.18. (a) Scheme of the model used for the thickness computations. (b) Example of the volumes separated for one of the thin-shell NPs. (c) Example of the volumes separated for one of the thick-shell NPs. (d) 2D chart for the distributions of shell thickness and NP size for the NPs separated in **Figure 3.15** and **Figure 3.17**.

thin-shell cases from, the thick-shell NPs present an inner structure and open pores connecting the core void and the exterior in many cases.

The analyses of **Figure 3.14**, **Figure 3.15**, **Figure 3.16** and **Figure 3.17** indicated that, although the NPs presented an overall spherical symmetry, the actual shell structure and size showed a certain degree of variability between the NPs segmented for the same sample composition. As a proxy for the shell thickness computation, a simpler model is proposed to represent each one of the particles separated from their respective bundles. The model represents each NP by a pair of concentric spheres. The larger one contains a volume equal to the total volume encompassed by the NP measured from the reconstruction. The smaller one is given a volume equal to that encompassed by the core region of the hollow structure of the NPs. As the majority of the NPs presented an open shell structure (a series of pores connected the central void in the core with the exterior), an approximation for the closing surface was required. A scheme for the equivalent volume construction is shown in **Figure 3.18**.

The spherical volumes (**Figure 3.18** (a)) computed from the segmented reconstructions (examples in **Figure 3.18** (b) and (c) for the thin-shell and thick shell NPs respectively) allow for the estimation of the shell thickness as the difference between the outer and inner radius of the spheres

$$r_i = \left(\frac{3V_i}{4\pi}\right)^{\frac{1}{3}} \tag{18}$$

where V_i stands for the volume of the sphere.

The resulting average value thickness value for the thin-shell NPs was 4.26 ± 0.31nm, whereas for the thick-shell case was 4.49 ± 0.34nm. Notice how close the resulting values are among them. Therefore, the actual differentiation between thick and thin shell samples is subtle. To complete the analysis, a 2D chart for the computed thickness vs the measured radius for the segmented NPs is included in **Figure 3.18** (d). The shaded areas correspond to an estimated probability density function for each one of the sets of segmented

NPs, assuming a gaussian distribution of probability around each datapoint (thick-shell NPs in red and thin-shell NPs in blue). Some of the datapoints are deemed as outliers for both distributions, and are accordingly shaded in grey. The analysis of this dual-parameter graph shows a much clearer separation between the thin and thick shells distribution.

To summarize the results, the use of the TVM methodology resulted in a set of high-quality morphological reconstructions for large bundles of NPs in both samples proposed, despite the limiting experimental conditions. From those reconstructions, several NPs could be automatically segmented by a semi-supervised algorithm (watershed), and the measurements from those separated volumes provided an insight of the distributions of shell thickness and sizes.

3.4.2.Analytical electron tomography. Iron - Nickel oxide NPs.

The sample of iron-nickel oxide NPs was expected to show a dual shell morphology, where the nickel oxide (NiO) shell should appear as an envelope structure of the inner iron oxide (γ-Fe_2O_3) shell, forming a dual-shell structure that enveloped the hollow core of the NP. However, the conventional EELS SI analysis showed inconclusive results in that regard (**Figure 3.19**, **Figure 3.20**). The non-negative matrix factorization (NMF) decomposition resolved a separation of Ni and Fe in the components with indices 3 and 4. The component labelled as #4 shows an important increment of the Ni WLs intensity with respect to Fe WLs, whereas component #3 shows a clear dominion of the spectra by the Fe WLs. However, the structure of the loading mappings for component #4 shows an uneven segregation towards the surface, instead of an enveloping shell structure. This 'clumps' of dominant NiO are observed clearly in the mappings overlays in both areas (**Figure 3.19** (c) and **Figure 3.20** (c)).

A detailed zoomed view of the Fe WLs is included in an inset on the spectrum areas in both cases (**Figure 3.19**(e) and **Figure 3.20**(e)), to highlight the apparent change in the oxidation state for the iron oxide accompanying the

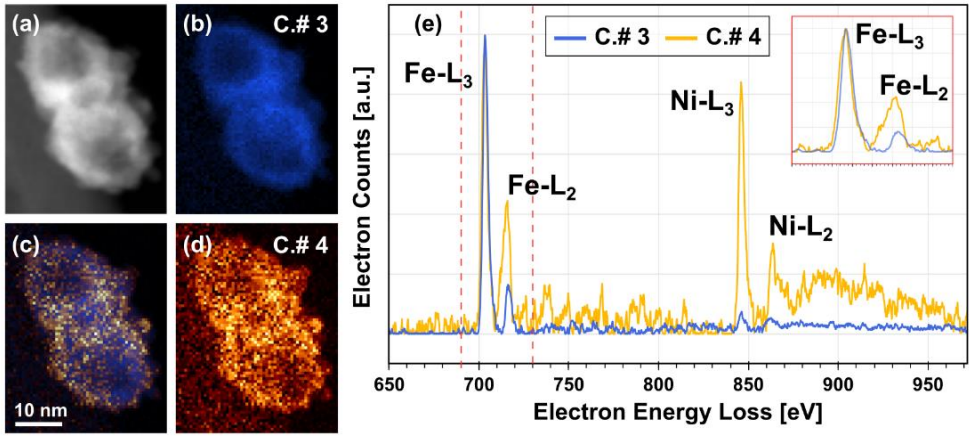


Figure 3.19. (a) HAADF co-acquired image for the first EELS-SI under analysis. The loading mappings for the NMF spectral factors (#3 and #4) shown in (e) are included in (b) and (d) (respectively). The inset of (e) shows a detailed view of the Fe WLs for the components resolved. (c) Overlay of loadings mappings for the components #3 and #4.

NiO segregation. The variation of the WL-ratio between component 4 and 3, and the displacement of the central position of the FeL_3 WL towards higher energy losses in component 3 are behaviours commonly described in the literature for the changes in the cation ELNES regions accompanying changes in the oxidation states[32]. Finally, it is important to mention that the spectra presented in both images (**Figure 3.19** and **Figure 3.20**) were independently scaled between (0,1) to allow an easier comparison of the resolved factors. In reality, component 3 presented in both areas an average number of electron-counts 2 orders of magnitude higher than component 4.

Moving on, a series of electron tomography experiments were carried out on the same sample to confirm that this apparent uneven distribution of NiO is not an artefact arising from the projection effect in a conventional EELS SI acquisition. Given that the Ni and Fe are close in the periodic table (the Z atomic numbers are 26 for the and 28 for the Ni), the regions with different compositions in the HAADF projection images used for the electron tomography reconstruction would not present any variation in contrast associated to elemental composition changes. Hence, a morphological electron

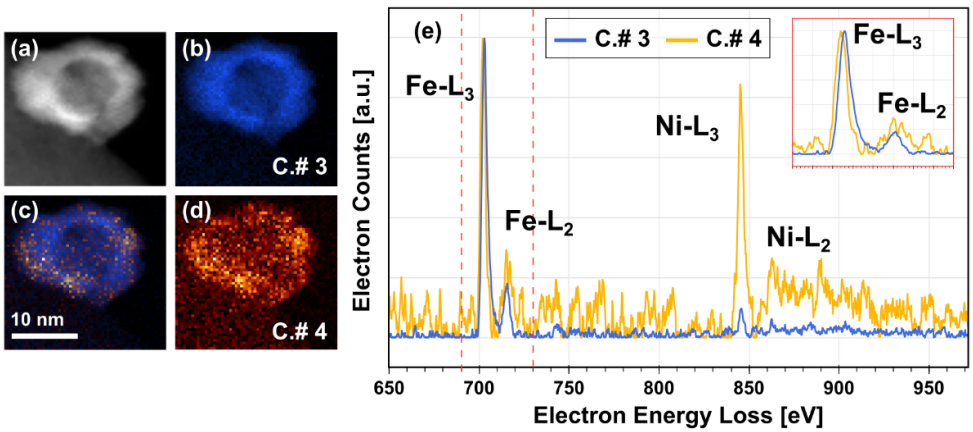


Figure 3.20. (a) HAADF co-acquired image for the second EELS-SI under analysis. The loading mappings for the NMF spectral factors (#3 and #4) shown in (e) are included in (b) and (d) (respectively). The inset of (e) shows a detailed view of the Fe WLs for the components resolved. (c) Overlay of loadings mappings for the components #3 and #4.

tomography experiment would not suffice to resolve the problem. To tackle this issue, the HAADF images were co-acquired with a series of X-EDS spectrum images. The mappings extracted from these spectrum images were later aligned using the HAADF projections (with a higher level of detail) and used to carry out an analytical electron tomography reconstruction, which retrieved a series of segmented volumes for each one of the elements of interest (Fe and Ni).

Once again, the problems with carbon contamination limited the maximum number of projections acquired to an angular range of ±65º and an angular step of 10º, limiting each projection to a 128x128 pixel-size area. The dwell time (pixel/time) was set as 20 milliseconds, resulting in an approximate 5min 30s per projection. As the spectral projections were acquired in a single scanning mode (a single illumination event per pixel), the electron beam current was increased to a value of 250pA to improve the SNR.

Each one of the projected SIs was identically modelled to extract the signal mappings for the Ni and Fe (same background model and integration windows for the reference lines). Both the Fe and Ni mappings were computed from the

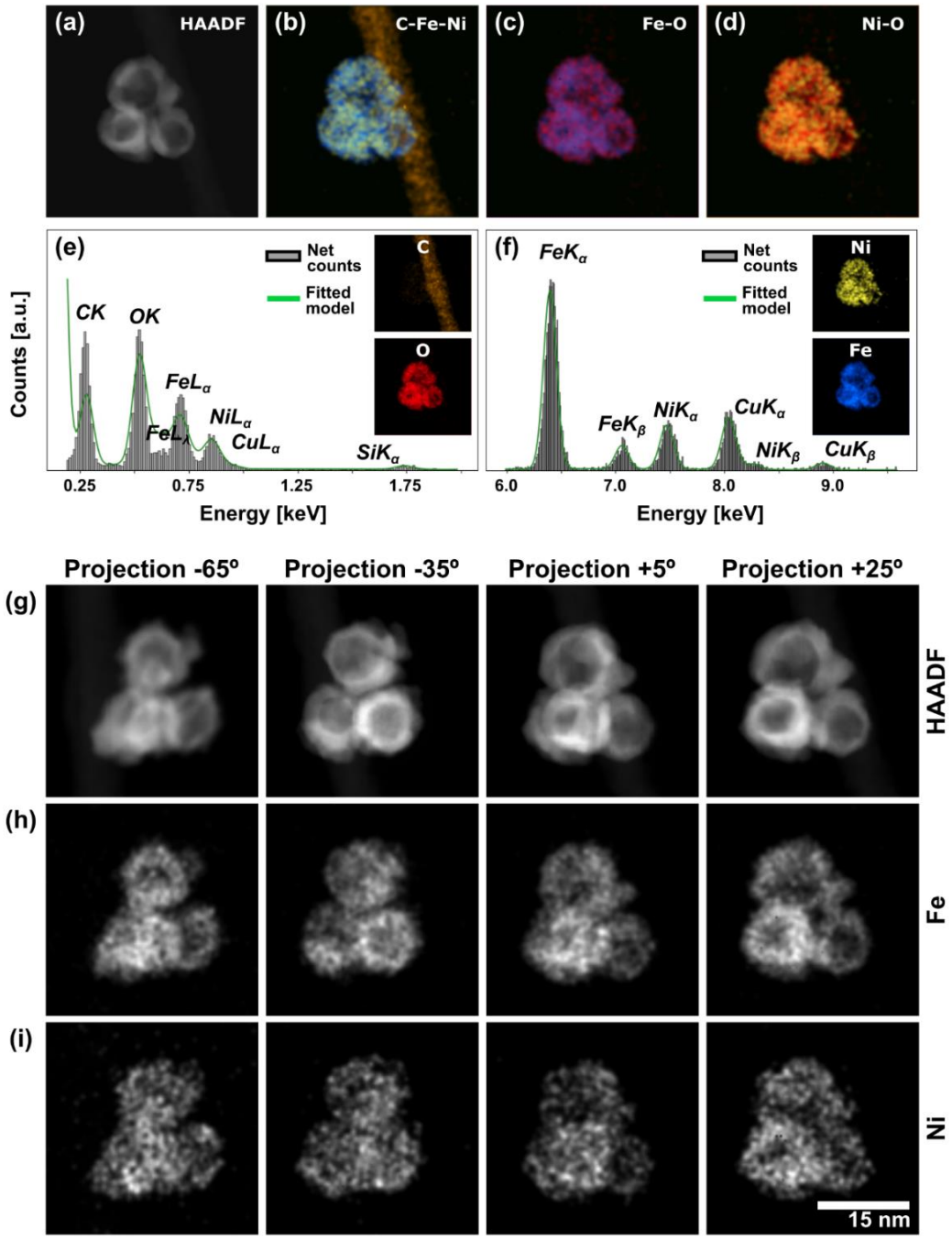


Figure 3.21. (a) HAADF projection for the NiO-γFe_2O_3 at an angle of 35º. The series of overlays (same projection angle) for the C-Fe-Ni (b), Fe-O (c) and NiO (d) elemental mappings were extracted from the spectral models shown in (e) for the low energy and (f) for the high energy regions. (g) Examples of projection images for the HAADF signal. (h) Examples of projection images for the Fe signal (elemental mapping). (i) Examples of projection images for the Ni signal (elemental mapping).

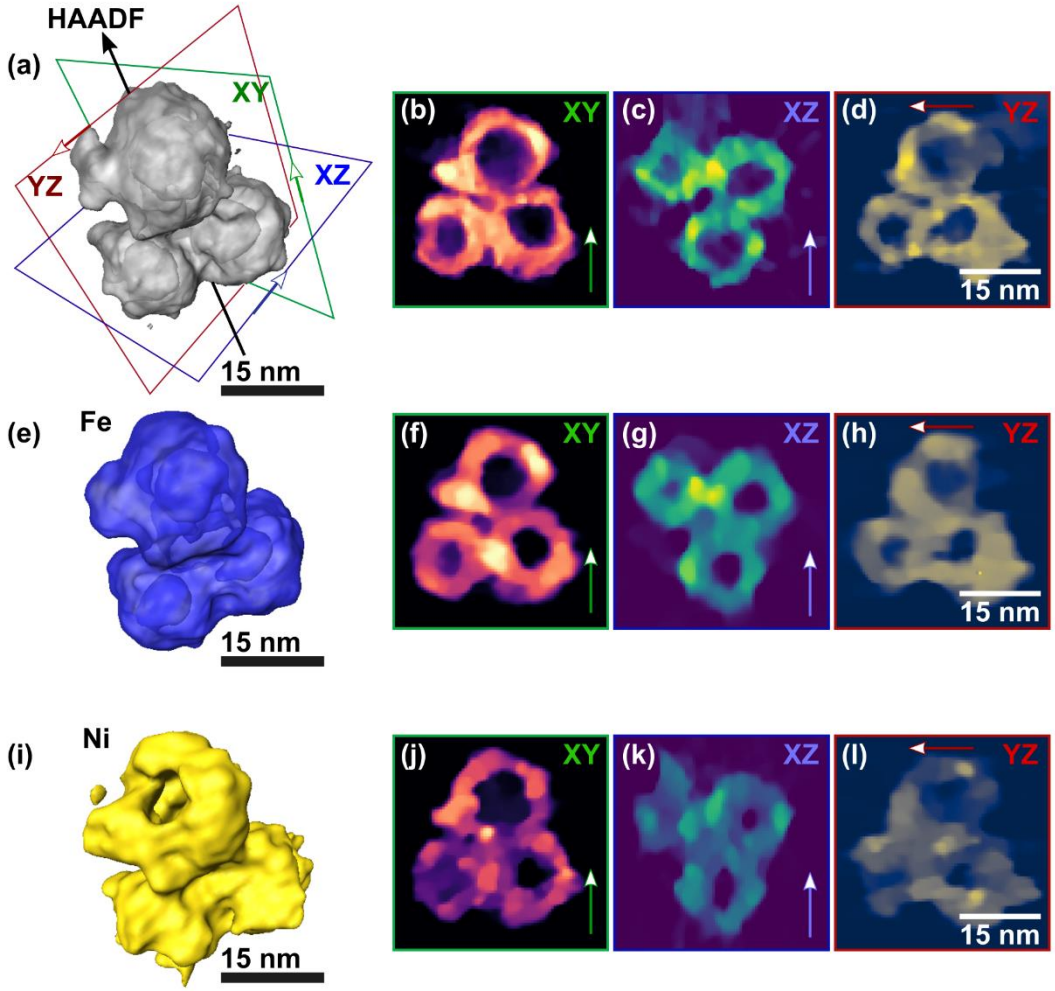


Figure 3.22. Results for the morphological and analytical electron tomography reconstructions for the NiO-γFe_2O_3 NPs. (a) Isosurface of the volume reconstructed from the HAADF projections. The rotation axis is indicated in the perspective representation. The slices through the HAADF volume are shown in (b) for the XY, (c) for the XZ and (d) for the YZ planes, as highlighted in the perspective volume in (a). (e) Isosurface of the volume reconstructed from the Fe projection mappings. The slices through the Fe volume are shown in (f) for the XY, (g) for the XZ and (h) for the YZ planes indicated in (a). (i) Isosurface of the volume reconstructed from the Ni projection mappings. The slices through the Ni volume are shown in (j) for the XY, (k) for the XZ and (l) for the YZ planes indicated in (a). All the slices share the same scale bar, indicated in all cases in the YZ mappings.

net intensity of the K_α and K_β lines in spectra, using the Velox software solution. Some examples of the X-EDS analysis previous to the electron tomography reconstruction are shown in **Figure 3.21**. Some examples of the elemental mappings for the oxygen and carbon signals are also included in **Figure 3.21** (b-d). However, these elements were not included as part of the analytical tomography reconstructions due to the low accuracy of the model proposed for the spectra in the low energy region (**Figure 3.21** (e)). Conversely, the model shows an accurate fit for the lines

in the high energy region (**Figure 3.21** (f)). **Figure 3.21** also includes examples of some of the HAADF projection images (g) used for the alignment and the reconstruction of the morphology of the NP bundle, alongside the mappings for the Fe and Ni elements at the same angular projections.

As in the case for the synthetic dataset in the previous section of this chapter, the HAADF reconstruction was done using a high value for the penalty parameter ($\mu = 2^8$) to reproduce a high level of detail of the structures and boundaries between volumes. Meanwhile, the segmented Fe and Ni reconstructions were done with a $\mu = 2^4$ to balance the severe undersampling of the elemental mappings extracted for the Fe and Ni signals. The isosurface representations of the reconstructed volumes are shown in **Figure 3.22**, alongside the slices through the volume in the three major planes (XY, XZ and YZ). The Fe volume follows closely the HAADF reconstruction, whereas the Ni volume presents a higher degree of granularity and some holes in the outer structure.

To complete the analysis and test if the Ni oxide shell covers completely the Fe oxide structure, a perspective view of the overlay of reconstructed volumes for the Fe and Ni signals is included in **Figure 3.23** (a). A quick look into the volume represented appears to confirm qualitatively that the NiO shell does not cover completely the Fe_2O_3 hollow inner-shell NP. The arrows in the isosurface renders from **Figure 3.23** (a) and (b) are also represented in this case over the orthoslices for the XY (**Figure 3.23** (c)), XZ (**Figure 3.23** (d)) and (**Figure 3.23** (e)) planes, to aid with the visual interpretation of the

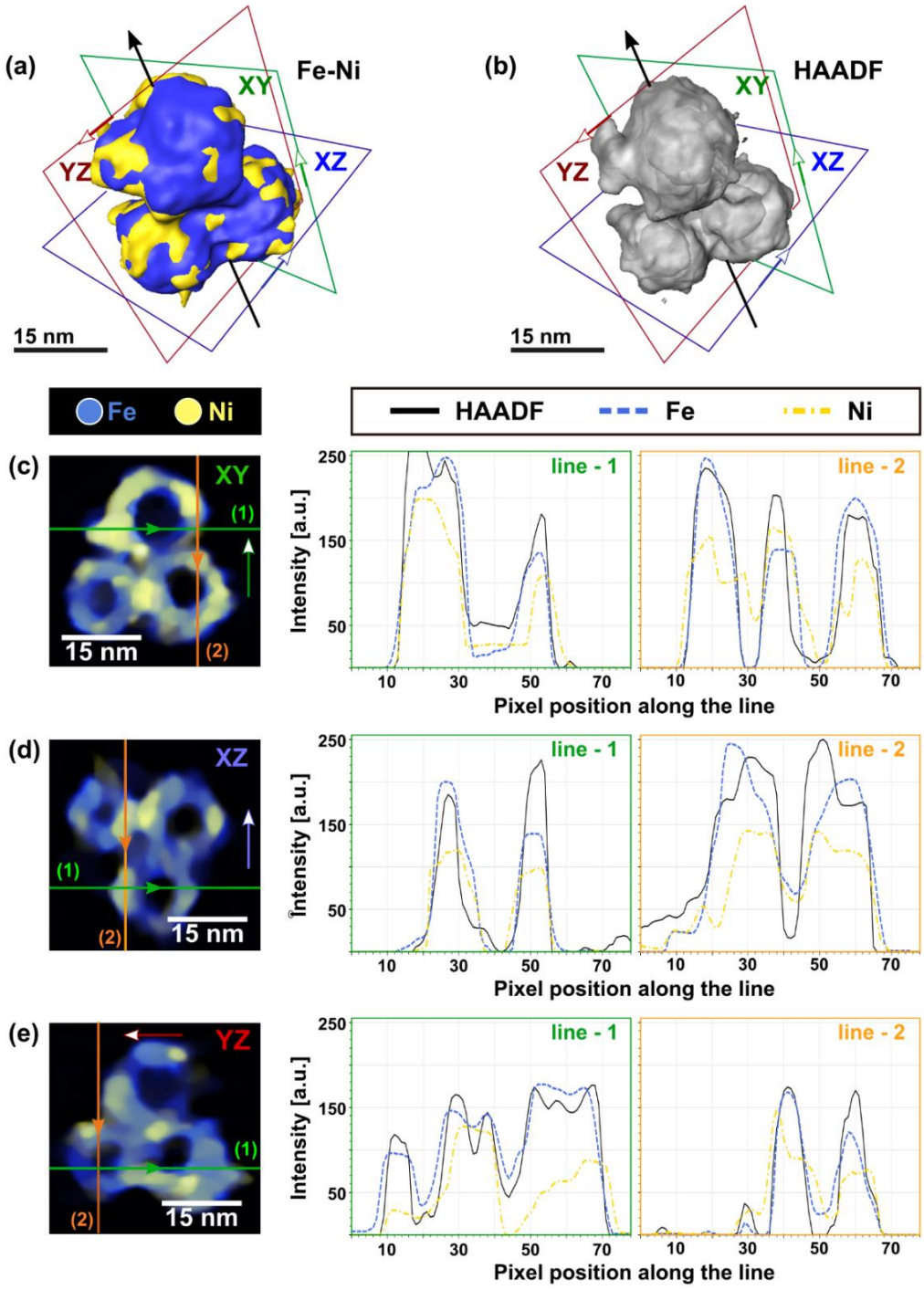


Figure 3.23. Results for the morphological and analytical electron tomography reconstructions for the NiO-γFe_2O_3 NPs. (a) Overlay of the isosurface renders for the volumes reconstructed for the Fe and Ni signals. (b) Isosurface for the morphological HAADF reconstruction. (c) Orthoslices and profiles for the XY plane. (d) Orthoslices and profiles for the XZ plane. (e) Orthoslices and profiles for the YZ plane.

orientation presented in them. A semi-quantitative analysis of profile lines over the sliced volume was also conducted for the 3 major planes.

At this point it would be important to remember the effects of the missing wedge on the different profiles through the reconstruction, already studied for the synthetic dataset in the previous section of this chapter (**Figure 3.7**, **Figure 3.8**, **Figure 3.9** and **Figure 3.11**). Generally, the rule of thumb is that the profiles along the illumination direction defined by the wedge present a smoothing artefact and loose the definition between boundaries of different objects. In this experiment, the missing wedge would be defined by the angle bisected by the YZ plane. Knowing that the rotation axis is defined by the intersection between the YZ and XY planes, the directions affected by this undersampling condition are those defined by the intersection of the YZ and XZ planes (i.e., precisely those directions marked with arrows for the XZ and YZ planes in **Figure 3.23** (a), (b) and in the orthoslices). More specifically, the profiles affected would be the vertical profile for the XZ plane along the line 1 in **Figure 3.23** (d), and the horizontal profile for the YZ plane along the line 2 in **Figure 3.23** (e). Furthermore, although both elemental reconstructions were done using the same reconstruction parameters, the differences in the SNR and image erosion between the sets of elemental mappings for the Fe and Ni (**Figure 3.21** (h) and (i) respectively) could affect the extent of the smoothing effect (as qualitatively observed for the synthetic dataset in **Figure 3.10**, **Figure 3.12**, **Figure 3.13**).

In any case, the profiles shown for the γ-Fe_2O_3-NiO NPs (**Figure 3.23** (c-e)), together with the isosurface render representations (**Figure 3.23** (a-b)), seem to corroborate the results from the EELS SI analysis in (**Figure 3.19**, **Figure 3.20**) that indicated a possible alloy of Fe and Ni oxides with occasional extrusions of the latter to the outer surface and inner cavity of the hollow NPs.

3.5. Conclusions

This chapter presents a review of the most relevant algorithms in the field of electron tomography (ART-SIRT). It also includes a comparison of SIRT (by far, the most popular 'go-to' solution for ET) with the more advanced TVM methodologies. The later are mostly based on the minimization of the total variation problem, that in the majority of ET applications is resolved by the TVAL3 algorithm (now made available in Python).

These new types of approaches were tested with a control synthetic dataset. It allowed the experimentation with challenging datasets for the reconstruction process, such as a those with a drastic reduction of the SNR of the projection images or the inclusion of severe erosion artefacts in the image to mimic the conditions of severe undersampling in some analytical mapping projections. TVM was also tested for several different number of initial projections used during the reconstruction, showing the high quality of the final results even for sets as limited as 11 total projection images. These tested conditions for the TVM would have certainly compromised the reconstruction results in the case of using the classic SIRT methodologies, as demonstrated by the large number of resulting artefacts recovered for the synthetic datasets.

Finally, the capability of the TVM methodology to produce quantitative results for the morphological ET and analytical ET was showcased through a series of experiments on iron oxide and nickel-iron oxide NPs.

3.6. References

[1] V. Lučić, F. Förster, W. Baumeister, STRUCTURAL STUDIES BY ELECTRON TOMOGRAPHY: From Cells to Molecules, Annu. Rev. Biochem. 74 (2005) 833–865. https://doi.org/10.1146/annurev.biochem.73.011303.074112.

[2] P. Torruella, J. Blanco-Portals, L. Yedra, L. López-Conesa, J.M. Rebled, F. Peiró, S. Estradé, Electron Tomography, Springer Ser. Mater. Sci. 308 (2021) 257–283. https://doi.org/10.1007/978-3-030-60473-8_11.

[3] P.A. Midgley, M. Weyland, 3D electron microscopy in the physical sciences: the development of Z-contrast and EFTEM tomography, Ultramicroscopy. 96 (2003) 413–431. https://doi.org/10.1016/S0304-3991(03)00105-0.

[4] S.R. Deans, The Radon transform and some of its applications, Courier Corporation, 2007.

[5] D.B. Williams, C.B. Carter, The Transmission Electron Microscope, in: Transm. Electron Microsc., Springer US, Boston, MA, 1996: pp. 3–17. https://doi.org/10.1007/978-1-4757-2519-3_1.

[6] R.F.F. Egerton, Electron Energy-Loss Spectroscopy in the Electron Microscope, Springer Science & Business Media., Boston, MA, 2011. https://doi.org/10.1007/978-1-4419-9583-4.

[7] L. Yedra, A. Eljarrat, J.M. Rebled, L. López-Conesa, N. Dix, F. Sánchez, S. Estradé, F. Peiró, EELS tomography in multiferroic nanocomposites: from spectrum images to the spectrum volume, Nanoscale. 6 (2014) 6646–6650. https://doi.org/10.1039/C4NR01100G.

[8] P. Torruella, R. Arenal, F. de la Peña, Z. Saghi, L. Yedra, A. Eljarrat, L. López-Conesa, M. Estrader, A. López-Ortega, G. Salazar-Alvarez, J. Nogués, C. Ducati, P.A. Midgley, F. Peiró, S. Estradé, 3D Visualization of the Iron Oxidation State in FeO/Fe_3O_4 Core–Shell Nanocubes from

Electron Energy Loss Tomography, Nano Lett. 16 (2016) 5068–5073. https://doi.org/10.1021/acs.nanolett.6b01922.

[9] C. Li, W. Yin, H. Jiang, Y. Zhang, An efficient augmented Lagrangian method with applications to total variation minimization, Comput. Optim. Appl. 56 (2013) 507–530. https://doi.org/10.1007/s10589-013-9576-1.

[10] B. Goris, W. Van den Broek, K.J. Batenburg, H. Heidari Mezerji, S. Bals, Electron tomography based on a total variation minimization reconstruction technique, Ultramicroscopy. 113 (2012) 120–130. https://doi.org/10.1016/j.ultramic.2011.11.004.

[11] J.M. Muñoz-ocaña, A. Bouziane, F. Sakina, R.T. Baker, A.B. Hungría, J.J. Calvino, A.M. Rodríguez-chía, M. López-haro, Optimization of STEM-HAADF Electron Tomography Reconstructions by Parameter Selection in Compressed Sensing Total Variation Minimization-Based Algorithms, Part. Part. Syst. Charact. 37 (2020) 2000070. https://doi.org/10.1002/ppsc.202000070.

[12] M. López-Haro, M. Tinoco, S. Fernández-Garcia, X. Chen, A.B. Hungria, M.Á. Cauqui, J.J. Calvino, A Macroscopically Relevant 3D-Metrology Approach for Nanocatalysis Research, Part. Part. Syst. Charact. 35 (2018) 1700343. https://doi.org/10.1002/ppsc.201700343.

[13] P.F. Gilbert, The reconstruction of a three-dimensional structure from projections and its application to electron microscopy. II. Direct methods., Proc. R. Soc. London. Ser. B. Biol. Sci. 182 (1972) 89–102. https://doi.org/10.1098/rspb.1972.0068.

[14] J. Radon, Über die Bestimmung von Funktionen durch ihre Integralwerte längs gewisser Mannigfaltigkeiten, in: Books.Google.Com, 1983: pp. 71–86. https://doi.org/10.1090/psapm/027/692055.

[15] G. Beylkin, Discrete Radon Transform, IEEE Trans. Acoust. 35 (1987)

162–172. https://doi.org/10.1109/TASSP.1987.1165108.

[16] P. Gilbert, Iterative methods for three dimensional reconstruction of an object from projections, J. Theor. Biol. 36 (1972) 105–117. https://doi.org/10.1016/0022-5193(72)90180-4.

[17] R. Gordon, R. Bender, G.T. Herman, Algebraic Reconstruction Techniques (ART) for Three-dimensional Electron Microscopy and X-ray Photography, J. Theor. Biol. 29 (1970) 471–481. https://doi.org/10.1016/0022-5193(70)90109-8.

[18] W. van Aarle, W.J. Palenstijn, J. De Beenhouwer, T. Altantzis, S. Bals, K.J. Batenburg, J. Sijbers, The ASTRA Toolbox: A platform for advanced algorithm development in electron tomography, Ultramicroscopy. 157 (2015) 35–47. https://doi.org/10.1016/j.ultramic.2015.05.002.

[19] P.A. Midgley, R.E. Dunin-Borkowski, Electron tomography and holography in materials science, Nat. Mater. 8 (2009) 271–280. https://doi.org/10.1038/nmat2406.

[20] Z. Saghi, G. Divitini, B. Winter, R. Leary, E. Spiecker, C. Ducati, P.A. Midgley, Compressed sensing electron tomography of needle-shaped biological specimens – Potential for improved reconstruction fidelity with reduced dose, Ultramicroscopy. 160 (2016) 230–238. https://doi.org/10.1016/J.ULTRAMIC.2015.10.021.

[21] P. Penczek, M. Marko, K. Buttle, J. Frank, Double-tilt electron tomography, Ultramicroscopy. 60 (1995) 393–410. https://doi.org/10.1016/0304-3991(95)00078-X.

[22] B. Bougher, Introduction to compressed sensing, Lead. Edge. (2015) 1256–1258. https://doi.org/http://dx.doi.org/10.1190/tle34101256.1.

[23] C.E. Shannon, Communication in the Presence of Noise, Proc. IEEE. 72 (1984) 1192–1201. https://doi.org/10.1109/PROC.1984.12998.

[24] M.W. Marcellin, M.J. Gormish, A. Bilgin, M.P. Boliek, An overview of JPEG-2000, (2002) 523–541. https://doi.org/10.1109/DCC.2000.838192.

[25] R. Leary, Z. Saghi, P.A. Midgley, D.J. Holland, Compressed sensing electron tomography, Ultramicroscopy. 131 (2013) 70–91. https://doi.org/10.1016/J.ULTRAMIC.2013.03.019.

[26] E. Candès, J. Romberg, Sparsity and incoherence in compressive sampling, Inverse Probl. 23 (2007) 969–985. https://doi.org/10.1088/0266-5611/23/3/008.

[27] C. Li, Compressive Sensing for 3D Data Processing Tasks: Applications, Models and Algorithms, 2011. https://www.caam.rice.edu/~optimization/L1/TVAL3.

[28] M.R. Hestenes, Multiplier and gradient methods, J. Optim. Theory Appl. 1969 45. 4 (1969) 303–320. https://doi.org/10.1007/BF00927673.

[29] F. Bleichrodt, T. van Leeuwen, W.J. Palenstijn, W. van Aarle, J. Sijbers, K.J. Batenburg, Easy implementation of advanced tomography algorithms using the ASTRA toolbox with Spot operators, Numer. Algorithms 2015 713. 71 (2015) 673–697. https://doi.org/10.1007/S11075-015-0016-4.

[30] E. Biermans, L. Molina, K.J. Batenburg, S. Bals, G. Van Tendeloo, Measuring porosity at the nanoscale by quantitative electron tomography, Nano Lett. 10 (2010) 5014–5019. https://doi.org/10.1021/nl103172r.

[31] J. Zečević, K.P. de Jong, P.E. de Jongh, Progress in electron tomography to assess the 3D nanostructure of catalysts, Curr. Opin. Solid State Mater. Sci. 17 (2013) 115–125. https://doi.org/10.1016/j.cossms.2013.04.002.

[32] P.A. Van Aken, B. Liebscher, V.J. Styrsa, Quantitative determination of iron oxidation states in minerals using Fe L2,3-edge electron energy-loss near-edge structure spectroscopy, Phys. Chem. Miner. 25 (1998) 323–327. https://doi.org/10.1007/s002690050122.

Chapter 4

Machine learning for EELS data analysis. Clustering and dimensionality reduction methods.

4.1. Introduction.

The rapid advances in transmission electron microscopy (TEM) instrumentation experienced over the past decades have led a dramatic increment in the size and complexity of the acquired datasets Every TEM characterization has been affected: from the ever-growing CCD[1,2] and CMOS electron detectors[3–5] for imaging modes, to the improvements in the electronics and control software for the scanning modes of the electron beam that allowed the advent of the dual core-loss / low-loss experiments in electron energy loss spectroscopy[6] (EELS), and the simultaneous acquisition of energy dispersive x-ray spectroscopy (X-EDS) and EELS datasets (even combined with dual-EELS)[7].

This is not an isolated issue concerning the TEM and materials science community, but rather a global trend towards the analysis and processing of increasingly larger volumes of data. In fact, these circumstances have prompted the appearance of the data mining, big data and data science fields of study[8,9]. Within them, the dimensionality reduction methods (DRM), clustering algorithms, neural networks and many more analysis techniques, commonly grouped in the category of (and referred as) machine learning (ML) algorithms, are constantly evolving towards the objective of a better, faster and more comprehensive (fully)data-driven analysis. Precisely these same methodologies have steadily been filtered back to the research community:

from clustering analysis in astrophysics[10], database visualization in cellular studies[11], or even the use of neural networks for the classification of TEM images[12].

Switching back to TEM, although the introduction of ML techniques influenced many of the available imaging and spectral characterization techniques, EELS data analysis has to be listed as one of the biggest beneficiaries from it, if not the biggest. In a way, the addition of ML into the standard data processing pipeline leaded to a paradigmatic change in the manner that many EELS experiments were conducted. For example, some experimental procedures consciously reduced the exposure times for the EELS spectral acquisition, trading-off a worst signal-to-noise ratio for the possibility of acquiring larger (but noisier) datasets without excessive electron doses being transferred to the sample. This opens the possibility of carrying out longer experiments on sensitive samples prone to beam damage, otherwise forced to be analysed by the use of a cryo-TEM[13] or to the acquisition of punctual spectra in some limited number of handpicked regions of the sample[14] (e.g., acquiring spectra of biological samples[15], or carrying out oxidation state experiments in large areas of easily reducible rare earth oxides[16]).

The exploration of ML techniques for EELS data analysis during the period of my doctoral thesis mainly revolved around dimensionality reduction and clustering analysis techniques. This self-contained chapter presents an exhaustive revision of the theory behind each one of the algorithms visited. Each one of them is tested with an experimental dataset of known composition: an iron-manganese oxide core-shell nanoparticle (NP) with two different oxidation states coexisting in the manganese oxide shell[17].

The chapter is organized as follows:

- ***Reference sample for the analysis.***

  The first section of the chapter is an introduction to the EELS spectrum image (SI) that will be used as the reference dataset to test the inner workings of the DRM and clustering algorithms.

- ***Dimensionality reduction methods (DRM).***

  This section is subdivided into matrix factorization problems and graph-based approaches. The introduction of the uniform manifold approximation and projection(UMAP) algorithm as a non-linear DRM for EELS data analysis is included alongside the description of some of the most commonly used algorithms.

- ***Clustering analysis for EELS datasets.***

  This third section includes a taxonomical revision of the main algorithms found in the literature concerning EELS data analysis, and the introduction of HDBSCAN as the superior option for a fully data-driven and unsupervised classification.

- ***Combination of dimensionality reduction and clustering analysis.***

  The fourth section describes the combined use of UMAP and HDBSCAN, and the possible inclusion of a NMF pre-analysis step for further separation of spectral features.

- ***Performance analysis: quantitative scores, noise resilience and outlier detection.***

  The fifth and final section is devoted to a series of controlled experiments with different synthetic datasets, devised to approximate numerically the performance of the different algorithms presented (and some of the possible combinations between them). Through these experiments, the qualitative analysis of the DRM and clustering segmentation results (i.e., the label maps resolved) are compared to quantitative numerical scores. Also, the noise and outliers (faint spectral characteristics) detection limits for an accurate segmentation using the combination UMAP-HDBSCAN are tested. For future reference and to put in perspective the times listed for each method, all the calculations in this work were done using a laptop with Intel® Core™ i5-8250 CPU at 1.6 GHz (boosted up to 3.4 GHz) with 16 Gb of RAM in a Windows 10 home (64 bits).

A Jupyter notebook for data treatment analysis using the techniques described in this work may be found here (http://hdl.handle.net/2445/179410).

4.2. Reference sample for the analysis.

In order to test the different algorithms described in this chapter and explore their innerworkings, an EELS spectrum image (SI) of an iron oxide – manganese oxide core-shell nanoparticle is utilized. This dataset is proposed as the reference sample due to its already well known composition and spectral features from previous characterizations [17,18]. A summary of these spectral and compositional properties is shown in **Figure 4.1**. The sample is composed by a monophase core region of iron oxide, surrounded by a dual phase shell of manganese oxide (see the difference in the signals for the pixels shown in **Figure 4.1** (a) and (b), and the composite images of **Figure 4.1** (c) and (d)).

The dual phase in the shell was resolved by a multiple linear least squares fitting (MLLS) experiment on the denoised spectrum image. The results of the MLLS showed an uneven distribution of weights (**Figure 4.1** (d)) for the two different manganese oxide reference signals extracted from the dataset itself (**Figure 4.1** (e)). These signals used for the fitting are the average EELS spectra for the regions of the shell marked in **Figure 4.1** (a)), which were selected according to their proximity to the iron oxide core. These reference signals already show a clear difference in the white lines (WL) L_3/L_2 ratio and a possible chemical shift towards higher energy loss values for the signal in area 2.

From the MLLS fitting, 3 quantitative maps were extracted (using the Oxide Wizard software solution[17]). **Figure 4.1** (f) shows the L_3/L_2 WL ratio, **Figure 4.1** (g) shows the full width half maximum (FWHM) values of the dominant L_3 line of the manganese edge, and **Figure 4.1** (h) shows the distance to the oxygen K edge of the Mn L_{32} edge onset. The latter is a more robust measurement of the chemical shift of the Mn L_{32} edge than the usual direct

measurement of the L_3 position, as it minimizes the eventual effects of random energy fluctuations during the acquisition.

The results shown are consistent with the spectral changes reported in the literature[19] for variations of the Mn oxidation state. In this case, area 1 (i.e., the outermost regions of the shell) was resolved as an almost pure MnO phase.

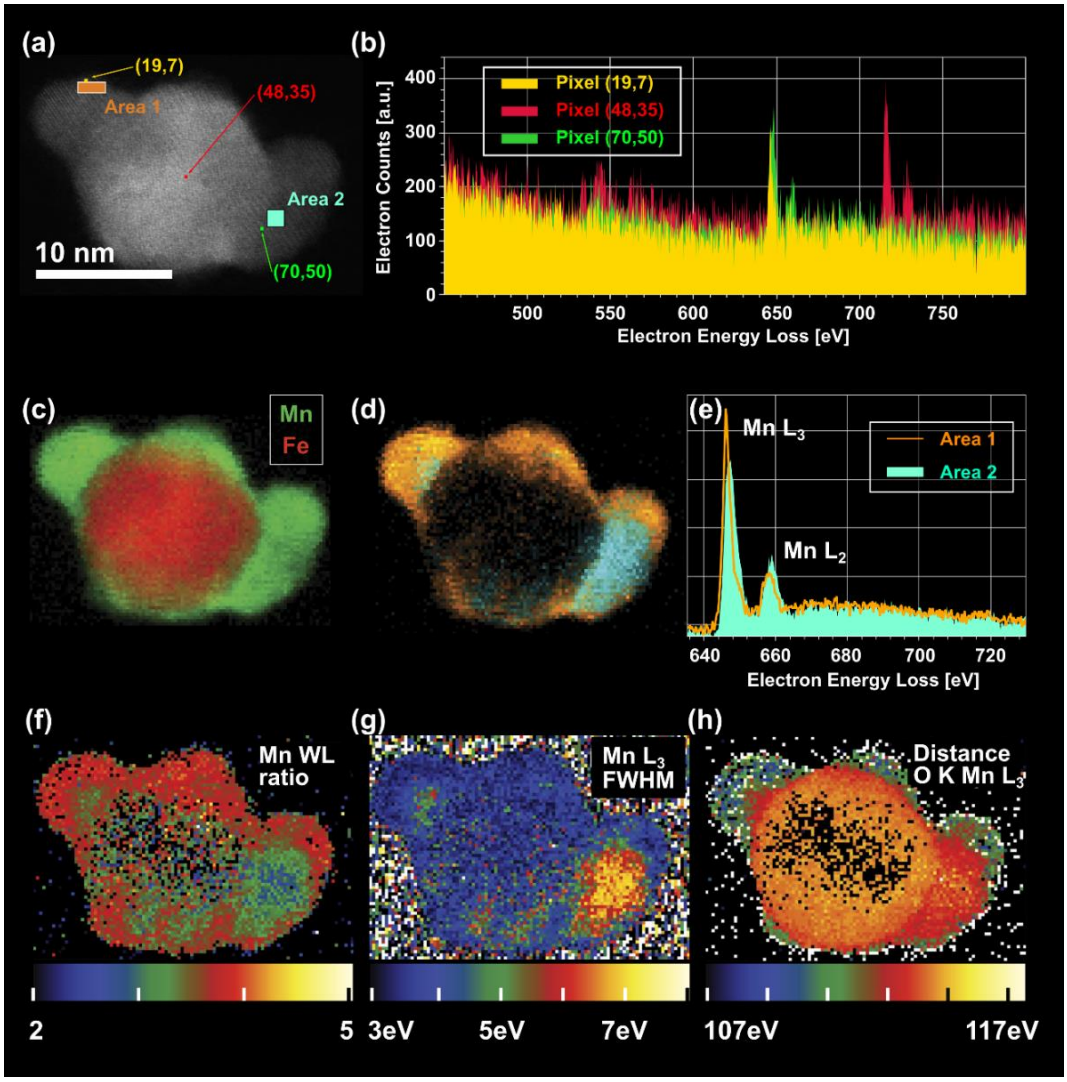


Figure 4.1. (a) High resolution HAADF image of the Fe-Mn oxide core-shell NP used as a reference dataset. (b) Raw spectra for the pixels (19,7), (48,35) and (70,50) marked in (a). (c) Colour composite image of the weights for the Mn_yO_x and Fe_3O_4 resolved by a MLLS fitting. (d) Composite image of the MLLS weights resolved for the shell signals (Mn_yO_x) shown in (e). (e) Reference average signals of the areas 1 and 2 marked in the HAADF image in (a), used in the MLLS fitting for (d). (f) Manganese L_3 / L_2 white line ratio, (g) full width half maximum of the Mn L_3 and (h) O K edge to Mn L_3 edge distance map resolved by Oxide Wizard [17].

Meanwhile, the spectra from area 2 (i.e., the innermost part of the shell) showed signal characteristics compatible with a mixed Mn^{2+} and $Mn^{2.66+}$ oxidation state[1], indicating that the electron path intersected both the MnO and Mn_3O_4 phases.

This sample is an excellent candidate for the test of DRM and clustering algorithms due to its well-defined spectral characteristics constrained to different areas of the spectrum image.

4.3. Dimensionality reduction methods (DRM).

Dimensionality reduction (or feature extraction) refers to the problem expressing as faithfully as possible the original dataset in a compressed representation capable of producing the same analytical results[9]. These types of methods have been routinely included in EELS data analysis for a long time, being principal component analysis (PCA)[15,18,20–23] and non-negative matrix factorization (NMF)[24–26] the most commonly used algorithms. Although initially conceptualized as endmember extraction methods for EELS data analysis (i.e., to represent the whole spectral dataset, commonly an EELS-SI, by a limited set of reference signals or archetypes), they have been reduced to a denoising pre-processing step most of the times in our work. This approach has been described in many occasions in the literature[27–29].

Up until recently, the vast majority of dimensionality reduction algorithms applied in EELS analysis could be ascribed the larger family of linear approaches known as matrix factorization techniques (to which PCA and NMF belong). However, some new non-linear approaches have recently made their way into the field, namely t-distributed stochastic neighbour embedding (t-SNE) [30,31] and uniform manifold approximation of projections

[1] Expressing the manganese with a fractional oxidation state $Mn^{2.66+}$ is a common notation used to express the 2:1 ratio of Mn oxidation states in the Mn_3O_4 phase, as it would actually present 2 Mn^{3+} and 1 Mn^{2+} ions per unit cell.

(UMAP)[32,33]. These methods do not seek to find some hypothetical set of endmembers that linearly combined recover the original EELS dataset, but they rather try to project into a lower dimensional embedding (typically $\mathbb{R}^2$or $\mathbb{R}^3$) the hypothetical manifold that describes the dataset in the original hyperdimensional space. This underlying manifold is often constructed from a connected graph computed from the dataset itself, and then projected into the lower dimension by a force directed layout[32]. Hence, these kinds of approaches are sometimes referred as the graph-based dimensionality reduction methods.

Each methodology and each specific algorithm have its advantages and shortcomings when it comes to dimensionality reduction of EELS datasets, all of which are described in the following pages.

4.3.1. Matrix factorization.

The majority of the dimensionality reduction methods used for EELS data analysis (usually of large datasets, such as SI) fall under the common definition of matrix factorization problems. A general description of a matrix factorization problem goes along the following lines (see **Figure 4.2**).

(1) The complete dataset is described by a matrix $\mathbf{X} \in \mathbf{R}^{m\times n}$ of m rows of datapoints ($\vec{\boldsymbol{x}}_i$), and n columns of properties or characteristics.

(2) The goal of a matrix factorization is to obtain a new representation matrix ($\mathbf{Z} \in \mathbf{R}^{m\times n}$) from the original dataset ($\mathbf{X}$), generally with a lower k rank. There are an infinite number of $\mathbf{Z}$ solutions to such problem. Thus, the challenge becomes the search for the optimum solution in an efficient and accurate way. The easiest (and first) approach would be to resolve a least squares minimization problem for the Frobenius distance between both matrices, which is described by the following equation

$$\mathrm{d}_{Fro}(\boldsymbol{X},\boldsymbol{Z}) = \frac{1}{2}\|\boldsymbol{X}-\boldsymbol{Z}\|_{Fro}{}^2 \qquad (1)$$

and so, excluding the constant terms in the actual minimization problem

$$\text{minimize}\left(\mathrm{d}_{Fro}(\boldsymbol{X},\boldsymbol{Z})\right) \sim \text{minimize } \|\boldsymbol{X}-\boldsymbol{Z}\|_{Fro}{}^{2} \tag{2}$$

$$\text{minimize} \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\boldsymbol{X}_{ij}-\boldsymbol{Z}_{ij}\right)^{2} \tag{2.1}$$

The squared norms in eq.((1)) and (2) will be expressed from now on as a summatory of the squared differences elementwise, as shown in eq (2.1).

(3) A convenient way to encode the constraint of the lower (or equal) k into the equation is to think about $\mathbf{Z}$ as the inner product between a loading or representation matrix $\mathbf{L} \in \mathbf{R}^{m\times k}$ and an archetypes matrix $\mathbf{A} \in \mathbf{R}^{k\times n}$, and the minimization problem becomes

$$\text{minimize} \sum_{i=1}^{m}\sum_{j=1}^{n}\left(\boldsymbol{X}_{ij}-(\boldsymbol{LA})_{ij}\right)^{2} \tag{3}$$

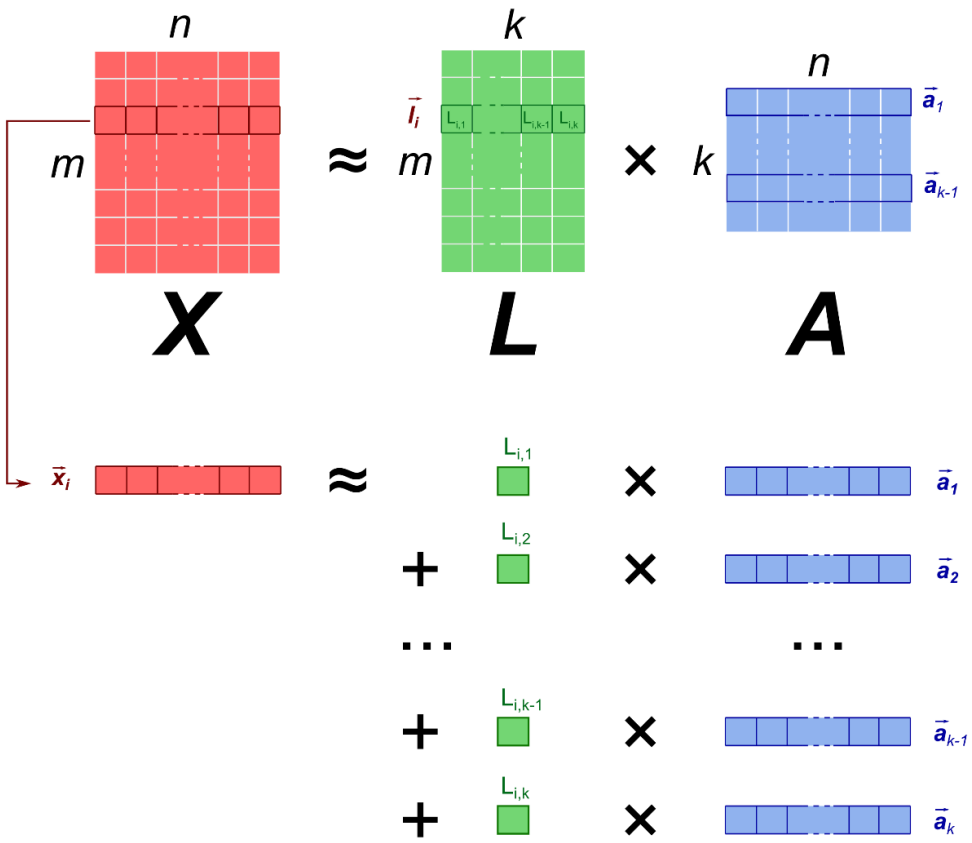


Figure 4.2. General schematics for a matrix factorization problem $\mathbf{X} \approx \mathbf{LA}$, where each datapoint is resolved as a linear combination of archetypes, via the coefficients in the representation matrix.

(4) Finally, each of the datapoints ($\vec{\boldsymbol{x}}_i$, row in $\mathbf{X}$) is expressed as a linear combination of the coefficients $\mathrm{L}_{i,j}$ (from a row $\vec{\boldsymbol{l}}_i$) in the representation matrix ($\mathbf{L}$) and the archetype vectors $\vec{\boldsymbol{a}}_j$ (row from matrix $\mathbf{A}$).

$$\vec{\boldsymbol{x}}_i = \sum_{j=1}^{k} \mathrm{L}_{i,j}\ \vec{\boldsymbol{a}}_j \tag{4}$$

Under the umbrella of this general definition one can find principal component analysis (PCA), sparse PCA (sPCA), non-negative matrix factorization (NMF or NNMF), and even K-means (commonly referred to as a clustering algorithm).

The differences between these different methods lie on the constraints applied on the general problem presented in eq.(3). For example, PCA only requires the lower rank constraint on Z, and can be described by the general equation in (3).

$$\mathbf{PCA} \ \rightarrow \ \min\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - (\boldsymbol{LA})_{ij}\right)^2\right) \tag{5}$$

Whereas NMF further imposes a non-negativity constraint in both the archetypes and representation matrices.

$$\begin{aligned}\mathbf{NMF} \ \rightarrow \ \min&\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij} - (\boldsymbol{LA})_{ij}\right)^2\right)\\ s.t. \quad & L_{ij} \geq 0 \ \ and \ A_{ij} \ \geq 0\end{aligned} \tag{6}$$

Sparse PCA imposes the 'sparsity' condition by limiting the maximum number of non-zero values on the representation matrix below the target k rank for $\mathbf{Z}$ (i.e., the number of non-zero values for each row of $\mathbf{L} \in \mathbf{R}^{\,m \times k}$ is lower than k). Thus, the number of archetypes ($\vec{\boldsymbol{a}}_j$) that represent each of the datapoints $\vec{\boldsymbol{x}}_i$ is reduced, as most L_{ij} are zeroed (see **Figure 4.2**), which may help with the interpretability of the results (i.e., lesser archetypes per datapoint facilitates the interpretation of the model resolved).

$$\mathbf{sPCA} \;\rightarrow\; \min\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij}-(\boldsymbol{LA})_{ij}\right)^2\right) \quad s.t.\;\; \|\boldsymbol{L}\|_2 = 1 \; and \; \|\boldsymbol{L}\|_0 \leq k \tag{7}$$

K-Means, understood as a matrix factorization problem, would be the extreme case by which any given row of the representation matrix is only allowed to have a single non-zero value (and so, each datapoint is classified by a single archetype, called centroid, in a single cluster)[2]

$$\mathbf{K-Means} \;\rightarrow\; \min\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij}-(\boldsymbol{LA})_{ij}\right)^2\right) \quad s.t.\;\; \|\boldsymbol{L}\|_2 = 1 \; and \; \|\boldsymbol{L}\|_0 = 1 \tag{8}$$

where the $\|\boldsymbol{L}\|_2$ gets the largest singular value (as in singular value decomposition, SVD) of $\mathbf{L}$ (i.e., the largest eigenvalue of $\mathbf{L}\cdot\mathbf{L}^*$), and $\|\boldsymbol{L}\|_0$ refers to the number of $L_{i,j}$ different from 0 in each row (sparsity condition).

Although all the problems have been formally introduced by assuming a minimization target function based on the Frobenius distance, some of the practical implementations for these problems may assume other distance measurement. For example, the NMF problem is sometimes revolved from a Kullback-Leibler (KL) divergence (distance)

$$\mathrm{d}_{KL}(\boldsymbol{X},\boldsymbol{Z}) = \sum_{i,j} X_{i,j}\log\frac{X_{i,j}}{Z_{i,j}} - X_{i,j} + Z_{i,j} \tag{9}$$

Once again, in a minimization problem the constant parameters can be simplified, and here the logarithmic expression can be expanded

[2] The target result of the minimization problem with these extreme constraints of sparsity seeks to classify all the datapoints by a single archetype each, which is equivalent to a clustering classification. Nonetheless, this is far from the actual K-Means computation. The reason to label it as 'K-means dimensionality reduction' will become clear once the generalized probabilistic approach to the matrix factorization problem is introduced, and after the description of K-Means in the clustering section is completed.

$$\text{minimize}\left(\mathrm{d}_{KL}(\boldsymbol{X},\boldsymbol{Z})\right) \sim \text{minimize}\left(\sum_{i,j} Z_{i,j} - X_{i,j}\log Z_{i,j}\right) \tag{10}$$

and recovering the constraints for the NMF problem and representation and archetypes matrices we get[3]

$$\textbf{NMF} \rightarrow \min\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\left((\boldsymbol{LA})_{ij} - X_{i,j}\log(\boldsymbol{LA})_{ij}\right)^2\right)$$
$$s.t.\;\; L_{ij} \geq 0 \;\; and \; A_{ij} \geq 0 \tag{11}$$

The formulation of exponential PCA[4] takes a probabilistic approach to the matrix factorization problem. It incorporates the information about the underlying probability distribution generating the dataset to the matrix factorization problem. Knowing how the data was generated allows for the selection of a suitable statistical descripting model (e.g., a dataset composed by a series of counts as the datapoints properties may be described by a Poisson distribution model).

In this approach, one must think about the reconstruction $\mathbf{Z}$ matrix not as a representation of $\mathbf{X}$ itself (factorized as $\boldsymbol{Z} = \boldsymbol{LA}$), but as the best model parameters ($\boldsymbol{\eta}$) for the accurate description of $\mathbf{X}$ (i.e., the parameters that will make $\mathbf{X}$ the most likely observation for the proposed model). Thus, $\mathbf{X}$ is described by a probability density function

$$\boldsymbol{X} \sim f(\,\cdot \mid \boldsymbol{\eta}\,) \tag{12}$$

where f is the conditional probability for $\mathbf{X}$ under the $\boldsymbol{\eta}$ parameters, and these model parameters play the equivalent role to the factorized result (i.e., knowing the model parameters is equivalent to resolving the low rank model proposed, and thus it plays the role of the factorization matrices)

---

[3] This becomes relevant once the probabilistic approach in introduced, with an underlying Poisson distribution for the modelled data.

[4] Exponential PCA or PCA with an underlying probability distribution described by a function from the exponential family of distributions.

$$\boldsymbol{\eta} = \boldsymbol{LA} \tag{13}$$

To find now the values for the parameters (i.e., resolve the low rank model that makes $\mathbf{X}$ as likely as possible) the loss function becomes the so-called negative log likelihood[5]

$$-\log\big(\mathcal{L}(\,\boldsymbol{X} \mid \boldsymbol{\eta}\,)\big)^{6} \tag{14}$$

and the probability function introduced in eq.(12) can be parameterized by any of the probability distributions from the exponential family of distributions, generally described as

$$f(\boldsymbol{X} = \boldsymbol{x}|\boldsymbol{\eta}) = h(\boldsymbol{x})\exp\Big(\boldsymbol{\eta}^{\boldsymbol{T}}\boldsymbol{T}(\boldsymbol{x}) - G(\boldsymbol{\eta})\Big) \tag{15}$$

Each of the distributions from the exponential family will have a different expression for the functions $h(\boldsymbol{x})$ (undelying measure), $\boldsymbol{T}(\boldsymbol{x})$ (sufficient statistics), $G(\boldsymbol{\eta})$ (log normalizer) and the natural parameters $\boldsymbol{\eta}$[34]. The important result here is that the negative log likelihood (i.e., the loss function in the new formulation for the minimization problem) becomes

$$-\log\big(\mathcal{L}(\,\boldsymbol{X} \mid \boldsymbol{\eta}\,)\big) \;\propto\; G(\boldsymbol{\eta}) - \boldsymbol{\eta}^{T}\boldsymbol{T}(\boldsymbol{X}) \tag{16}$$

Thus, the minimization problem expressed in familiar terms becomes

[5] Usually, in statistics the maximum likelihood estimation (MLE) is a method to approximate the parameters of a probability distribution from the likelihood function for a measured dataset. For a parameterized family of probability density functions ($\boldsymbol{x} \to f(\,\boldsymbol{x} \mid \boldsymbol{\eta}\,)$), the likelihood function is $\boldsymbol{\eta} \to f(\,\boldsymbol{x} \mid \boldsymbol{\eta}\,)$. Formally written, $\boldsymbol{\mathcal{L}}(\,\boldsymbol{\eta} \mid \boldsymbol{x}\,) = f(\,\boldsymbol{x} \mid \boldsymbol{\eta}\,)$ (i.e., when the probability density function f is viewed as a function of the parameters $\boldsymbol{\eta}$ of a parameterized function of the family of density functions, it is called the likelihood function). If the density function is dependent of multiple parameters (i.e., $\boldsymbol{\eta}$ is a vector of parameters), the likelihood function is the product of likelihoods for each individual parameter. This prompts the use of the logarithmic scale, as a product conveniently becomes a summation. The log likelihood is expressed as $\boldsymbol{l}(\boldsymbol{\eta} \mid \boldsymbol{x}\,) = \log(\boldsymbol{\mathcal{L}}(\,\boldsymbol{\eta} \mid \boldsymbol{x}\,))$ Furthermore, most of the common probability distributions (including, notably, the exponential family) are logarithmically concave, which is important for their use as cost functions in MLE[80]. Furthermore, as log functions grow monotonically, to maximize $\boldsymbol{\mathcal{L}}$ is equivalent to maximize $\boldsymbol{l}$**.** In the matrix factorization problem, the objective is the minimization of a loss function.

[6] As the log-likelihood is positively defined, one can instead use the negative log-likelihood $-\log(\boldsymbol{\mathcal{L}}(\,\boldsymbol{\eta} \mid \boldsymbol{x}\,))$.

$$\min\left(-\log\left(\mathcal{L}(\boldsymbol{X}\mid\boldsymbol{\eta})\right)\right) \sim \min\left(\sum_{i=1}^{m}\sum_{j=1}^{n} G\left((\boldsymbol{LA})_{i,j}\right) - X_{i,j}\cdot(\boldsymbol{LA})_{i,j}\right) \tag{17}$$

And the function $G(\boldsymbol{\eta})$ can be looked upon for any specific distribution of the exponential family desired.

For instance, if one supposes that a Gaussian distribution (single parameter, mean $\boldsymbol{\mu}$,with known unit variance) is likely to accurately describe the dataset, the log normalizer function becomes

$$G(\boldsymbol{\eta}) = \frac{\mathbf{1}}{2}\boldsymbol{\eta}^{\mathbf{2}} \tag{18}$$

and, thus, we get the normal matrix factorization minimization problem

$$\min\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{2}\left((\boldsymbol{LA})_{i,j}\right)^2 - X_{i,j}\cdot(\boldsymbol{LA})_{i,j}\right) \tag{19}$$

As indicated before, adding or subtracting constant values to the minimization problem is allowed, and thus we can add $\frac{1}{2}X_{i,j}^{\ 2}$ and complete squares to get

$$\min\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\frac{1}{2}\left(X_{i,j} - (\boldsymbol{LA})_{i,j}\right)^2\right) \tag{20}$$

without constraints. That is, we recover the PCA matrix factorization (see eq.(5)). Thus, PCA matrix factorization is assuming a normally distributed dataset around a mean μ value encoded in $\boldsymbol{L} \in \boldsymbol{R}^{m\times k}$ and $\boldsymbol{A} \in \boldsymbol{R}^{k\times n}$ (which is always a good first approach if no extra information is known about the analysed dataset).

Trying now a Poissonian distribution (with an arrival rate λ as the sole parameter $\boldsymbol{\eta}$), the log normalizer function becomes

$$G(\boldsymbol{\eta}) = e^{\boldsymbol{\eta}} \tag{21}$$

and thus,

$$\min\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\exp(\boldsymbol{LA})_{i,j} - X_{i,j}\cdot(\boldsymbol{LA})_{i,j}\right) \tag{22}$$

Recovering now the formulation for NMF minimization problem where the KL-distance was the loss function (eq.(11)), one can clearly see the relation between both equations (a re-parametrization of the Poisson distribution with log (λ) recovers the original NMF formulation, and the obvious non-negativity constraint is translated to the non-negativity of the arrival rate). In this case (see eq.(22)), no constraints are applied directly on the minimization problem and, thus, the problem calculations are simplified.

All this new statistical formulation is introduced here to highlight how some of the most common matrix factorization methods can be grouped together under the exponential PCA formulation, and also how the different methods may be inadvertently assuming a certain statistical distribution for the dataset (e.g., PCA assumes an underlying normal distribution and the KL approach of NMF assumes a Poissonian distribution). This statistical considerations on the underlying nature of the datasets will surface once again when discussing the clustering algorithms classified as parametric (e.g., K-means, hierarchical agglomerative) and the resulting consequences for the datapoint classification.

Back to the specifics of matrix factorization for EELS data analysis**, Figure 4.3** shows a scheme of the correspondence between the formal terminology introduced above for a generalized matrix factorization problem and an example of PCA dimensionality reduction analysis for an EELS SI. The experimental data in (a) corresponds to the results of the decomposition (factorization), targeting a number of 14 components (rank k = 14), on the spectrum image of the iron oxide - manganese oxide core - shell nanoparticle showcased in **Figure 4.1**.

In EELS data analysis, the m rows of $\mathbf{X}$ correspond to the complete set of datapoints (spectra) acquired, and the n values X_{ij} in each $\vec{\boldsymbol{x}}_i$ (row) are the electron counts for each one of the channels in the spectrometer. In the case of

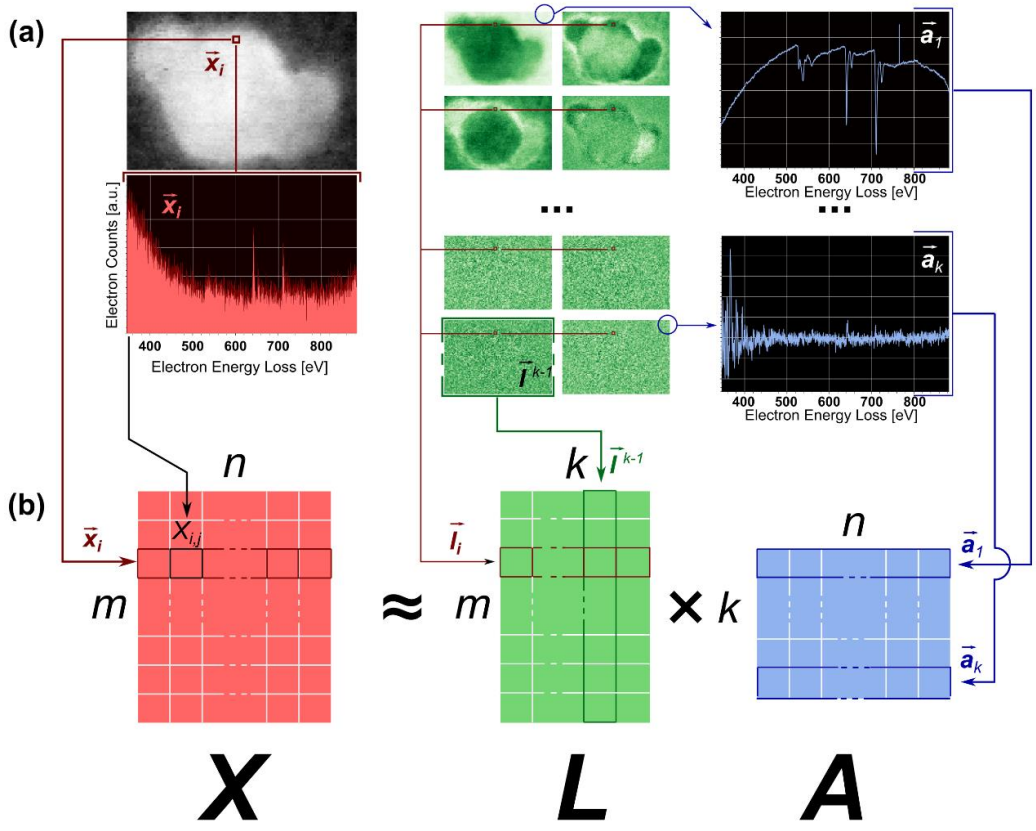


Figure 4.3. (a) EELS SI of the core shell iron-manganese oxide nanoparticle and the results of the PCA decomposition via the sklearn implementation of the algorithm. (b) Schematics of the general problem of matrix factorization, linked to the real-case experimental problem of (a).

a spectrum line (SL), m is the number of points along the line scanned. In a SI (as the one in the figure), the image itself losses the spatial 2D dimensionality as it is 'flattened' (reorganized) into a single dimension (no specific ordering is required, as the spatial distribution of pixels in the image does not affect the problem, as long as all the columns are flattened cohesively). Usually, the SI is organized following a left-to-right and top-to-bottom pattern. After the PCA decomposition, the loadings matrix $\mathbf{L} \in \mathbf{R}^{m \times k}$ is often presented as a set of k images, recovered from the columns $\vec{\boldsymbol{l}}^{j=1,\ldots,k}$ (i.e., reorganizing the column vectors as the original 2D SI). The archetypes matrix (commonly called the factors matrix among the EELS community) $\mathbf{A} \in \mathbf{R}^{k \times n}$ is presented as the set of factoring spectra $\vec{\boldsymbol{a}}_{j=1,\ldots,k}$.

As already mentioned, the same dataset from **Figure 4.1** and **Figure 4.3** (Fe-Mn oxide core-shell NP) will be the test subject to exemplify the effects of the most commonly used matrix factorization techniques in EELS data analysis. In all cases, a total number of 7 dimensions was set as the objective k-rank (underlying dimension), as a dual oxidation state for the manganese oxide (MnO and Mn_3O_4) and a single magnetite phase (Fe_3O_4) for the iron oxide were expected (the 3 individual phases and all the possible combinations between them, plus the background signal, add up to 7 possible different regions). The algorithms chosen are listed in **Table 4.1**, alongside the specific hyperparameters selected for all of them in the scikit-learn implementation used. The results for each of the included algorithms are showcased in **Figure 4.4**, **Figure 4.5**, **Figure 4.6** and **Figure 4.7**

Table 4.1. Parameters for the sklearn implementation of the algorithms. The selected number of components in all cases was 7. The field 'other' is included to further specify the options selected for some of the characteristic parameters of each algorithm

Algorithm	***Solver***	***Tolerance***	***Max # Iterations***	***Others***
PCA	SVD solver – 'randomized' [81]	-	-	-
sPCA	Coordinate descent[82] 'cd'	10^{-5}	500	Alpha factor 0.075
NMF	Multiplicative update[83] 'mu'	10^{-5}	500	Kullback-Leibler (KL) beta-loss
K-Means	-	10^{-5}	300	25 initializations kmeans++[62]

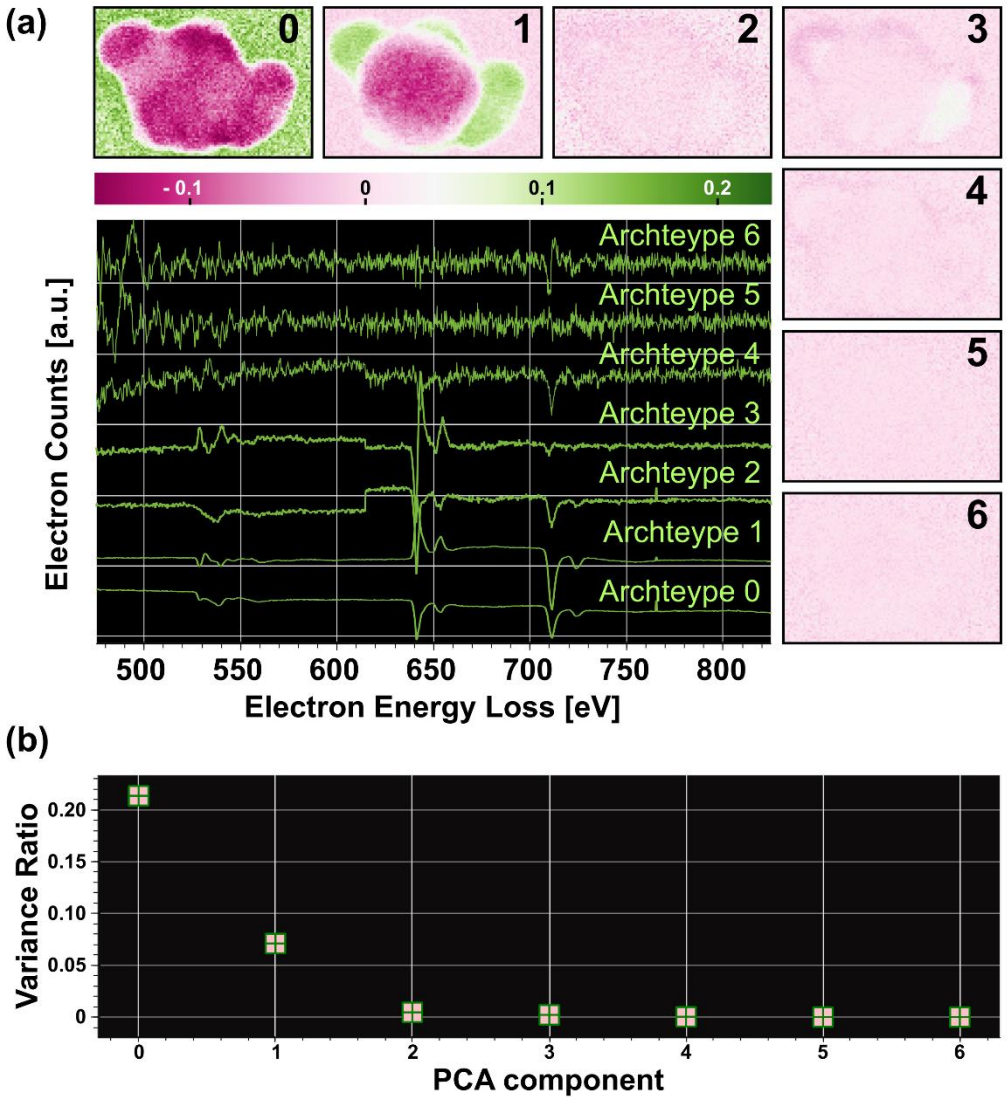


Figure 4.4. (a) PCA results for the Iron-Manganese oxide nanoparticle EELS SI (L_2-normalized). The colour scale is shared by all images. The archetypes or factor signals are not re-scaled, but they are displaced on the y-axis to fit in a single plotting area. (b) Scree plot for the explained variance ratio of the PCA decomposition.

The interpretation of the PCA decomposition (**Figure 4.4**) results is complicated. The representation (loadings) matrices (i.e., the component images) share the same colour scale, and a clear correlation between the matrix values and the scree plot (b) for each component is clear. The higher the variance ratio, the higher the average values for the m pixels in the representation matrix (**L**) for a given component k. In a classic PCA

decomposition analysis, one would be tempted to cut out the components below the knee of the scree plot (b), clearly located on component number 1. Nonetheless, components 2, 3 and 4 show some structure in the images, and the archetypes (the factor curves) also indicate the presence of spectral features of interest. For instance, component 3 could be interpreted as the one signalling a possible change in the manganese oxide oxidation state, by showcasing a displacement of the Mn white-lines to higher energy loss values and changes in the oxygen edge ELNES. Conversely, components 5 and 6 could probably be entirely ascribed to spectral noise.

There are however several problems associated to a PCA decomposition like the one displayed in **Figure 4.4**, the first one being the already mentioned difficulty of interpretability. In a general PCA problem, having a dense L matrix means that for each datapoint all the archetypes resolved have a non-zero weight (i.e., each datapoint is a linear combination of each and every one of the archetypes, and not a combination of a limited number of them). In EELS data-analysis this means that each pixel in the image is a combination of several reference signals (archetypes) with a non-zero weight value (loadings), some of which present opposing physical characteristics in the spectra. Evaluating how these interact with each other often becomes a cumbersome job, and is clearly an activity prone to suffer from confirmation bias (i.e., the posterior analysis of the resolved model is likely done with some a-priori information about the dataset analysed).

Furthermore, the archetypes are oftentimes plagued with non-physical spectral features, such as negative electron counts, inverted white-lines, sharp 'squared' shoulders and 'count valleys'. This would be a clear red-flag if, for instance, one would like to utilize them as the endmembers references for a MLLS fitting in a quantitative analysis.

Also, archetypes 5 and 6 (i.e., the noise components), show wider amplitude of random oscillations at the lower energy loss channels. This is an indication of Poissonian noise (higher count values produce a larger noise response), as these areas on the spectra present higher electron counts due to the

background decay from the low loss region. This noise effect is mitigated in some cases by applying a variance-stabilizing transform over the data, such as the Anscombe transform[35,36]. It is usually done previous to the decomposition step, to effectively translate the Poissonian noise into normal Gaussian (white) noise. Nonetheless, this imposes a non-negativity constraint on the electron counts data, which may be an issue in datasets post-processed by subtracting the dark-current signal (automatically done in some equipment and usually referred at as dark-correction).

Some extra concerns have been raised about the use of PCA in EELS data analysis in the literature. For example, PCA imposes orthogonality conditions on the components resolved implicitly (one of the reasons for the appearance of non-physical attributes in the spectral archetypes). This was tackled by some authors by formulating refined matrix factorization techniques, such as spectral unmixing (SU)[37].

Figure 4.5 shows the results for the sPCA decomposition. In the sklearn implementation, a penalty component is introduced using the L_1-norm of the archetype matrix ($\mathbf{A}$)[38].

$$\min\left(\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij}-(\boldsymbol{LA})_{ij}\right)^2+\alpha\left|A_{i,j}\right|\right)$$
$$s.t.\ \ \|\boldsymbol{L_s}\|_2=1\ \forall\ 0\leq\ s<k \tag{23}$$

Alpha is the penalty parameter that controls the promotion of sparsity during the decomposition. The constraints are slightly different to the general problem described in eq.(7), but still correspond to a sparse PCA minimization. One of the curious effects that this implementation causes on the dataset, is that the zero-entries in the representation matrix (the images) for an EELS dataset are all resolved for the same dimension (matrix column or, equivalently, loadings image). It also enforces flat curves for the archetypes of these components (4 to 6). One improvement over the PCA counterpart is the easier interpretation of the factorized results. Archetype 0 is the main component for the oxygen edge and the background signal. Archetypes 1 to 3

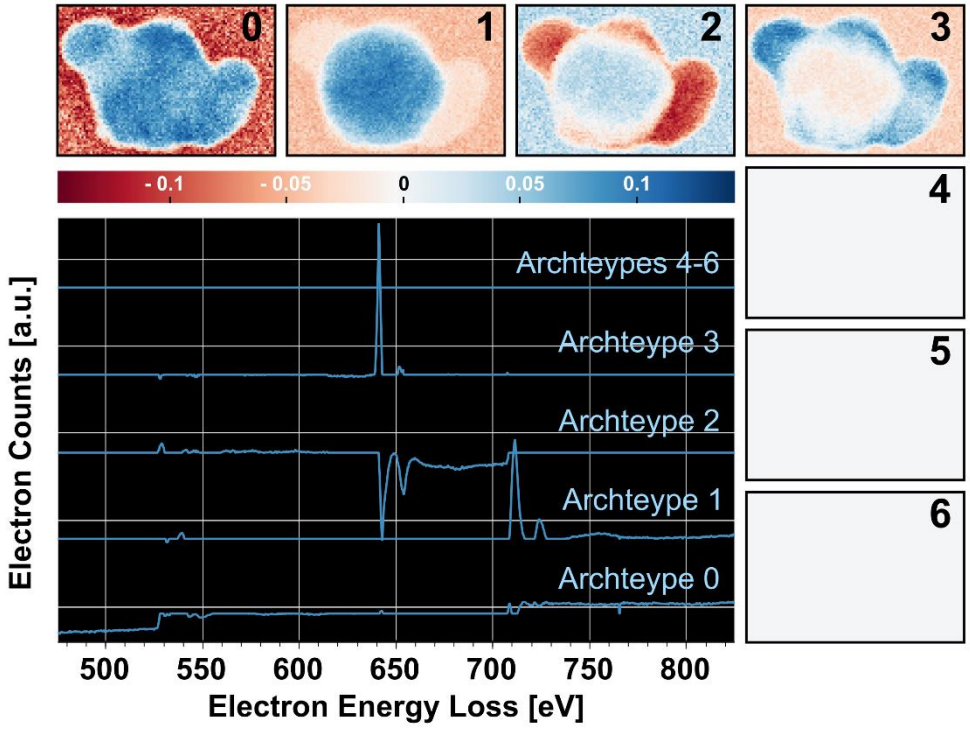


Figure 4.5. sPCA results for the Iron-Manganese oxide nanoparticle EELS SI (L_2-normalized). The colour scale is shared by all images. Notice that the images 4-6 are entirely composed by zeros. The archetypes or factor signals are not re-scaled, but they are displaced on the y-axis to fit in a single plotting area. The archetypes for the 4^{th} ,5^{th} , and 6^{th} components are the same curve, a flat one.

correspond to the iron white-lines (located mainly at the NP core) and the manganese white-lines (located mainly on the shell parts of the NP). The differences between 2 and 3 are a shift on the energy loss values, a change in the relative ratio of white-lines and some minor peaks at the oxygen edge. All these variations can easily be linked to the transition from Mn_3O_4 to MnO (change in the Mn oxidation state) from the innermost to the outermost areas of the shell, respectively[17].

Still, the archetypes presenting non-physical features, the lack of noisy components (and noise in the components themselves) and the piling of zeros on specific components, raise some concerns about the potential use of the decomposition results of sPCA for quantitative analysis, as some information appears to be missed from the original dataset.

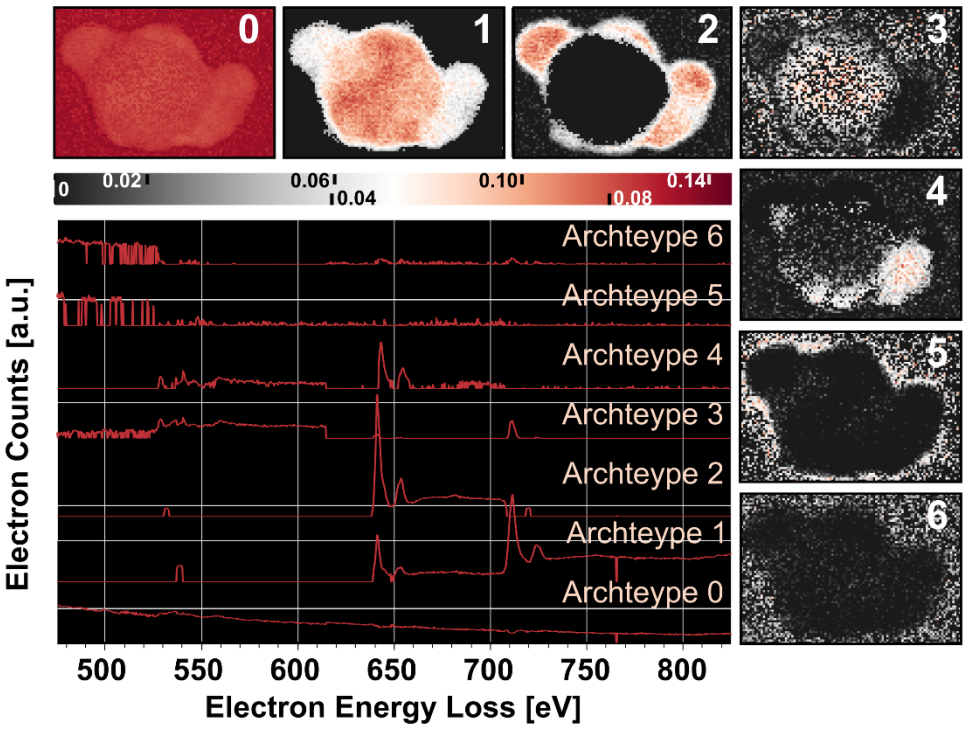


Figure 4.6. NMF results for the Iron-Manganese oxide nanoparticle EELS SI (L_2-normalized). The colour-bar shows two scales. The one on top (labels to the left of the ticks) is only for the component 0, and the one at the bottom (labels to the right of the ticks) is shared by the rest of the components resolved. The archetypes or factor signals are not re-scaled, but they are displaced on the y-axis to fit in a single plotting area.

The NMF decomposition is shown in **Figure 4.6**. Notice the dual scale in the colour-bar. In order to improve the contrast in the image representations a second scale (top numbers) is included exclusively for component 1, as its variance value is somewhat higher than the values retrieved for the rest of the components.

The algorithm from sklearn includes the option of extra regularization terms on top of the general formulation of eq.(6), but are all set to 0 in this case. The KL distance was selected for the loss function, to coherently include the Poissonian statistical description of NMF explained in eq.(22), together with the statistical normal distribution for PCA. As the acquired dataset was dark-corrected, some fictitious negative counts were present in the EELS-SI. As a shortcut, a rescaling of the dataset was done previous to the L_2-normalization,

to set the minimum spectral counts at 0 (a translation in the y-axis). Again, non-physical features are observed all along the archetypes resolved, but interestingly enough several areas on the images are now set to 0. The nonnegative double singular value decomposition[39] (NNDSVD) initialization (by default) promotes sparsity, transferred to the resolved representation matrix (loadings). The interpretability of the factorization is easier now than in the case of the sPCA (**Figure 4.5**). Some spectral regions are clearly separated, such as the Mn oxide shell from component 2, and the presence of a mixed oxidation state from the WL shift and change in ratio in component 4. Nevertheless, from component 1 one may interpret that iron oxide is present across the whole NP volume, but the non-physical drop in archetype-2 may be an attempt to compensate for this issue.

Finally, the results for a sklearn K-means clustering analysis are shown in **Figure 4.7**. This preview of a K-means analysis is included in the matrix factorization subsection because, as it was indicated in the theoretical introduction above, K-means is formally equivalent (eq.(8)) to the extreme case of enforcing a single archetype per datapoint in a sparse PCA decomposition.

The colour scale is different from the one displayed in figure 4.4,figure **4.5** and **figure 4.6**. Now, it indicates the fictional label assigned to each one of the clusters resolved. The actual representation matrix would be composed of ones (in the positions of pixels belonging to a given cluster) and zeros (for the pixels that do not belong to a specific cluster in a given column). The archetypes are the centroids of the clusters (i.e., the mean spectral values for the datapoints belonging to a given cluster).

One clear advantage of k-means is that all the centroids have physical meaning. Consequently, they are fitting candidates to become the endmembers of a MLLS fitting process (this is, assuming that one wants to carry out a quantitative analysis after the image segmentation).

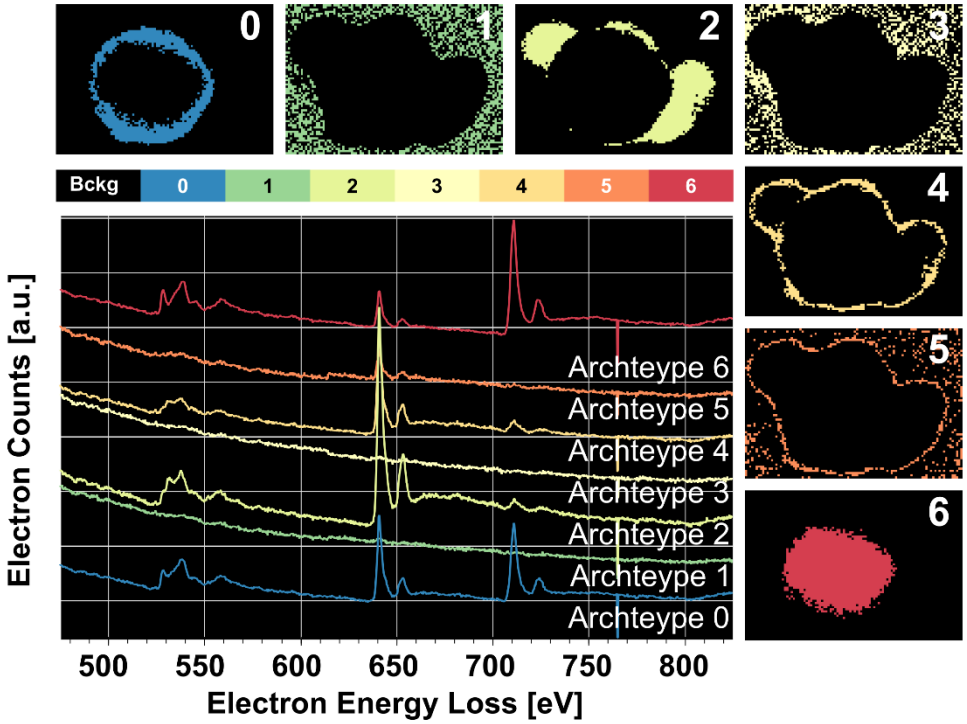


Figure 4.7. K-Means results for the Iron-Manganese oxide nanoparticle EELS SI (L_2-normalized). The colour scale for the images marks the cluster label number, and not the values for each component in the loadings matrix (which would be all 1). The archetypes or factor signals are not re-scaled, but they are displaced on the y-axis to fit in a single plotting area.

Conversely, the information conveyed in the representation matrix is minimized. Each pixel is represented by a single archetype now. This would, in principle, invalidate its integration in a tomographic reconstruction methodology, something that has been demonstrated possible from a simpler PCA decomposition[22]. Nevertheless, it would present an interesting case for an electron tomography experiment using DART[40].

In **Figure 4.7**, a clear pattern of random divisions of pixels for the background areas of the EELS SI (components 1 and 3) is visible. A concentric ordering of pixels in an 'onion-like' structure for components 4 and 5 is noticed as well. These are instabilities caused by noise and the uneven distribution of electron counts owed to the differences in thickness in the sample, respectively. These effects will be explored in depth in the clustering subsection.

To close the matrix factorization section, it is convenient to mention that the use of this dimensionality reduction family of algorithms has been relegated in most cases to a denoising pre-processing step for EELS data treatment[27–29]. Instead of selecting a reduced number of objective dimensions (the rank k of the new **Z** matrix), the procedure begins by setting a number of dimensions equal to the original number of electron channels.

After the decomposition, the components with a lower variance value are cut out at a threshold value. This is done under the assumption that precisely those lower variance components are mainly composed by noise information. This practice has been signalled as a source of possible bias in the literature (most notably in the PCA case), as fainter spectral features may be randomly redistributed over the noisy lower-variance components resolved[41–43].

4.3.2.Non-linear neighbour-graph approaches.

A second group of dimensionality reduction methods would be the non-linear neighbour-graph approaches. Their importance arises from the realization that linear methods for dimensionality reduction (e.g., a matrix factorization) do not always provide an appropriate solution when the underlying structure of the data lies in a curved manifold (of whichever dimension it may be).

These methods are all grouped together by the underlying core concept of reducing the dataset dimensionality by embedding a neighbours graph, constructed in the original higher dimensional space where the data resides, into a lower dimensional objective space.

Thus, every single one of these methods can be reduced to the following conceptually basic steps [49]:

i. **Construction of the k-neighbours graph in the higher dimensional space**

   The first step is to specify how the weights for the edges between the nodes of the graph (the datapoints) are constructed, in order to

preserve the structure of the dataset and to deal with the inherent asymmetry resulting from the construction a k-neighbours graph.

i. ***Projection of the graph on the objective lower dimensional space.***

Once the k-neighbours graph is constructed in the higher dimensional space, a lower dimensional representation is required to minimize a given objective function, preserving as much of the k-nearest graph information as possible.

The specifics of how these two steps are tackled are the actual basis for the formulation of different methods. It will also govern the computational costs for each algorithm.

Among the algorithms that can be listed as part of this family of DRM, only two have been tested in EELS datasets: t-distributed stochastic neighbour embedding[45] ***(t-SNE)*** and uniform manifold approximation and projection [33]***(UMAP).***

Although formally conceptualized as a visualization method for high dimensional datasets in 2D or 3D representations, t-SNE [30,31] is considered by many the state-of-the-art neighbours-graph based dimensionality reduction method.

It can be described in a single sentence as a symmetrized version of stochastic neighbour embedding (SNE)[46], where the mapping of the lower dimensional space is done by 1-degree of freedom Student t-distribution. A partially detailed description of the algorithm would go along the following lines.

In SNE, the k-neighbours graph of the original dataset (a vectorial space of datapoints)

$$\boldsymbol{X} = \left\{\vec{x}_1, \vec{x}_2, \dots, \vec{x}_i, \vec{x}_j, \dots, \vec{x}_N\right\} \tag{24}$$

is constructed by *similarity* and not by an actual hyperdimensional Euclidean metric. The concept of similarity is defined as the conditional probability of a given datapoint $\vec{x}_i$ setting $\vec{x}_j$ as its nearest neighbour, modelled by a Gaussian

probability distribution centred at each $\vec{x}_i$ with a variance σ_i (a sort of radial-basis-function (RBF), kernel with varying radius). The neighbouring measurement is then provided by the density of probability around $\vec{x}_i$. The formal equation for the conditional probability reads as

$$p_{j|i} = \begin{cases} \dfrac{\exp\left(-\left\|\vec{x}_i - \vec{x}_j\right\|^2 / 2 \cdot {\sigma_i}^2\right)}{\sum_{k \neq i} \exp\left(-\|\vec{x}_i - \vec{x}_k\|^2 / 2 \cdot {\sigma_i}^2\right)}, & \forall\, i \neq j \\ 0, \quad otherwise \end{cases} \tag{25}$$

Notice that the equation actually formulates a Gaussian pair-wise similarity constructed from the euclidean distances between two points $\vec{x}_i, \vec{x}_j$ (numerator), with a normalization term (denominator). Notice also that σ_i is variable for each $\vec{x}_i$. t-SNE (and SNE as well) computes σ_i via the so-called perplexity of the probability distribution

$$perplexity(P_i) = 2^{H(P_i)} = 2^{-\sum_j p_{j|i} \log_2 p_{j|i}} \tag{26}$$

where P_i is the conditional probability over all datapoints for a given $\vec{x}_i$, and $H(P_i)$ is the expression for Shannon entropy. The σ_i is computed by fixing the value of the perplexity as a hyperparameter in the algorithm (scikit-learn implementation).

Sometimes, the perplexity is described in plain words as a sort of measurement for the effective number of neighbours considered in the k-neighbours graph.

As $p_{j|i}$ in eq.(25) is asymmetric, t-SNE implements the following symmetrisation (and further normalization) to define the joint probability distribution[7]

$$p_{j,i} = \frac{p_{j|i} + p_{i|j}}{2N} \tag{27}$$

[7] Notice that the formulation for the joint probability distribution of t-SNE in eq.(27) does not follow the actual mathematical definition. Formally, the joint probability distribution $p_{b,a} = p_{a,b}$ relation with the conditional probability distribution $p_{b|a}$ (as the one in eq.(25)) is $p_{b,a} = p_{b|a} \cdot p_a$. Thus, t-SNE is using a heuristic definition for the joint probability distribution, and not a mathematically sound one.

which helps on the prevention of instabilities caused by outliers.

Considering now the low dimensional space (the mapping space)

$$\boldsymbol{Y} = \left\{\vec{y}_1, \vec{y}_2, \dots, \vec{y}_i, \vec{y}_j, \dots, \vec{y}_N\right\} \tag{28}$$

mapping the new positions by the same type of Gaussian distributions described in eq.(25) could induce a crowding problem, due to the potential mismatch between the actual dimensions of the original and projected data-spaces[30,47]. T-SNE addresses this problem by introducing a 1-degree-of-freedom Student distribution (i.e., a Cauchy distribution) in the joint probability distribution definition for the mapping space (i.e., the low dimensional projection space).

$$q_{i,j} = \begin{cases} \dfrac{\left(1 + \left\|\vec{y}_i - \vec{y}_j\right\|^2\right)^{-1}}{\sum_{k \neq l}(1 + \left\|\vec{y}_k - \vec{y}_l\right\|^2)^{-1}}, & \forall\, i \neq j \\ 0, \quad otherwise \end{cases} \tag{29}$$

Notice that the normalization term (denominator) is not expressed in terms of a specific index i (as in the element-wise normalization in eq.(25)), but instead indicates a matrix-wise normalization.

Finally, to find a solution t-SNE minimizes through a gradient descent (GD) method the Kullback-Leibler divergence (KL-div) between the two joint probability distributions (KL-div is commonly used as a score to measure similarity between probability distributions, and here plays the role of the cost function for the GD).

$$C = KL(P||Q) = \sum_i \sum_j p_{i,j} \log\left(\frac{p_{i,j}}{q_{i,j}}\right) \tag{30}$$

The change from the use of conditional probabilities (as the one described in eq. (25) for the SNE) to joint probabilities in t-SNE simplifies the computation of the gradient of the cost function.

Two optimizations (via hyperparameters) are included in the algorithm for an accurate and faster convergence. The *early compression* adds a L_2 penalization

term to the cost function that forces the mapped points ($\vec{y}_i$) to remain closer (low $q_{i,j}$) allowing a global restructuration of the possible clusters of data in the early stages of the optimization. The *early exaggeration* is a multiplier of the $p_{i,j}$ that, in the early stages of optimization, forces the mapping of large $p_{i,j}$ to large $q_{i,j}$ (global structure preservation).

Uniform manifold approximation and projection (UMAP) for dimension reduction is the direct competitor of ***t-SNE*** for the actual first position in the podium of the state-of-the-art k-neighbours graph approaches for dimensionality reduction[48].

Although formally classified in the same category as t-SNE, the construction of UMAP takes a completely different approach to the problem. The actual formulation justifies each of the steps in the hyperspatial-graph construction and the low dimensional layout with a mathematical topological background[49]. Nonetheless, in order to provide a coherent picture with regard to the t-SNE introduction above, a similar formulation is followed here to describe UMAP (Appendix-C at [49]).

The similarities in the high dimensional space, as conditional probabilities, are

$$p_{j|i} = \begin{cases} \mathrm{e}^{\frac{-d(\vec{x}_i,\vec{x}_j)-\rho_i}{\sigma_i}}, & \forall\, i,j \;\leq k \\ 0, & otherwise \end{cases} \tag{31}$$

Where $d(\vec{x}_i,\vec{x}_j)$ is the distance between two datapoints in the original hyperdimensional space $\boldsymbol{X}$. Notice that $d(\vec{x}_i,\vec{x}_j)$ is not constraint to be the Euclidean distance. Any valid distance formulation is allowed in principle. Also, only the k-nearest neighbours are considered. The parameters ρ_i (distance to the nearest neighbour of $\vec{x}_i$) and σ_i (normalization factor) are calculated as follows.

$$\rho_i = \mathrm{minimum}\big(d(\vec{x}_i,\vec{x}_j)\;\; s.t.\;\; 1 \leq j \;\leq k \; and \; d(\vec{x}_i,\vec{x}_j) > 0\big) \tag{32}$$

$$\sum_{j=1}^{k} \mathrm{e}^{\frac{-\max(0\,,\;\; d(\vec{x}_i,\vec{x}_j)-\rho_i)}{\sigma_i}} = \log_2 k \tag{33}$$

The symmetrisation of the probability is now

$$p_{i,j} = \left(p_{i|j} + p_{j|i}\right) - p_{j|i} \cdot p_{i|j} \tag{34}$$

with no further normalization required. Regarding the low dimensional embedding, the similarities by a symmetrized probability distribution are formulated as

$$q_{i,j} = \left(1 + a\left\|\vec{y}_i - \vec{y}_j\right\|^{2b}\right)^{-1} \tag{35}$$

where a and b are hyperparameters (positively defined), with optimum values given by the algorithm implementation as a = 1.929 and b = 0.7915 (notice that a, b = 1 recovers the Student distribution for $q_{i,j}$).

Finally, UMAP resolves the optimum mapping in the low dimensional embedding by a stochastic gradient descent methodology using as the cost function the cross entropy

$$C_{UMAP} = \sum_{i \neq j} p_{i,j} \log\left(\frac{p_{i,j}}{q_{i,j}}\right) + \left(1 - p_{i,j}\right) \log\left(\frac{1 - p_{i,j}}{1 - q_{i,j}}\right) \tag{36}$$

Although the formulation presented here for UMAP uses the same notation of t-SNE, and follows a similar pathway, the actual implementation of the algorithm is different. It begins with the construction of a directed weighted graph $\bar{G}(V, E, w)$, symmetrized to an undirected weighted graph $G(V, E, w')$, and projected to the low dimensional space by a force directed layout algorithm. The vertices of the $\bar{G}$ graph are the datapoints themselves, the edges are drawn between datapoints according to the local number of neighbours selected, and the actual vicinity of each datapoint defines the weight of those said edges (each datapoint draws a set of edges to the k-nearest neighbours directionally, setting a local metric space around each one of them, which is later ensembled in a single graph and symmetrised to get G). The repulsive and attractive forces applied in the low dimensional embedding are derived from the minimization of the cross entropy between both high and low dimensional representations of the dataset.[49]

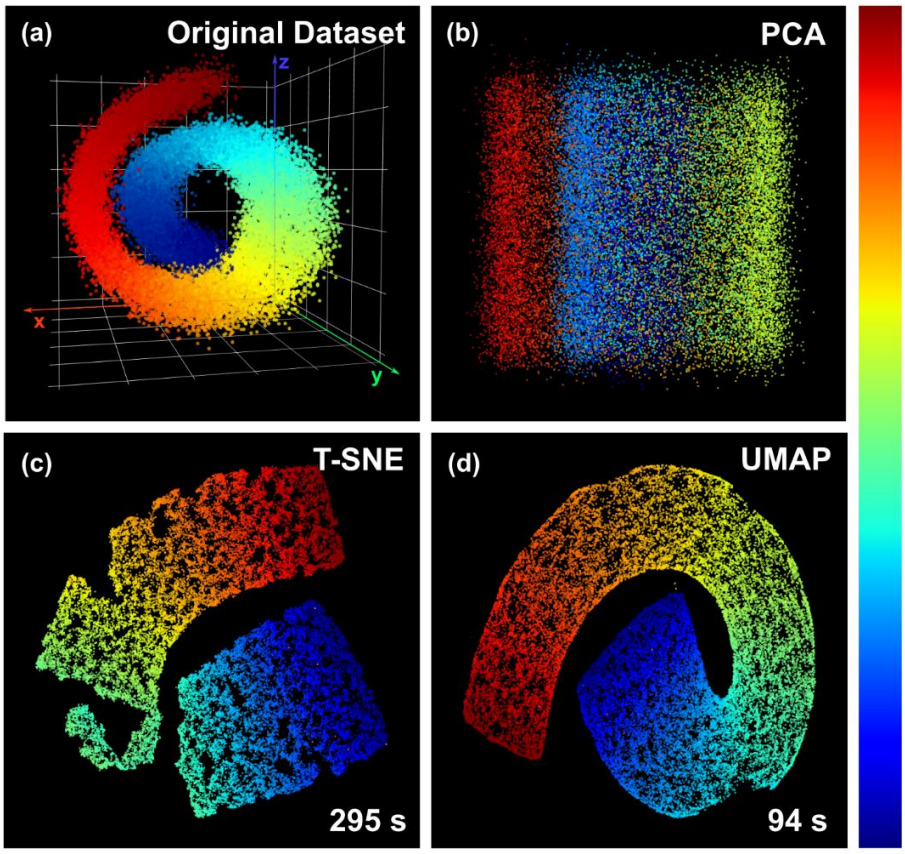


Figure 4.8. Dimensionality reduction of the 'Swiss-Roll' (a) 30000 points dataset (with noise) to a 2D embedding via (b) PCA, (c) t-SNE and (d) UMAP. The colour scale is shared by all 4 images.

As an example of the capabilities of t-SNE and UMAP, the 'Swiss-Roll' 3D-phantom dataset is projected into a 2D space, as in a dimensionality reduction problem. The results are shown in **Figure 4.8**. The 30000 points are originally organized in a sort of 3D spiral (a), and they are labelled with a colour according to their position (for a later identification in the projection embeddings). The underlying structure of the dataset is inherently non-linear (a curved 3D manifold), and thus is not a surprise to observe that the PCA decomposition (b), included to compare non-linear approaches and matrix factorization methods, results in a poor representation of the original dataset.

This 2D graph of the PCA decomposition (**Figure 4.8** (b)) is constructed from the representation matrix $\mathbf{L} \in \mathbf{R}^{30000\times 2}$ **.** Each datapoint is plotted into the $\mathbb{R}^2$ plane according to the $L_{i,j}$ loadings values (i.e., the new pointwise positional vector $\vec{l}_i$). The new vectorial basis for the principal components resolved can be read from the values of the archetype matrix.

$$\boldsymbol{A} = \begin{bmatrix} -0.7635 & 0.0099 & -0.6457 \\ -0.0408 & 0.9971 & 0.0635 \end{bmatrix} = \begin{bmatrix} \vec{a}_0 \\ \vec{a}_1 \end{bmatrix}$$

The archetype 0, or $\vec{a}_0$ (i.e., the new x axis) is close to the bisecting axis of the z = 0 and x = 0 planes, presenting an approximate 40º angle measured from the x axis. The archetype for component 1, or $\vec{a}_1$ (i.e., the new y axis) is almost identical to the original y axis (i.e., almost $\vec{a}_1 = (0,1,0)$)

The results for the t-SNE embedding are shown in **Figure 4.8** (c). The perplexity for t-SNE algorithm is set as 100 (way over the recommended values in the documentation[30]) and the early exaggeration is set as 12 (fairly large number to accelerate convergence times in the 30000 points dataset). The large perplexity value is justified by the expected absence of significant local structure (i.e., complete absence of clusters in the dataset, as it is a homogeneous manifold with induced Gaussian noise in the coordinates values). Thus, the importance of the global structure is exaggerated. Nevertheless, it is observable the tendency of t-SNE to clump datapoints together and even to break the manifold in several regions. Overall, no significant overlapping between the different colours occurs in the projected map.

The UMAP embedding is shown in **figure 4.8** (d). The structure of the manifold is clearly preserved in the 2D embedding. The number of k-neighbours (*n_neighbours* hyperparameter in the python implementation of the algorithm) was set as 100 (to mimic the perplexity in the t-SNE projection), and the minimum distance (*min_dist*) was set as 0.25. The latter, normalized between 0 and 1, is the main hyperparameter driving the repulsive forces in the 2D embedding between data clusters in the original space (lower values

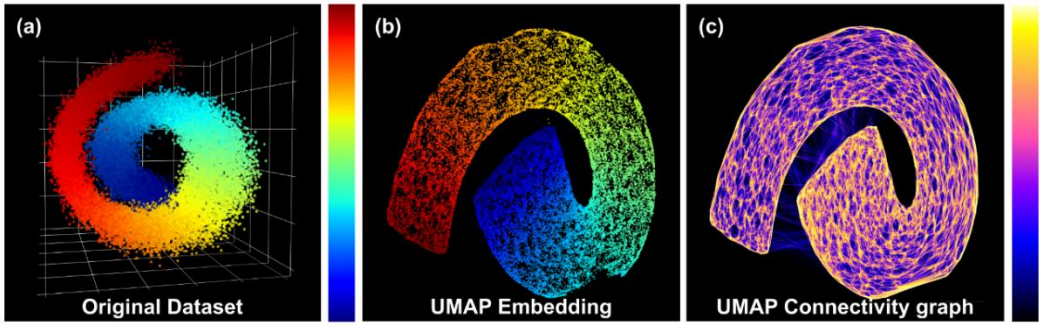


Figure 4.9. (a) Original 30000 points 'Swiss-Roll' dataset. (b) 2D embedding obtained via UMAP (100 k-neighbours and minimum distance of 0.25). (c) Connectivity graph for the embedding resolved in (b). Subfigures (a) and (b) share the same colour scale, 3D-positional indices, and (c) colour scale is identified with the weights of the symmetrized edges.

induce higher repulsion, and vice versa). In this example we already observe the ability of UMAP to preserve the global structure of the manifold. Also, in a limited 30000 points dataset UMAP is already significantly faster than t-SNE. The difference of convergence times between t-SNE and UMAP is incremented in favour of the latter as the dataset grows.

Although t-SNE and UMAP yield better results for the 2D embedding, the interpretability of the actual new mapping axes is not as straight forward as in the case of a matrix factorization dimensionality reduction. This has to be considered when dealing with real EELS datasets. Nevertheless, the non-parametric nature of the graph-algorithms implies that no assumption about the underlying distribution of probability is made along the dimensionality reduction process (data-driven dimensionality reduction). The translation of the local information about the probability density function (PDF) into a lower dimensional embedding, t might be exploited by a posterior density-based clustering analysis, as will be demonstrated later.

Deepening in the analysis of UMAP now, as the actual implementation of the algorithm is built upon the construction of a graph preserving the local and global topology of the dataset, the symmetrized edges of such higher dimension graph between datapoints (vertices) can be drawn into the lower dimensional embedding. This allows the user to assess if there is any actual

local structure in the dataset (bundles of edges and datapoints, representing local clusters) and visually evaluate the projection result. An example is shown in **Figure 4.9**. The connectivity graph in (c) shows a fairly homogeneous web, with no significant clumps of datapoints. Notice also the tenuous edges (blue coloured in (c)) between the blue and the red-orange map points in (b) (i.e., the innermost and outermost parts of the roll in (a), respectively). These connections between the outermost and the innermost regions are the driving forces behind the 'curl' in the embedding in (b) and (c).

As already mentioned, the interpretability of the resolved embedding axes in graph-based approaches is not as straight forward as in the case of a matrix factorization problem (i.e., extracting meaningful information about the components resolved, other than the actual local/global structure of the data as an abstract object, is difficult). Thus, the analysis of an EELS-SI dataset embedding cannot be completed (in terms of spectral analysis) without further applying complementary techniques, such as a clustering classification (on the low dimensional map resolved). Examples of experimental core-loss EELS-SI dataset embeddings (the core-shell iron-manganese oxide nanoparticle showcased in **Figure 4.1**) are shown in **Figure 4.10**. For the UMAP embedding (**Figure 4.10** (a), in red), the key hyperparameters were set as 45 k-neighbours and 0.35 d-min. For the t-SNE embedding (**Figure 4.10** (b), in green) the perplexity was set as 30 and the early exaggeration scaling value was set as 12 (default). Overall, UMAP presents a smoother result, owed to the higher degree of global structure preserved. On the other hand, t-SNE shows a higher tendency to aggregate datapoints together. It creates areas of clusters with a large granularity (see the yellow arrow in (b)), but it is unable to separate them effectively. This would in principle degrade a posterior density-based clustering analysis, as many dense regions resolved are mainly driven by spurious embedding artefacts (i.e., noise related clustering of datapoints).

Some extra cautionary notes are to be made before generalizing their use on EELS data analysis, aside from the interpretability problem.

t-SNE is not a generalized dimensionality reduction method. Its formulation allows the algorithm to project accurately the higher dimensional local structure into a 2D or 3D objective embedding dimension. However, the heavy tails of the Student PDF (i.e., the mapping statistics function) become a problem in higher objective dimensions, hindering the preservation of such local structure[30].

Furthermore, t-SNE tries to resolve a non-convex problem through the Kullback-Leibler divergence minimization. Hence, although well optimized and fairly robust to accurately describe the local structure of the data, different initializations varying slightly the hyperparameters may produce different results (i.e., reproducibility is not ensured, something of paramount importance in the analysis of scientific results)[30].

Regarding UMAP, its major weakness is that it will try to find the manifold structure of the dataset even when such structure might not exist at all (e.g., it may try to find structure within the noise of the dataset). More importantly, UMAP carries out some numerical calculations to speed up convergence[49]. These approximations might become relevant when the dataset analysed

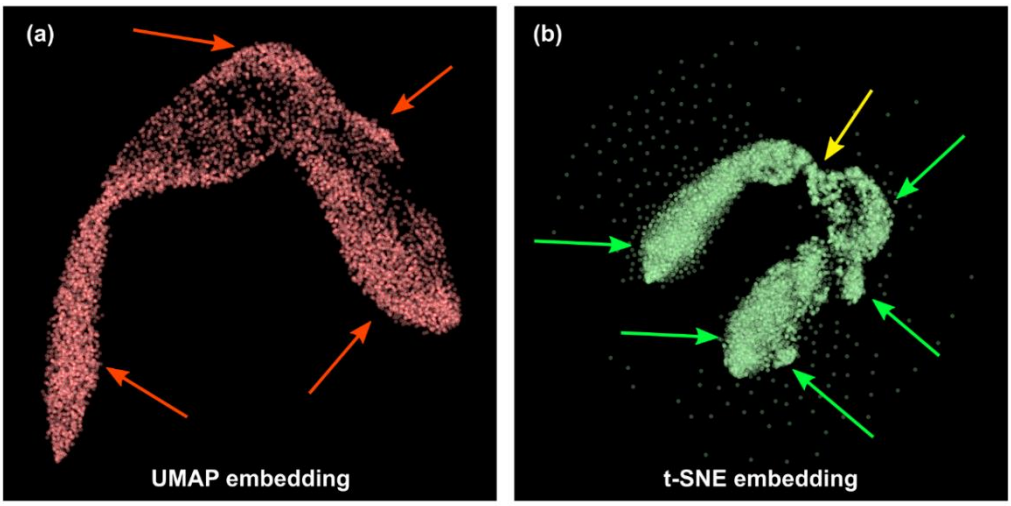


Figure 4.10. Non-linear graph embeddings of the iron-manganese oxide core-shell NP. (a) UMAP 2D embedding (red). The arrows point at regions with an apparent higher density of datapoints. (b) t-SNE 2D embedding (green). The green arrows point at regions with an apparent higher density of datapoints, and the yellow one points at a clumped region of several clusters.

contains a low number of datapoints, as the quality of the embeddings is greatly affected by the size of the data space.

4.4. Clustering analysis.

4.4.1. General formulation.

Formally, clustering (i.e., cluster analysis) can be described as the process to subdivide a certain dataset into a given number of subsets[9], aiming to achieve a higher homogeneity within them with respect to the properties utilized for the classification. The original dataset is viewed, thus, as a set of instances or objects of any given nature (e.g., a set of individuals taking part in a medical study, or in the case that will occupy us later, a collection of EELS spectra belonging to an acquired spectrum image), and the result of applying clustering analysis is an aggrupation of those instances in 'clusters' by similarity. The problem would be described formally as follows[50]:

An initial collection S of N number of datapoints (or vectors, in vectorial spaces)

$$S = \{\vec{\mathrm{x}}_i\}_{i=1,\dots,N} \qquad s.t. \qquad \vec{\mathrm{x}}_i = \left(x_{i1}, x_{i2}, \dots, x_{ip}\right) \tag{37}$$

where $\left(x_{i1}, x_{i2}, \dots, x_{ip}\right)$ are the p-dimensional data points (i.e., p characteristics defining each point); is subdivided into a collection C of clusters

$$C = \left\{C_j\right\}_{j=1,\dots,k} \quad s.t. \tag{38}$$

$$S = \bigcup_{j=1}^{k} C_j \tag{38.1}$$

$$C_i \cap C_j = \emptyset \quad \forall C_{i,j} \in C \tag{38.2}$$

Being each C_j the subset of data points denominated cluster, and C the overall clustering structure resolved by the cluster analysis, eq.(38). The original data space is recovered by the union of all the resolved clusters, eq.(38.1), and any object from the initial S dataspace will then belong exactly to one cluster or subset, eq. (38.2).

4.4.2. Distance measurements and metrics.

To separate the different clusters, some sort of measurement is required to distinguish between similar and dissimilar objects. This measurement is often presented as a pairwise distance $d(\vec{x}_i, \vec{x}_j)$ between any two given datapoints $\vec{x}_i$, $\vec{x}_j$ from S, and with respect to every other single object in S (i.e., a datapoint $\vec{x}_i$ will have *N-1* measured distances to the rest of $\vec{x}_{j\neq i} \in S$). For a given formulation of similarity to be considered as a valid distance measurement (for clustering), it is required to be symmetric (i.e., $d(\vec{x}_i, \vec{x}_j) = d(\vec{x}_j, \vec{x}_i)$) and to assign the minimum values to the case of identical datapoints (i.e., $d(\vec{x}_i, \vec{x}_j) = \min(d(\vec{x}_i, \vec{x}_j)\ ;\ \forall\ i, j \in \{1, .., N\}) \rightarrow \vec{x}_i = \vec{x}_j$). This distance measurements are commonly referred as metrics[8], although to actually conform a mathematically formal metric two extra requirements must be fulfilled: (1) the triangle inequality (i.e., $d(\vec{x}_i, \vec{x}_k) \leq d(\vec{x}_i, \vec{x}_j) + d(\vec{x}_j, \vec{x}_k)\ \forall\ \vec{x}_i, \vec{x}_j, \vec{x}_k \in S$), and (2) the minimum distance 0 only applies to equal datapoints (i.e., $d(\vec{x}_i, \vec{x}_j) = 0 \rightarrow \vec{x}_i = \vec{x}_j\ \ \forall \vec{x}_i, \vec{x}_j \in S$).

The most common distance measurement is the Euclidean distance, followed by the Manhattan distance (or taxicab-distance). Both are special cases of Minkowski metrics, formulated for two p-dimensional $\vec{x}_i = (x_{i_1}, x_{i_2}, \dots, x_{i_p})$ and $\vec{x}_j = (x_{j_1}, x_{j_2}, \dots, x_{j_p})$ vectors as follows.

$$d(\vec{x}_i, \vec{x}_j) = \left(\sum_{k=1}^{p} \left| x_{i_k} - x_{j_k} \right|^g \right)^{\frac{1}{g}} \tag{39}$$

$$d(\vec{x}_i, \vec{x}_j) = \left(\sum_{k=1}^{p} \left| x_{i_k} - x_{j_k} \right|^2 \right)^{\frac{1}{2}} \tag{39.1}$$

[8] In many implementations of clustering algorithms (such as the sklearn library in Python), a hyperparameter denominated metric is passed along the dataset to be clustered, to specify the distance measurement method, even if such measurement method does not constitute an actual formal metric.

$$d(\vec{x}_i, \vec{x}_j) = \left(\sum_{k=1}^{p} \left| x_{i_k} - x_{j_k} \right| \right) \tag{39.2}$$

Being eq.(39),(39.1) and (39.2) the general Minkowski (g), Euclidean (g = 2) and Manhattan (g =1) metrics respectively.

Other measurement of distance commonly used for clustering of vectorial spaces is based on the cosine similarity (e.g., classification texts by word counting[51])

$$s(\vec{x}_i, \vec{x}_j) = \frac{(\vec{x}_i \cdot \vec{x}_j)}{\|\vec{x}_i\| \, \|\vec{x}_j\|} \tag{40}$$

where $\vec{x}_i \cdot \vec{x}_j$ is the inner product of vectors. Thus, the similarity function measures the cosine of the angle between those datapoints in a p-dimensional space. In order to qualify as a valid distance measurement, the actual cosine 'metric'[9] is expressed as $1 - s(\vec{x}_i, \vec{x}_j)$. Hence, (1) for equal vectors the distance measures 0 (i.e., the angle between them is 0, and the cosine is 0), (2) for opposing vectors the distance measures 2 (i.e., datapoints with an angle of π rad and -1 cosine), and (3) for the rest of the possible angles the value is between 0 and 2, monotonically increasing from 0 to π rad. The cosine 'metric' is presented here because of its ability to cluster vector spaces disregarding the scale of such vectors (the modules). Although not exactly a formal metric, the cosine distance can be approximated by applying a pre-normalization (L_2) step on the data and then measuring the standard p-dimensional Euclidean distance (i.e., using a formal metric as a proxy for the cosine distance). Indeed, if we consider L_2-normalized vectors

$$\begin{aligned} \hat{x}_i = \frac{\vec{x}_i}{\|\vec{x}_i\|} \quad &\rightarrow \quad \|\hat{x}_i\|^2 = \hat{x}_i \cdot \hat{x}_i = 1 \\ \hat{x}_j = \frac{\vec{x}_j}{\|\vec{x}_j\|} \quad &\rightarrow \quad \|\hat{x}_j\|^2 = \hat{x}_j \cdot \hat{x}_j = 1 \end{aligned} \tag{41}$$

[9] The cosine 'metric' does not fulfil the triangle inequality and, thus, it is not a formal metric. Nonetheless, it is included as a 'metric' in several implementations of the algorithms.

the squared euclidean distance between them becomes

$$\begin{aligned} d^2(\hat{x}_i, \hat{x}_j) &= \left\|\hat{x}_i - \hat{x}_j\right\|_2^2 = (\hat{x}_i - \hat{x}_j) \cdot (\hat{x}_i - \hat{x}_j) \\ &= \hat{x}_i \cdot \hat{x}_i - 2\hat{x}_i \cdot \hat{x}_j + \hat{x}_j \cdot \hat{x}_j \\ &= 2 - 2\hat{x}_i \cdot \hat{x}_j \\ &= 2 - 2\cos\angle(\hat{x}_i, \hat{x}_j) \end{aligned} \tag{42}$$

As the algorithms of clustering commonly use the euclidean distance, the euclidean distance between normalized vectors is

$$\begin{aligned} d_{Euc}(\hat{x}_i, \hat{x}_j) &= \left\|\hat{x}_i - \hat{x}_j\right\| \\ &= (2)^{1/2} \left(1 - 1\cos\angle(\hat{x}_i, \hat{x}_j)\right)^{1/2} \end{aligned} \tag{43}$$

and thus, the measurements by the cosine distance are proportional to the ones of the euclidean distance of the normalized dataset.

In EELS SI data-analysis, it is often the case that in certain localized regions of a SI the scale of the spectra (i.e., the L_2-norm) is much larger than in other spatial regions. The possible causes for such variations of the modules can be several, most notably thickness variations in the sample[52,53]. To

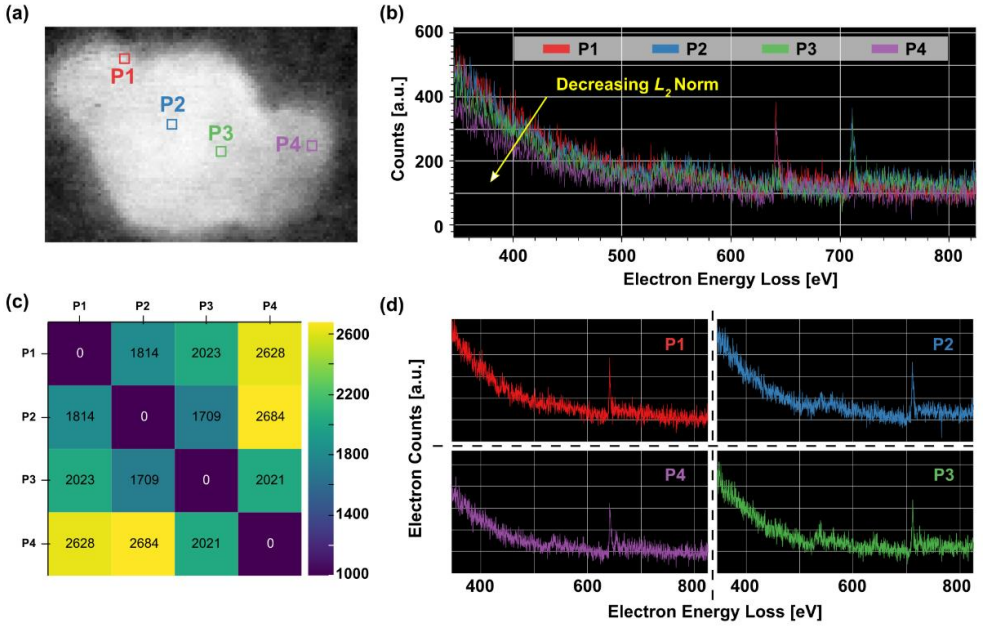


Figure 4.11. (a) Spectrum image of an iron-manganese oxide core-shell NP. 4 pixels are highlighted as $P_{1,2,3,4}$. (b) Overlay of the spectra corresponding to $P_{1,2,3,4}$. (c) Pairwise distance table for the euclidean distance between the spectra of $P_{1,2,3,4}$. (d) Spectra of $P_{1,2,3,4}$ separated in different plotting areas.

understand how the different metrics may affect a clustering classification, **Figure 4.11** and **Figure 4.12** show the results of calculating the pairwise distances for 4 different spectra of the iron-manganese oxide core-shell NP (already shown in the previous section) using the raw and L_2-normalized data. The locations of these datapoints are highlighted in the SI by squares (**Figure 4.11** (a) and **Figure 4.12** (a)). The 4 spectra are chosen by pairs: P_1 and P_4 correspond to areas of the shell (manganese oxide) with different thicknesses, and P_2 and P_3 correspond to areas of the core (iron oxide). The L_2-norms and electron count sum per spectra values for these pixels are shown in **Table 4.2**.

Beginning with the non-normalized (raw) data, **Figure 4.11** (b) shows an overlay of the spectra for the $P_{1,2,3,4}$ pixels. From it, it is clearly visible how the decline of the L_2-norm and sum of electron counts from P_1 to P_4 (following the cardinal order, see **Table 4.2**) is mainly linked to the differences in the downslope trend of electron counts preceding the oxygen edge at 532 eV of energy loss. For instance, in the thinner area where P_4 (purple) is measured, a lower number of electron-counts from the low-loss region of the spectrum, and ascribed to sample thickness affecting the plasmon excitation, are added to the core loss spectrum. Conversely, P_1 (red) is measured in a thicker region of the shell, and both the low-loss contribution to the background signal and an increment in the WL signal intensity are visible in the spectrum.

Table 4.2. Values of the L_2-norms for the spectra of the pixels highlighted as $P_{1,2,3,4}$ in **Figure 4.11** and **Figure 4.12**.

	P₁	***P₂***	***P₃***	***P₄***
L_2-*norm*	9193.73	9054.99	8171.31	7013.88
Sum	3.48e5	3.48e5	3.14e5	2.70e5

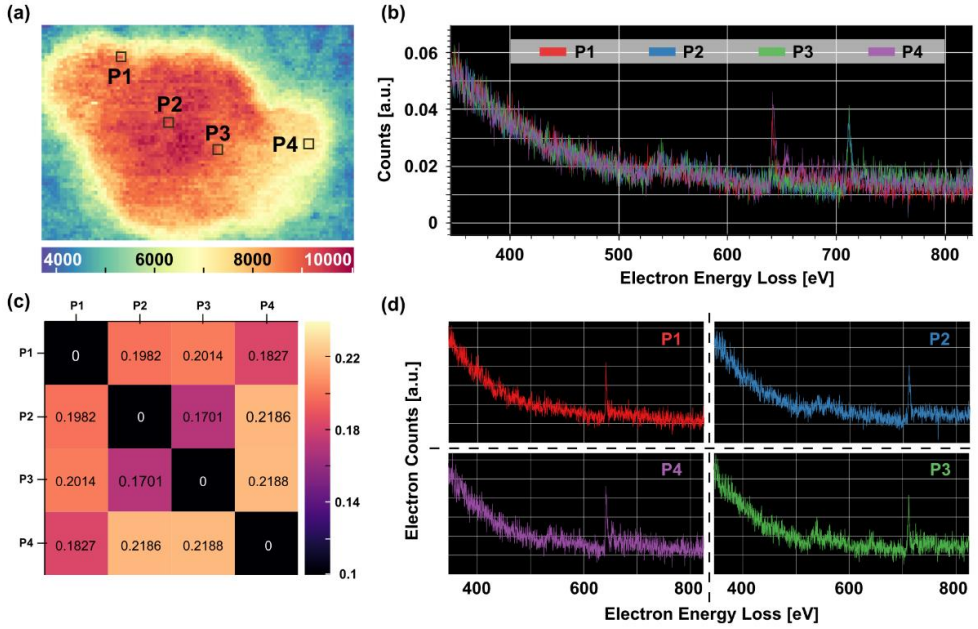


Figure 4.12. (a) L_2-norm image for each pixel in the SI of iron-manganese oxide core-shell NP. 4 pixels are highlighted as P1,2,3,4. (b) Overlay of the l_2-normalized spectra corresponding to $P_{1,2,3,4}$. (c) Pairwise distance table for the euclidean distance between the normalized spectra of $P_{1,2,3,4}$. (d) l_2-normalized spectra of $P_{1,2,3,4}$ separated in different plotting areas (identical coordinates axes)

Figure 4.11 (c) shows a 'table of sorts' with the pairwise distance between the selected datapoints organized in a matrix. Although $P_{1,4}$ share the same ELNES spectral features (manganese white lines at approximately 640 eV of energy loss, see **Figure 4.11** (d)), the euclidean distance measured between them has the higher value. In fact, P_1 is almost as close (distance wise) to P_2 as P_3 is to P_2 (sharing the latter two the same spectral features of iron WLs at approximately 710 eV of energy loss, see **Figure 4.11** (d)). Under the right conditions, in a clustering experiment this could mean a grouping of $P_{1,2,3}$ together, even if elementwise P_1 and P_4 should belong to the same cluster separated from P_2 and P_3[10].

[10] Or, at least, it will contribute negatively to the clustering analysis for spectral segmentation if thickness is not of particular interest.

Figure 4.12 has the same structure of **Figure 4.11**, but this time showing the L_2-normalized spectra. The SI area in (a) is coloured according to the L_2-norm values of the spectra before the actual normalization. The effects of the pre-normalization are clear, as all now the spectra share a similar background signal (albeit the random noise fluctuations, see **Figure 4.12** (b)). The heatmap (**Figure 4.12** (c)) contains the pairwise euclidean distances measured for the normalized spectra. Now it shows the expected structure, as the lower values of measured distances are for the P_2- P_3 and P_1- P_4 (i.e., the spectra from areas sharing the same elemental composition, thus, with similar spectral features).

Comparing now the euclidean distances, on the raw and normalized dataset, and the cosine distance on the raw dataset (see **Figure 4.13**), one can observe the parallelism between the latter two (as already explained by eq.(42) and (43)), although the actual values are an order of magnitude apart. From **Figure 4.11**, **Figure 4.12** and **Figure 4.13**, the differences in the ranking of pairwise distances between the same EELS datapoints (spectra) utilizing different formulations, as well as the effects that different scales in the spectra in an EELS SI may have over the intended segmentation, are evident. Hence the importance of the metric selection (or the L_2-normalization as a proxy for non-metric distance measures) for a satisfactory clustering analysis.

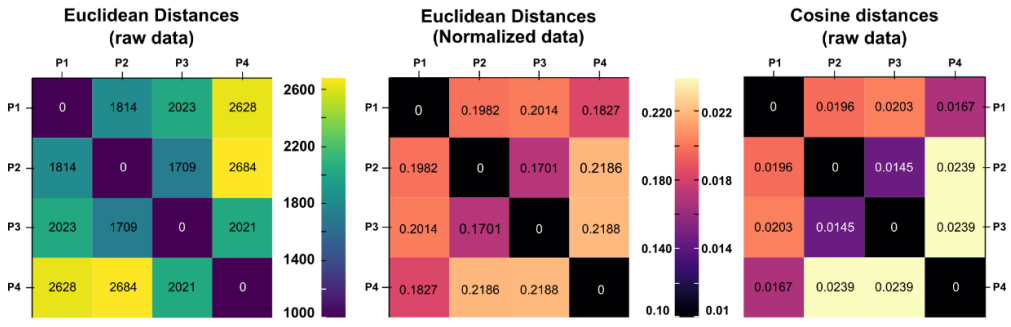


Figure 4.13. Tables of pairwise distances for the points $P_{1,2,3,4}$ presented in **Figure 4.11** and **Figure 4.12**. The tables of euclidean distances of L_2-normalized spectra and cosine distances for the raw spectra share the same colorbar. Nonetheless, notice the change of scale between them (euclidean distances for normalized data are an order of magnitude larger than the cosine distances for the raw data).

4.4.3. Taxonomical classification of clustering algorithms.

Given the general formulation of a clustering problem, and having chosen an appropriate distance measuring method, each clustering algorithm will encode an evaluation criterium (i.e., the specific formulation of each clustering algorithm will determine how the selected metric is applied to resolve the partitioning)[9,50]. Generally speaking, this selection criterium is what is often used for the classification of the different cluster algorithms into different subcategories. This sort of taxonomical classification is useful, as it allows any user to quickly select the clustering algorithm that better suits the dataset under analysis. However, as many given evaluation criteria may have features from several different categories, every classification of clustering algorithms is often subjected to discussion[9,50,54] (e.g., a common classification differentiates between partitioning, hierarchical and density-based algorithms and, thus, a density-based hierarchical algorithm, such as OPTICS[55], would fall into two different categories). Moreover, having accepted as valid the general concept of cluster and clustering analysis discussed above, an actual specific definition of the notion of cluster is still debated[50,56]. This contributes even further to the proliferation of different algorithms (with wild variations of the induction principle for the segmentation) and, thus, increasing the difficulty of classifying them into separated categories.

In this work the clustering algorithms (which will be applied to EELS data analysis) are classified in 4 basic classes, arranged in a 2D grid of natural properties (see **Table 4.3**).

The algorithms will be classified in one axis as parametric or density-based. An algorithm will be deemed parametric when the formulation of the evaluation criteria is such that assumptions about the shape of the clusters, in whichever p-dimensional space the data may live, are implicitly or explicitly made (i.e.,

Table 4.3. Classification of clustering algorithms, listing the paradigmatic methods for each class.

	Single partitioning	***Hierarchical***
Parametric	K-Means	Agglomerative (ward)
Density-based	DBSCAN	OPTICS, HDBSCAN

the underlying probability distribution is presupposed). Conversely, the ultimate objective of a density-based clustering algorithm is to resolve a probability density function (PDF) that models the underlying probability distribution from which the datapoints are drawn (usually, a mixture of several distributions)[50,57].

On the second axis, the algorithms are classified as hierarchical or single partitioning. The hierarchical methods resolve the clustering by recursively dividing datapoints in categories. Hierarchical agglomerative (HA) clustering follows a bottom-up approach, beginning from a cluster for each datapoint and merging them together at each iteration given a similarity measure. Hierarchical divisive is the complementary top-down approach (separating clusters from single cluster at the beginning). Single partitioning clustering algorithms work using a single reference value (a 'flat' cut, usually provided as a hyperparameter in the implementation of the algorithm) as the cluster separation criteria (e.g., K-means requires the objective number of clusters to be provided, and will always retrieve a number of clusters equal to the value provided).

K-means

The prototypical clustering algorithm and first to be described in any basic text of data-science (probably the simplest of all clustering algorithms as well) has

to be K-Means (KM). It is the most relevant representative of the intersection of single partitioning and the parametric clustering families (see **Table 4.3**). KM is classically built upon the sum of squares errors and resolved as an error minimization problem

$$\min\left(\sum_{k=1}^{K}\sum_{\forall\vec{x}_i\in C_k}^{N_K}\|\vec{x}_i-\vec{\mu}_k\|^2\right) \tag{44}$$

where N_k is the number of elements of cluster K and $\vec{\mu}_k$ is the average value (property-wise) of the elements $\vec{x}_i \in C_K$ (cluster K)

$$\vec{\mu}_k=\frac{1}{N_k}\sum_{j=1}^{N_k}\vec{x}_j \quad \forall\vec{x}_j\in C_k \tag{45}$$

Once described the general problem to be solved, it is worth noticing that the actual implementation of the sklearn KM algorithm (the one used throughout all this work) is based upon the so-called Lloyd's algorithm[58,59].

It was suggested in the dimensionality reduction subsection that a sparse matrix factorization problem that imposes a single archetype per datapoint (i.e., sPCA carried to the extreme) it is formally equivalent to a KM clustering. Indeed, recalling the general formulation for a matrix factorization problem (eq. (3)) and according to the schematics of **Figure 4.1**, each datapoint $\vec{x}_i$ is approximated by the low rank model

$$\vec{x}_i=\sum_{j=1}^{k}L_{i,j}\vec{a}_j \tag{46}$$

And therefore, from (eq. (8)), the $\|\mathbf{L}\|_0=1$ condition equivales to

$$\forall\, i=1,\dots,m \mid L_{i,k}\neq 0 \rightarrow L_{i,j}=0\;\forall\, j\neq k \tag{47}$$

Hence

$$\vec{x}_i \sim L_{i,k}\vec{a}_k \tag{46}$$

The equivalence with the KM formulation is then straightforward, as both are problems minimizing the sum of squares errors and imposing the $\|\mathbf{L}\|_0 = 1$ condition on the representation matrix and $\|\mathbf{L}\|_2 = 1$ (indirectly, via the average value condition for the centroids of eq. (45)). Hence, from the eq.(44) and (45), the representation matrix $\mathbf{L}$ and archetypes matrix $\mathbf{A}$ would be such that

$$\forall \vec{x}_q \in C_k \quad L_{q,k} = 1 \quad L_{q,j\neq k} = 0 \tag{47}$$

$$\forall\, \vec{a}_k = \sum_{j=1}^{m} A_{k,j} \;\mid\; \vec{a}_k = \vec{\mu}_k \tag{48}$$

and at each iteration of the minimization problem, changes in the $\mathbf{L}$ matrix are equivalent to changes of cluster membership for the datapoints, and changes in the $\mathbf{A}$ matrix are equivalent to the computation of the new archetypes by getting the new centroids (average value $\vec{\mu}_k$).

Linking KM to sPCA has the collateral effect of hinting the actual parametric nature of the algorithm. As a matter of fact, KM is assuming that the datapoints are drawn from a hyperdimensional Gaussian distribution of probability, centred around some hyperdimensional $\vec{\mu}_k$ centroids, imposing isotropic clusters (similar in sizes and shape). This type of parametric imposition will be denominated from now on as the Gaussian-ball assumption. A side effect of the k-means parametrization (deepening on the actual meaning) is that, by using the euclidean distance as measurement and expecting isotropic clusters, it becomes very sensitive to the scale of the dataset properties[60].

In the case of an EELS analysis, it translates to an imbalanced weight for certain spectral features, such as the background signal (often the dominant source of electron counts in the core-loss spectrum region). A good proxy to approximate the expected segmentation from a k-means clustering in a non-normalized dataset, is the analysis of the spectral L_2-norm (in a vectorial space for the spectra, the L_2-norm becomes the modulus, i.e., the scale). For instance, **Figure 4.14** shows the dominance that the L_2-norm has over the resolved

clustering for the non-normalized dataset. The contours-levels calculated from the L_2-norm image in (a) are almost perfectly translated into the clustering segmentation map (or label-map) in (b). This contour levels are generated by simply dividing the range of values for the L_2-norm by 5 equidistant cuts. From the centroid signals (average value of spectra for each cluster) in (d) one can observe the weight that the spectral background has on the L_2-norm, as the areas end up separated by thickness and not by the actual position of spectral features (e.g., the red cluster, number 4, contains both areas from the core and shell of the nanoparticle, being the latter only composed by manganese oxide and the former by iron oxide)[17]. Notice that the clusters label numbers are assigned at random (c), and do not follow any actual pattern related with the L_2-norm or any other spectral characteristic.

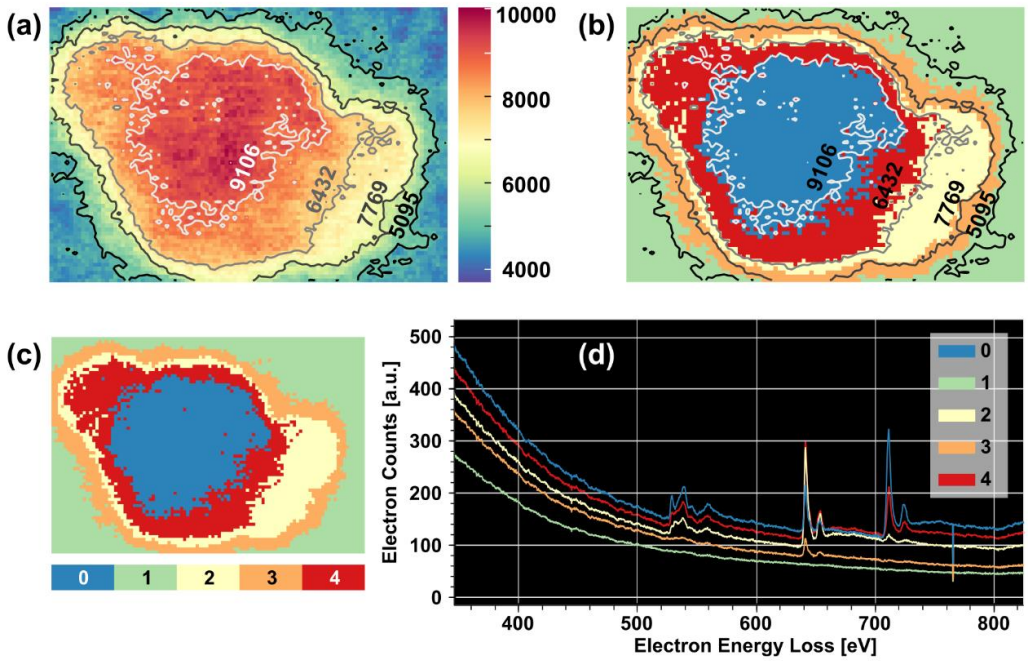


Figure 4.14. (a) 5-fold contour levels from the L2-norm values of the spectra from the EELS SI of the manganese-iron oxide NP, overlapped with the actual L2-Norm colourmap. (b-c) Label map resolved for a 5-clusters k-means segmentation of the raw data for the Mn-Fe oxide NP, overlapped with the same contour levels from (a) in (b). (d) Centroids (average signals) per cluster resolved.

After observing the weight of the background signal on the clustering segmentation, one may be tempted to perform a background removal procedure before the actual clustering analysis. An example of this practice is shown in **Figure 4.15**. The L_2-norm mapping (**Figure 4.15** (a)) for the background-subtracted SI does no-longer present a smooth gradient from the thicker part of the NP to the carbon grid, and it is also much noisier than the raw-data counterpart from **Figure 4.14** (a), owed to the lower signal-to-noise (Q) ratio. Precisely, the granularity of values in the L_2-norm translates into a higher difficulty for the interpretation of the contour levels in **Figure 4.15** (b). The 5 k-means clustering separates the background, core and shell regions successfully (notice that cluster 0, in blue, only shows the manganese white-lines). However, the core (clusters 1, green, and 3, orange) and the carbon background (clusters 2, yellow, and 4, red) regions are subdivided in two different clusters each, and from the overlay of contour-level and clustering

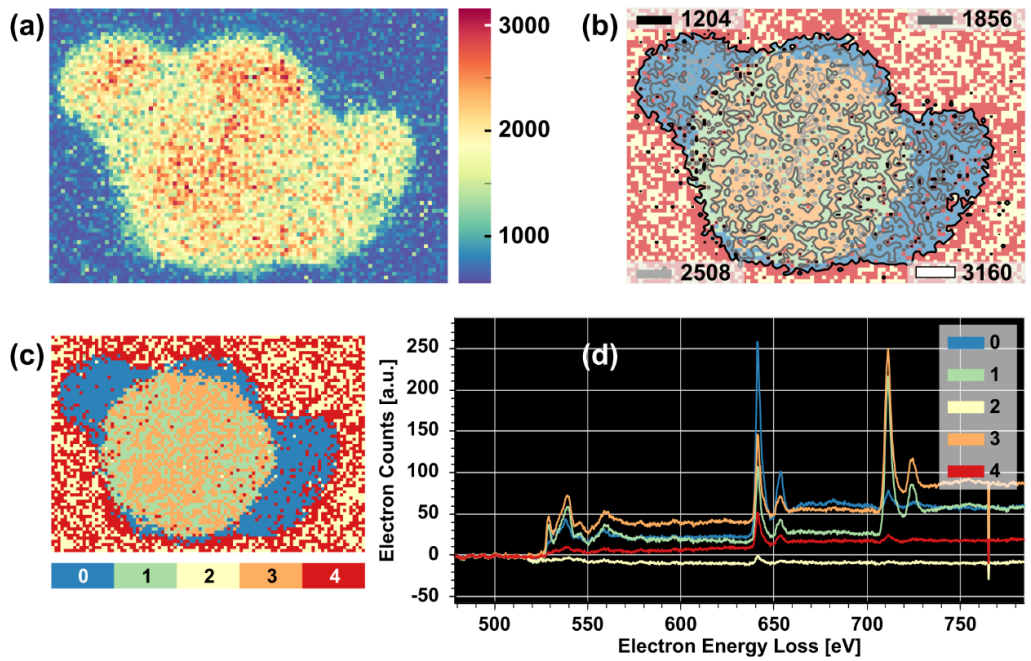


Figure 4.15. (a) L_2-norm values colormap from the spectra of the EELS SI of manganese-iron oxide NP, without background previous to the oxygen edge. (b-c) Label map resolved for a 5-clusters k-means segmentation of the no-background raw data for the Mn-Fe oxide NP, overlapped in (b) with the 5-levels contours resolved for the L_2-norm image in (a). (d) Centroids (average signals) per cluster resolved.

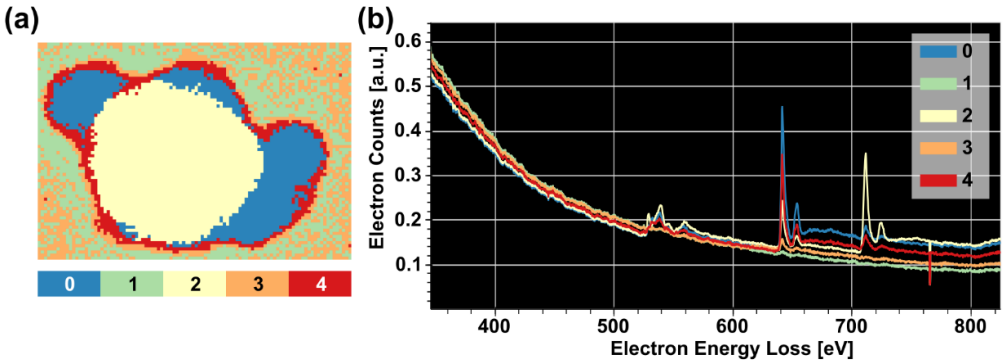


Figure 4.16. (a) Label map resolved by a 5-clusters k-means segmentation for the L_2-normalized data of EELS SI of the Mn-Fe oxide nanoparticle. (b) Centroid values for the normalized dataset (per cluster).

labels in **Figure 4.15** (b) it is easy to observe the direct connection with the L_2-norm once again. Furthermore, parts of cluster 4, in red and predominantly located in the carbon background areas of the SI, are visible as inclusions in the shell and core areas. The contour plot in **Figure 4.15** (b) reveals that those pixels present in their majority an L_2-norm lower than 1204 (black line) (i.e., lower norm than the surrounding pixels, and closer to the background ones).

Besides, this type of approach is also problematic in the sense that retrieving a background-removed spectrum image from a noisy dataset is oftentimes challenging. Notice how in cluster 2 that the average spectral signal (i.e., the centroid 2 from **Figure 4.15** (d)) shows a drop below 0 electron counts, a clear non-physical feature. This is a consequence of the overestimation of the number of electron counts when calculating the power-law parameters, mainly caused by noise fluctuations in the spectral region used for the background computation.

The use of a cosine metric is a better approach to the scaling problem , or, as demonstrated by **Figure 4.11**, **Figure 4.12** and **Figure 4.13**, the equivalent method of applying a L_2-normalization before the clustering process. These results are shown in **Figure 4.16**. From **Figure 4.16** (a) and **Figure 4.16** (b), the clustering clearly resolves the core-shell structure, although the imposition of 5 natural clusters in k-means still subdivides the background

signal in two different clusters. This division is probably driven by the residual noise fluctuations, but it does not affect the quality of the rest of clusters resolved.

This slight overestimation of the number of k clusters required leads naturally to the discussion about the number of clusters to be considered in a real case scenario (i.e., analysing an actual EELS dataset). Given the fast convergence of KM, the usual consensus is to carry out a serialized classification of the dataset selecting a range of incremental k-values and plotting the inertia of each clustering classification with respect to the objective number k of clusters specified. The inertia of a KM clustering is defined as the sum of squared distances from the datapoints (in this work, the EEL spectra) to the closer centroid resolved (i.e., the centroids defining each datapoint membership to a given cluster). The resulting plot should be similar to the one shown in **Figure 4.17** (d), which is the inertia curve for the KM classification of the iron-manganese oxide NP for the k-values in a range between 1 and 14 clusters (inclusive). The optimum value for the number of k clusters would be then the so-called knee (or elbow) of the curve. In this particular case, the number of clusters 3 occupy that position. This is the criterium employed to resolve the number of relevant clusters (and centroids) later used as endmembers (reference spectra) in a special type of multiple-linear-least-squares (MLLS) fitting called k-MLLS[61], although from the initial clusters only those that were linearly independent from the rest of the vectorial base were kept (i.e., the centroids that could not be retrieved as a linear combination of the rest of them). However, this criterium is based on heuristic assumptions, and may bury some interesting information by underestimating the number of clusters. For instance, notice that when the number of clusters is increased bellow the 'knee' some interesting spectral features are likely to be resolved (e.g., 9 objective clusters resolve the possible existence of a mixture of oxidation states for the manganese oxide, as a pre-peak appears and the ratio of some features changes for oxygen edge, (f), and the ratio and width of the Mn WLs changes, (g))[17]. Notice that the reference spectra for each cluster in (e) are

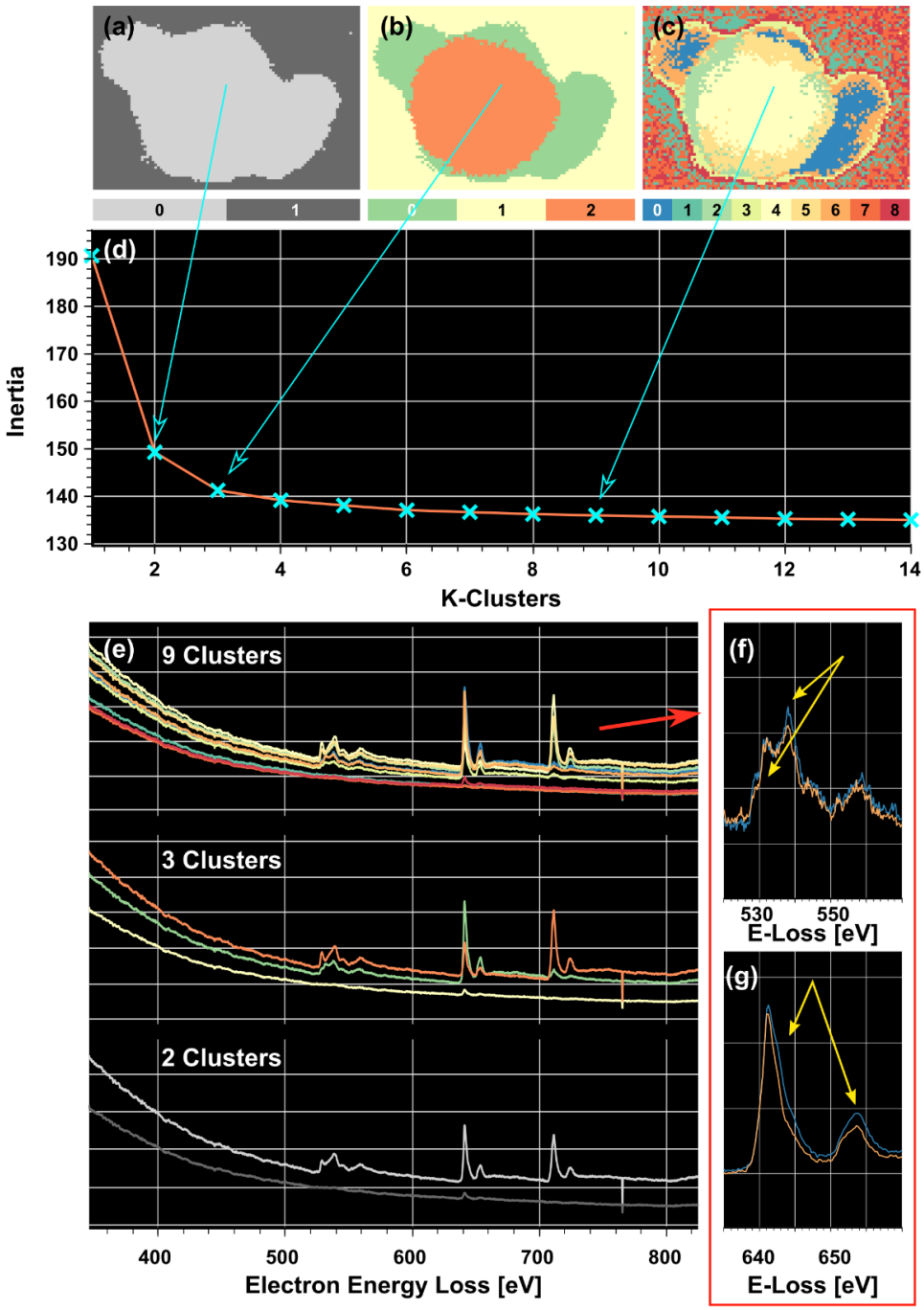


Figure 4.17. (a-c) KM label-maps of the clustering classifications of the normalized EELS SI of the Fe-Mn core-shell NP for 2, 3 and 9 k objective clusters respectively. (d) Inertia plot of the KM clustering for a range of k = 1 to 14 clusters. (e) Average values of the non-normalized spectra per cluster resolved for the 2, 3 and 9 clusters cases. (f-g) Detail of the average spectral values for clusters 0 (blue) and 6 (orange) for the 9 clusters case, for two regions of the spectrum (oxygen and manganese edges respectively). The yellow arrows indicate areas of interest in both edges showcased.

not the actual centroids, as the clustering classification was carried out for the normalized dataset and the spectra shown are the average values of the raw-data for each cluster resolved from the normalized dataset.

A final note on KM regarding the reproducibility of results is required. In order to speed up convergence (getting times equivalent to $\circ(N)$ for large datasets), most implementations (including the one from sklearn utilized in this work) resort to find a local optimum solution by initializing the centroids in random points within the dataspace of the sample (e.g., in EELS datasets, the first iteration sets $\vec{\mu}_k$ as spectra from pixels randomly selected). This makes the algorithm sensitive to get stuck in local minima without achieving an actual global minimum, depending on how it is initialized. A walkaround for this problem is to iteratively carry out several clustering classifications at random initial states, keeping at the end the one with the lower minimum value for the objective function after convergence. In all the examples presented in this work, the k-means++[62] implementation is precisely used to tackle this problem. Nonetheless, as the number of initializations is always a discrete number, and given that providing a very large number of initializations will compromise the actual convergence time gains and pose a memory intensive task (all the initializations are kept in memory up until the final selection), the risks of convergence to local minima are still present. Thus, using KM one may expect different results for different initializations.

Hierarchical Agglomerative clustering (Ward's linkage method).

The hierarchical agglomerative (HA) clustering constructs a hierarchy of clusters in a bottom-up approach, that begins by assigning an individual cluster to each datapoint and merging them iteratively. The end-result is a dendrogram, a tree-like structure that reproduces the cluster hierarchy, nesting groups by similarity of their properties. The actual clustering segmentation is obtained by a flat cut of the tree, which may be specified as an objective number of clusters (a hyperparameter in the sklearn default implementation) or done at a given level after observing the clustering structure.

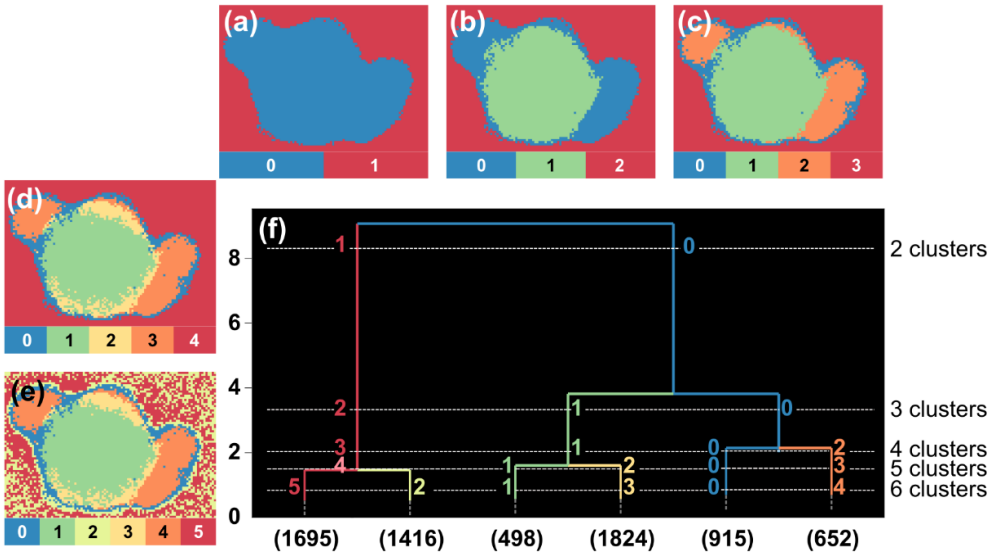


Figure 4.18. (a-f) Clustering label-maps for the hierarchical agglomerative algorithm implemented in sklearn, setting the objective number of clusters as 2, 3, 4, 5 and 6 respectively. (g) Dendrogram for the hierarchical clustering structure using the Ward linkage. Only the 6 larger clusters are drawn.

An example of HA-Ward (linkage) clustering is shown in **Figure 4.18** for the L_2-normalized EELS SI of the iron-manganese oxide NP. The colours of the dendrogram (f) are adapted to the clusters shown in (a-e). The dendrogram structure is pruned, meaning that only the top tree structure (up to a separation of 6 clusters) is shown, although the actual construction is done in a bottom-up approach. Also, the usual consensus in the data-mining and data-science community is to set the optimal cut of the dendrogram at the place with the larger vertical branches between horizontal intersections. As it is observed from the results (a-e), sometimes setting a lower cut reveals more information about the structure of the dataset, especially in cases of such high dimensionality as in an EELS SI. If one sets the optimal cut in this example in the 2 clusters limit, all the information about the core-shell structure is lost.

The merging of clusters that results in the dendrogram (i.e., the linkage method), can be done by a variety of different approaches (most of the times, user-selected). The most common method is known as Ward's linkage[63]. In it, the cluster-to-cluster distance is defined at each step as the average value of

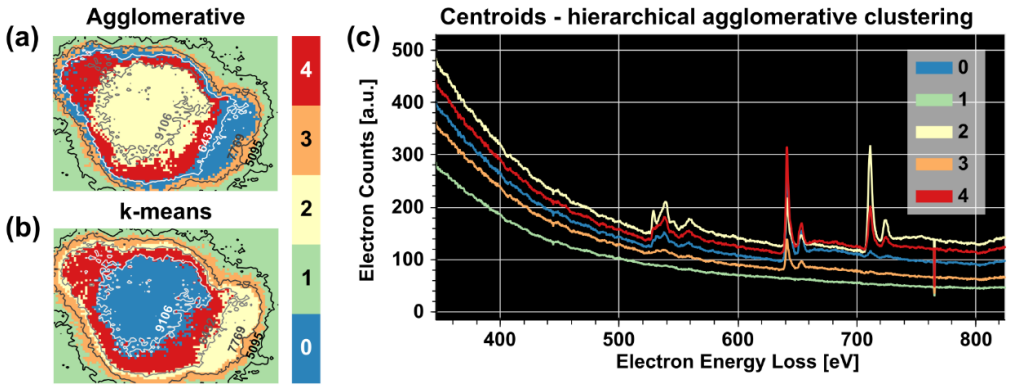


Figure 4.19. (a) HA-Ward clustering label-map and (b) KM clustering label-map, for the raw non-normalized dataset and setting the objective number of clusters as 5, overlapped in both cases with the L_2-norm contour levels. (c) Centroid signals for the HA-Ward clustering in (a).

the pairwise-distances between every datapoint member of a given cluster and every other datapoint belonging to a different cluster. The actual merge of clusters is resolved by the optimization of the sum of squares differences (errors) criterium. Hence, the parametric nature of this algorithm is derived from the same optimization criterium as in the case of the k-means clustering and, thus, the Gaussian ball assumption is also in place. For a non-normalized EELS dataset, this translates once again to a dominance of the L_2-norm values of the spectra for the clusters formation (**Figure 4.19** (a), (c)). Notice that KM produced clusters closer to the actual contour levels of the L_2-norm image (b) for the same number of clusters, but, still, the average spectra for each cluster in the HA show the same background signal dependence (c).

One major problem with this algorithm (and, as a matter of fact, any other HA clustering algorithm) is that the convergence times scale as $O(N^2)$.[9,50] That is a substantial increment compared to the sklearn implementation of k-means scaling as $O(N)$ for larger datasets. In a fairly small spectrum image, as the one analysed as an example in several figures of this text of 70x100 pixels (i.e., only 7000 datapoints), the time for HA-Ward is already 5 times larger than the one for KM (see **Figure 4.20** (a)(b)). The clustering structure is not identical in both cases (c), as HA-Ward seems to resolve better the C-background as a

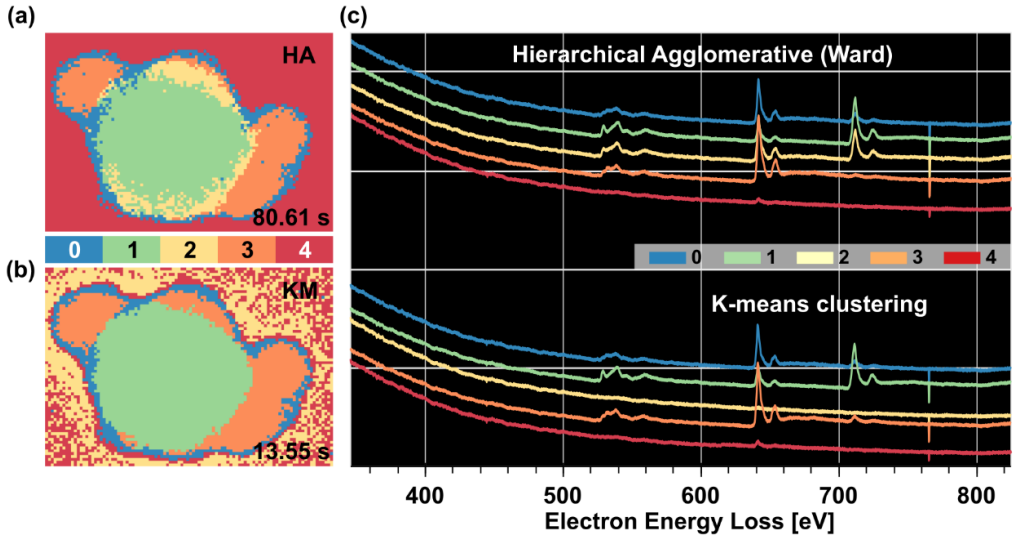


Figure 4.20. (a) Hierarchical agglomerative (Ward linkage) clustering. (b) K-means clustering. (c) Centroids for the KM clustering resolved in (b), and average signal spectra for the clustering resolved by HA-Ward in (a). Notice the lack of y-axis, as the reference EELS signals are separated artificially for a better observation of the spectral features.

singular group and discovers some structure within the iron oxide core (i.e., resolves 2 clusters for the core, depending on the proportion of iron-manganese oxide signals, given the overlapped nature of the core-shell structure).

HA-Ward also lacks back-tracking capabilities (i.e., presents an inability to escape from any possible convergence to local-minima and produces invariant deterministic results). Although getting a consistent (deterministic) result may sound as the better option for a classification method of data with scientific purposes (i.e., ensuring reproducibility), in noisy datasets, as the ones presented here and as it is usually the case with large EELS SI, having a way to escape from spurious local minima convergence might help on the discovery of spectral structures, otherwise opaqued.

Density-based clustering. DBSCAN

Up to this point, both HA-Ward and KM presented a similar approach to the resolution of a clustering classification, regarding the explicit/implicit

parametrization they carry over the dataset. As such, the data is always analysed as a set of instances extracted from the union of Gaussian probability distributions (i.e., the Gaussian-ball assumption dominates the analysis). In general, this would pose a problem in 2D and 3D datasets with a clear non-spherical shape. In EELS data analysis the problem runs deeper, as the actual dimensionality of the dataset is so high (potentially as high as the number of channels in an energy loss spectrum) that the visualization of the actual cluster shapes is no longer an option without a dimensionality reduction step (the introduction of UMAP will later be demonstrated as a key feature in clustering analysis)[33].

Density-based algorithms attempt a different approach to clustering analysis. In a nutshell, they try to identify dense regions in the data-space (clusters) and separate them by lower density regions (sparse areas)[9]. As such, the new notion of cluster as a dense area of datapoints (according to some parameters) is the key for a non-parametric approach to the problem, as no assumptions about the actual shape of these dense regions are made. Different algorithms will implement this notion of density in a variety of ways. One key feature of most of them is that they allow the classification of datapoints as noise, whenever those said instances do not belong to any of the resolved clusters with regard to some specified set of constraints.

In a language closer to statistical analysis, these algorithms will try to resolve an a-priori unknown continuous probability density function (PDF), governing an underlying and theoretical data-generation mechanism. They do so from the discrete (finite) dataset itself (data-driven methodology) and without imposing any parametric assumption along the way[64].

Among the density-based family of algorithms, one of the most popular representatives is the density-based clustering of applications with noise, DBSCAN for short[65,66]. Briefly, it constructs the clusters by analysing the ϵ-neighbourhood of each datapoint of the dataset D (being ϵ a radius measurement centred in each datapoint, employing any given metric), and declaring them as core points (that will define the actual clusters themselves)

if a number of datapoints equal or higher than a minimum value (min-points) is inside the ϵ-radius (including itself, the datapoint defining the centre).

In actuality, the notion of cluster described by DBSCAN has a higher degree of complexity, as it will also have to describe those datapoints which are in the neighbourhood of core-points but do not fulfil the core-datapoint condition (i.e., the border points). First, the density-reachability notion is introduced, where a datapoint $\vec{p}$ is density-reachable from a core datapoint $\vec{q}$ if a chain of core datapoints can be traced between them, fulfilling the ϵ-radius and min-points conditions (i.e., two border datapoints are never density-reachable between them, but the combination of two core datapoints or a core and a border datapoint may describe a density-reachable pair). Second, the density-connectivity notion that symmetrize the measurements[9,65,67] is introduced, which states that any two given datapoints $\vec{p}_1$ and $\vec{p}_2$ are densely-connected if they can be density-reached by a common $\vec{q}$ core datapoint. The clusters are finally defined as subsets of the dataspace $C_k \subseteq D$ such that any pair $\vec{p}_1, \vec{p}_2 \in C_k$ is densely connected, and only those $\vec{p}_{i=1,\dots,K} \in C_k$ can be densely-connected to any other given datapoint from the same cluster (i.e., datapoints from different clusters, or noise datapoints, are never densely-connected to datapoints from other clusters). The inner workings of the algorithm are meticulously described elsewhere[65].

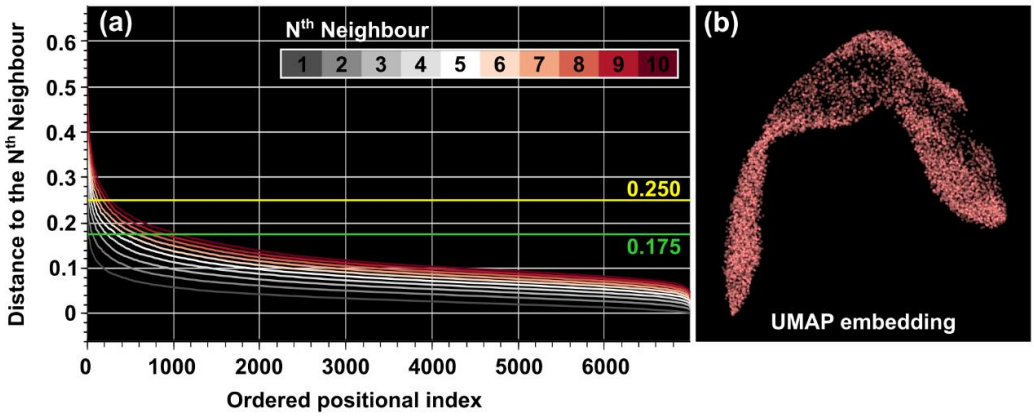


Figure 4.21. (a) Distance to the Nth neighbour for every datapoint in the UMAP embedding of the Fe-Mn oxide NP EELS SI shown in (b), and already showcased in **Figure 4.10**. The positional indices, the x-axis, are ordered so the distances are displayed from larger to smaller.

DBSCAN is considered also a single-partitioning algorithm because it requires from the user the input of the epsilon parameter (a single epsilon value). Furthermore, the value of min-points is also required as an input hyperparameter. The problem becomes then to obtain a successful clustering with two unknown variables ϵ and *min-points*.

A heuristic estimation method for these parameters is already given alongside the original formulation of the algorithm[65], and consist of the analysis of the pairwise-distances to the N^{th} neighbour for each datapoint. The ordering of these distances from higher to lower would ideally resemble the graph shown in **Figure 4.21** (a). Setting the ϵ value as the value defining a horizontal line intersecting the knee of the curve for the N^{th} nearest neighbour (NNN), and min-points equal to N, one would get a fair approximation to the optimum hyperparameter values. At the bottom of this methodology, one is basically analysing the distribution for the pairwise distances of datapoints. Notice that the N^{th} NN-graph in (a) is in fact calculated from the UMAP embedding showcased in (b) and in **Figure 4.10**. An actual N^{th} NN-graph for the normalized (or raw) dataset will most likely resemble the one shown in **Figure 4.22** (a) for any given EELS dataset. The detailed zoomed region in **Figure 4.22** (b) shows how closely packed the NNN-curves are, a by-product of the so-called 'curse of dimensionality'[68,69] (i.e., the distribution of distances for high dimensional datasets is flattened, and the side effect is that every datapoint becomes effectively equidistant to any other given datapoint). The results of DBSCAN on such EELS datasets, **Figure 4.22** (c), showcase the inability to separate relevant clusters, other than the differentiation between NP and C-background spectra (being the latter always classified as noise for more restrictive epsilon values, and not as a distinctive cluster).

Finally, representative clusters will only be resolved under a sensible parameter selection[9]. As the general convergence times are in the order of $O(N^2)$ (down to $O(N \log(N)$ in indexable databases, which would not apply to high dimensional EELS SI)[9,65], an iterative approach for a raw (or normalized) EELS dataset (as in the case of KM) is no longer commendable.

Hence, the use of dimensionality reduction techniques previous to the density-based clustering algorithms will always be recommended.

Hierarchical density-based clustering. OPTICS and HDBSCAN

The 'flat' nature of DBSCAN posed the immediate problem of selecting appropriate values for the hyperparameters epsilon and minimum-(number of)-points (a similar problem to the selection of k-clusters for KM). As demonstrated through the example on **Figure 4.22**, the clustering structure of a high dimensional dataset may not always be well described by a single set of density parameters that impose, thus, some global values to every subset of datapoints from the original hyperdimensional dataset. As with the case of KM, an alternative arises from the introduction of a hierarchical structure (i.e., completing the last box from the taxonomical classification proposed in **Table 4.3**). They are density-based clustering algorithms (i.e., non-parametric approaches). Here, the hierarchy is referred to the density-based clustering structure (i.e., a hierarchical ordering of clustering sets characterized by density measurements). The differences in the construction of the hierarchy will characterize the different available algorithms.

The first one described in this section has to be OPTICS (ordering points to identify the clustering structure)[55], as it is among the most successful algorithms for density-based clustering classification and the first of its kind to be used for EELS data analysis (low loss regions)[45]. It is common to see OPTICS described as a generalization (or extension) of DBSCAN, where an infinite number of neighbourhood distances (epsilon) are processed simultaneously[9,55]. Indeed, OPTICS foundations rely on the fact that the clusters extracted from DBSCAN are 'monotonic' with respect to the density definition (i.e., for the same value of minimum-points, a dense cluster defined for a neighbourhood ϵ_1 may contain subsets of denser clusters defined by a set of parameters $\epsilon_{2,3,\dots,k}$ such that $\epsilon_{2,3,\dots,k} < \epsilon_1$, as in the example from **Figure 4.22** (c) where the smaller ϵ values produced smaller and denser clusters -0-contained inside the sparser ones from larger ϵ values). Conversely, the min-

points-neighbourhood for different datapoints may be defined by different radial ϵ values.

Hence, instead of relying on a fixed ϵ value to define the min-points-neighbourhood of every $\vec{p}$ datapoint, OPTICS defines two new magnitudes to construct a hierarchy of densities. (1) The core-distance $d_{core}(\vec{o})$ for every $\vec{o}$ is defined as the ϵ' (upper bounded by a user provided ϵ) that ensures that the neighbourhood of $\vec{o}$ contains at least min-points (as a hyperparameter once again). (2) The reachability-distance (d_{reach}) to $\vec{q}$ from $\vec{p}$ (notice the directionality, non-symmetric magnitude) is defined as the minimum distance value that makes $\vec{q}$ density-reachable from $\vec{p}$, or mathematically

$$d_{reach}(\vec{p},\vec{q}) = \begin{cases} \max\{\, d_{core}(\vec{p}) \,, d(\vec{p},\vec{q}) \,\}, & if\ \exists\ d_{core}(\vec{p}) \\ undefined, & otherwise \end{cases} \tag{49}$$

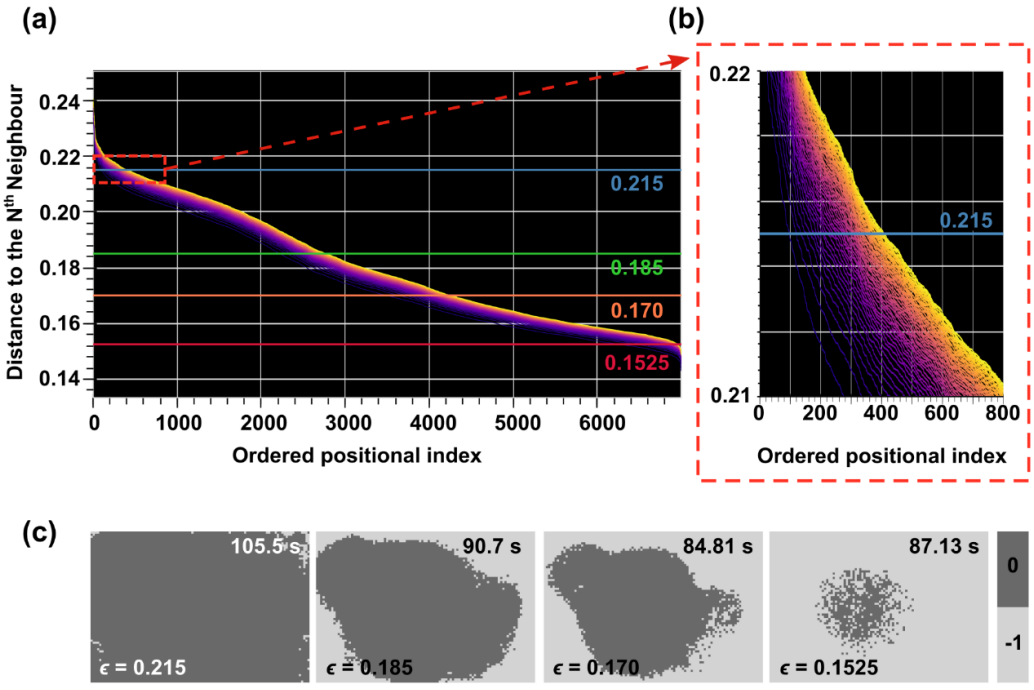


Figure 4.22. (a) Distance to the Nth (up to 50) neighbour for every datapoint in the L_2-normalized Fe-Mn oxide NP EELS SI. The positional indices, the x-axis, are ordered so the distances are displayed from larger to smaller. (b) Detail of the area squared in red in (a), showcasing the same distance curves up to the 50th neighbour. (c) Label maps resolved by DBSCAN, setting 50 as the min-points and epsilon as the values displayed in the horizontal lines in (a).

(i.e., using the same density-reachability notion as DBSCAN, it is the minimum distance that, being $\vec{p}$ a core object, sets $\vec{q}$ in the neighbourhood of $\vec{p}$). Both magnitudes might be set by OPTICS as undefined for any given datapoint, as ϵ' is effectively upper-bounded (e.g., $\vec{p}$ not being a core object would make any reachability-distance to $\vec{q}$ undefined from $\vec{p}$, or a given $\vec{o}$ with an $\epsilon' > \epsilon$ would have a core distance undefined). Also, from the definition itself several reachability-distances are likely to be defined for any given $\vec{p}$, as it may be directly density-reachable from several different core-points. The lower values are linked to the closer core-point and are, thus, the ones that will influence to a higher degree the construction of the hierarchical structure.

Finally, OPTICS will build a clustering ordering (assigning indices to each datapoint) from these newly introduced magnitudes. This ordering will cluster together datapoints with small values for the reachability-distance with respect to the same core datapoints (being the 'stronger' core-points those with the lower values of the core-distance magnitude, as they will represent a significant number of the connections via the reachability-distance). The end result is usually displayed in the so-called reachability plot (1D), being the x-axis the ordering index for each datapoint and the y-axis the reachability-distance. The specifics of the actual algorithm innerworkings are described elsewhere[9,50,55]

Parallel to the approach taken for the introduction of the non-linear dimensionality reduction techniques (UMAP and t-SNE, see **Figure 4.8**), a phantom 2D-dataset (**Figure 4.23**) is presented to help on the understanding of how a clustering classification is obtained from OPTICS. This synthetic 2D dataset is composed by 6 blobs of datapoints drawn from spatial Gaussian probability distributions with different sigma values and centres **Figure 4.23** (a). The reachability plot resolved by OPTICS is showcased in **Figure 4.23** (c) and **Figure 4.23** (e). The deeper valleys (lower d_{reach} values) correspond to the denser areas in the original dataset **Figure 4.23** (a). The actual clustering is obtained from flat cuts on the reachability plot. As expected, different level-cuts produce different clustering structures, shown in **Figure 4.23** (b) and

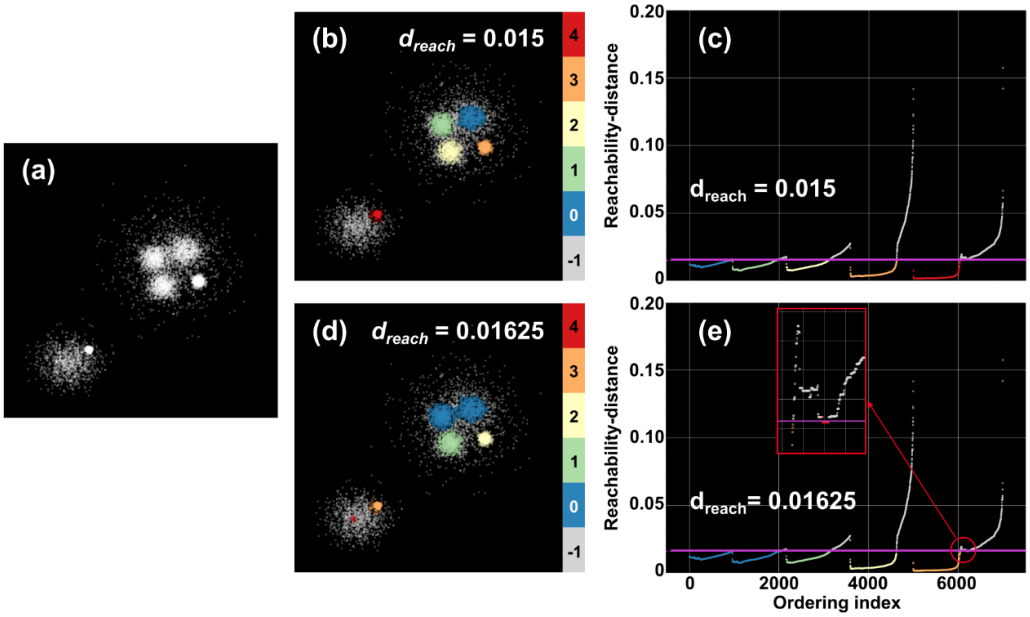


Figure 4.23. (a) Phantom 2D-dataset composed by 6 Gaussian blobs of different sigma values. (b) Colouring of the phantom dataset according to the flat cut at 0.015 on the reachability plot resolved by OPTICS in (c). (d) Colouring of the same dataset according to the flat cut at 0.01625 on the reachability plot resolved by OPTICS in (e). Red inset: zoom of the area encircled in red, corresponding to the separation of clusters 3 and 4.

Figure 4.23 (d). In both cases, the datapoints with reachability-distance values above any given flat cut are classified as noise. Notice that the graphs in **Figure 4.23** (c) and **Figure 4.23** (e) are coloured according to the clusters resolved in **Figure 4.23** (b) and **Figure 4.23** (d) by the flat cuts at d_{reach} = 0.015 and d_{reach} = 0.01625, respectively. The inset in (e) shows a zoomed area of the reachability graph, where the flat cut produces an extra cluster in **Figure 4.23** (d). This single level-cut imposition hinders the retrieval of every original blob at the same time. For instance, the differentiation of clusters 3 (orange) and 4 (red) in **Figure 4.23** (d), by setting epsilon to 0.01625, comes at the cost of merging clusters 0 (blue) and 1 (green) from **Figure 4.23** (b), resolved for an epsilon value of 0.015.

In order to test OPTICS with a higher dimensionality dataset, the normalized Fe-Mn oxide NP EELS SI is brought up again for analysis (with a dimensionality equal to the number of energy loss channels in the spectra). The results are

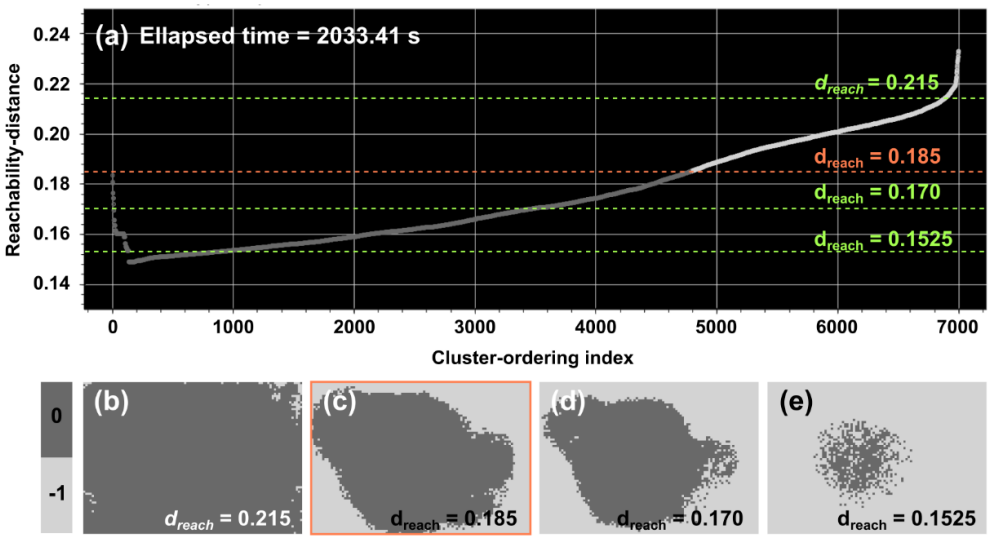


Figure 4.24. (a) Reachability plot calculated via the OPTICS algorithm for the L2-normalized dataset of the Fe-Mn oxide NP. (b-e) Label-maps extracted from several cuts on the reachability plot shown in (a).

shown in **Figure 4.24**. The reachability plot in **Figure 4.24** (a) shows a single valley and a monotonically crescent set of values for the d_{reach} of the rest of datapoints. This behaviour is once again (as in the DBSCAN case) a collateral effect of the curse of dimensionality, still a major issue using OPTICS in extreme cases of high dimensionality[66]. Some clustering classifications are included in **Figure 4.24** (b-e), corresponding to flat level cuts on the reachability plot. The d_{reach} values selected are equal to the ϵ values in **Figure 4.22**. Notice how there exist an almost perfect one-to-one correspondence between the cluster structures resolved by the OPTICS algorithm (**Figure 4.24** (b-e)) and the ones resolved by DBSCAN (**Figure 4.22** (c)), highlighting the close relation between the flat cuts at d_{reach} in the OPTICS hierarchy and the ϵ for DBSCAN. This behaviour further supports the interpretation of OPTICS as a generalized DBSCAN for multiple epsilon values. Only a handful of small differences are observable in the cluster membership for some boundary pixels of cluster 0 for $d_{reach} \mid \epsilon = 0.185$.

The second candidate, and the actual state-of-the-art density-based hierarchical algorithm would be hierarchical-DBSCAN (HDBSCAN), that has

been described by its own authors as a conceptual improvement upon OPTICS[70–72], and also recently introduced for EELS data analysis (core-loss SI)[our paper]. Much like OPTICS, it constructs a whole hierarchy of clusters of varying densities that can be summarized as running DBSCAN for a range spamming to every possible epsilon value (from infinite to zero).

As DBSCAN and OPTICS, it is a non-parametric clustering algorithm that avoids making any assumption about the subjacent probability density function (PDF). The PDF is estimated from the data itself (data-driven methodology) using the same notion of core-points distances $d_{core}(\vec{o})$ as OPTICS. However, HDBSCAN substitutes the asymmetric reachability distance by the symmetric counterpart, the mutual reachability distance (d_{mreach})[73].

$$d_{mreach}(\vec{p},\vec{q}) = \max\{\, d_{core}(\vec{p}) \,, d_{core}(\vec{q}) \,, d(\vec{p},\vec{q}) \,\} \tag{50}$$

Notice that $d(\vec{p},\vec{q})$ is not required to be any given specific metric (could be euclidean, Manhattan, etc.). As HDBSCAN is not upper-bounded the d_{mreach} is always defined (i.e., ϵ can effectively run from 0 to infinite, or, equivalently, the cluster density λ runs from infinite to 0). Also, as this magnitude fulfils the triangular inequality (unlike d_{reach} in OPTICS) it constitutes an actual well-defined metric by itself, which is a convenient characteristic that allowed the implementation of the accelerated version used throughout this work[33,71].

Conceptually, HDBSCAN constructs (in a top-down approach) a hierarchy of density-based clusters from a mutual reachability graph[70], whose vertices are the datapoints and the edges between them are drawn with weights according to the d_{mreach} values. The specifics about the computations carried out by the algorithm are detailed elsewhere[70,72]. Up to this point, HDSCAN would have constructed the equivalent of a single-linkage tree from the minimum points value required to define the core-distance for each datapoint. Any given epsilon value afterwards acts as a 'flat-cut' at a density-level in the hierarchical construct, calculated via the elimination of the edges with a weight

value above epsilon from the graph[11]. Thus, any finite epsilon value will separate the connected components in the graph (clusters) from the rest of datapoints (noise). In the practical implementation of HDBSCAN, as it happened for DBSCAN and OPTICS, increasing the value of m_{pts} has a smoothing effect over the density estimates (i.e., a conservative approach to the cluster designation).

One critical problem with other hierarchical algorithms, especially dealing with noisy datasets, is the consideration of every single split from a cluster as a new cluster of its own, causing an excessive branching of the tree. This situation is especially grave in larger datasets, resulting in a very convoluted hierarchical structure of difficult interpretation (e.g., a dendrogram plot with thousands of branches, most of them corresponding to non-meaningful cluster splits from background noise). HDBSCAN simplifies the hierarchical structure by considering the concept of rigid clusters[74], by which any given connected component in the graph (cluster) evolves according only to 1 of 3 possibilities when the density (λ) requirement is increased (i.e., the ϵ parameter value is decreased): (1) the cluster shrinks in size but remains connected (is still the same cluster with the same label) up to a density threshold level, at which (2) the cluster splits in two new sub-sets or (3) it disappears. Thus, HDBSCAN produces a sort of 'pruned' dendrogram tree, where only the meaningful clusters and cluster splits are represented and the noisy datapoints fall from the tree branches (clusters) as the density requirement is increased. The practical implementation of HDBSCAN calls this structure a 'condensed tree plot', and a hyperparameter called minimum-cluster-points ($m_{Cl\text{-}pts}$) defines the meaningful cluster size[71]. The single linkage structure and the ability to select any given d_{mreach} radius (flat cut) is still provided, which is the equivalent of having access to the infinite range of DBSCAN clustering solutions for an

[11] Again, the inverse proportionality relation of density-epsilon means that increasing the density requirement of a level-set (cluster) in HDBSCAN is equivalent to decreasing the epsilon value.

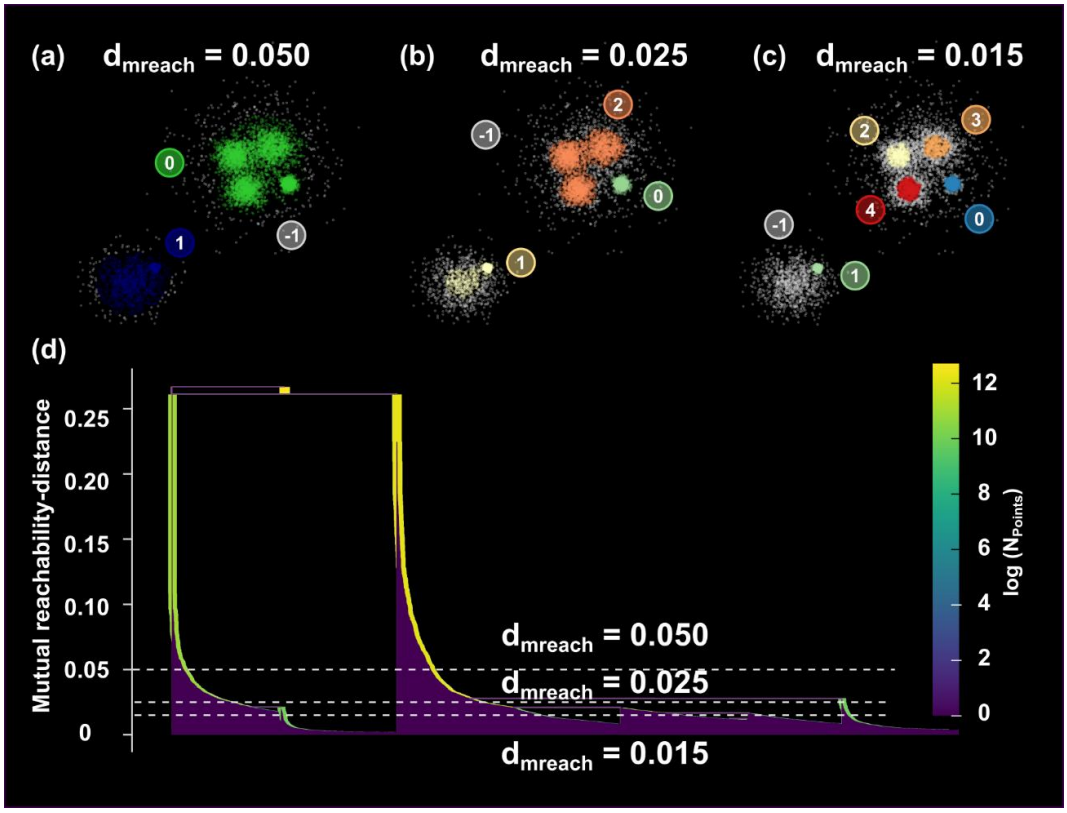


Figure 4.25. HDBSCAN clustering results extracted from the single linkage graph calculated for a mutual reachability distance value of (a) 0.05, (b) 0.025 and (c) 0.015. (d) Mutual reachability distance single linkage graph calculated by HDBSCAN.

infinite number of ϵ values. An example of this feature utilizing the phantom dataset from **Figure 4.23** is included in **Figure 4.25**.

Although a large quantity of information may be extracted from the exploratory analysis of the hierarchical structure resolved by HDBSCAN (as in the case of the reachability plot and OPTICS), the end goal in many applications (EELS data analysis included) is the extraction of viable clusters. HDBSCAN implements a method to discover the so called 'natural' clusters, other than the common flat level-cut[75] (as those shown in **Figure 4.25**) through the hierarchy employed in OPTICS (i.e., flat cut through the reachability plot, **Figure 4.23**) or HA-Ward (i.e., flat cut through the dendrogram, **Figure 4.18**). These 'natural' clusters are selected from the branches in the condensed hierarchical structure presenting the larger persistence (i.e., the longer living branches of the tree in terms of density, λ). This allows the algorithm to select

different level-sets in the hierarchy and, thus, HDBSCAN is capable of identifying simultaneously areas with very different local densities[70].

An example of HDBSCAN and the hierarchical structure it produces is shown in **Figure 4.26**, using the synthetic 'blobs' dataset shown already in **Figure 4.23**. The natural clusters resolved by HDBSCAN are highlighted in **Figure 4.26** (b), and are extracted from the condensed tree plot shown in **Figure 4.26** (c). Notice that the separation of 1 and 2 (labels from HDBSCAN) cannot be achieved by any given flat cut through the hierarchy without losing the separation of clusters 4 and 5. This effect was already observed when analysing the same dataset via OPTICS (**Figure 4.23**). This is a major winning point for the natural cluster selection of HDBSCAN.

The results of HDBSCAN in higher dimensionality dataset, the same EELS SI for the Fe-Mn oxide NP, are shown in **Figure 4.27**. Flat cuts on the single linkage

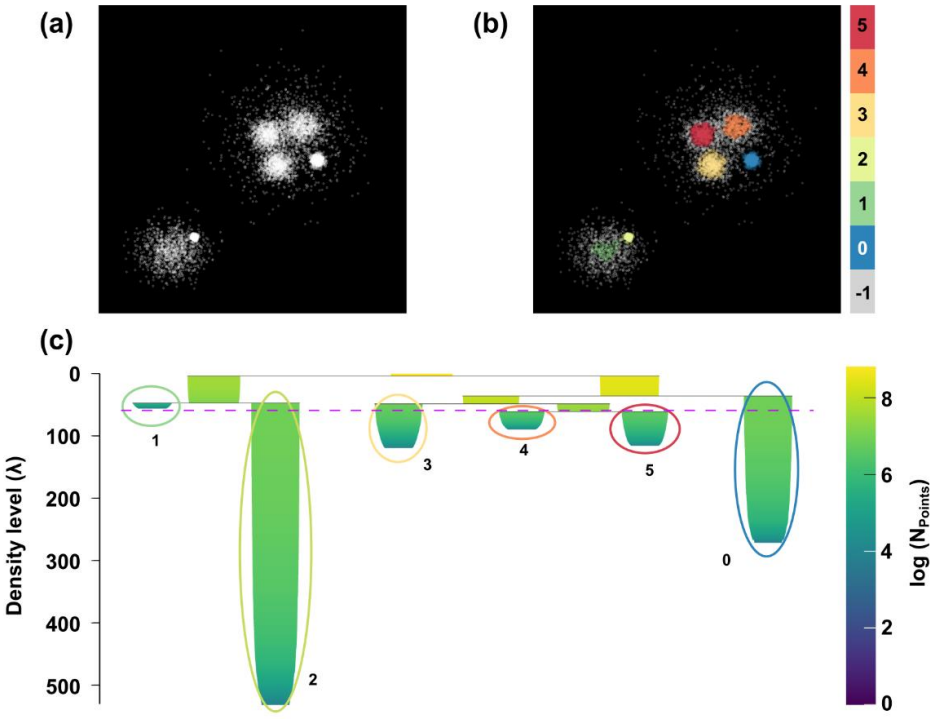


Figure 4.26. (a) Original phantom 2D dataset, shown also in **Figure 4.23.** (b) HDBSCAN natural clustering classification. (c) HDBSCAN condensed tree plot for the dataset in (a), setting a minimum cluster size of 50 and equal minimum number of samples (i.e., m_{pts}). The branches resolved as the natural clusters are highlighted with the same colour-code of (b).

tree for the mutual reachability distance calculated by HDBSCAN are included from **Figure 4.27** (a) to (c), equivalent to the flat cuts done on the reachability plot by OPTICS in **Figure 4.24** and the different epsilon radius selected for DBSCAN in **Figure 4.22**. The natural clustering separation **Figure 4.27** (d) fails for the EELS dataset and for the minimum cluster size selected (50, equal to the m_{pts}), as every datapoint is grouped on the same branch in the condensed-tree-plot **Figure 4.27** (e). The reason, once again, the curse of dimensionality[66]. However, a positive takeaway point is that to obtain an equally flawed segmentation HDBSCAN took only 197 s, whereas OPTICS took over 2000 s.

One last comment on the HDBSCAN capabilities is that it includes a soft-clustering method, where all the noise datapoints are classified in one of the natural clusters discovered. This is a very interesting feature for EELS data analysis, as the pixels classified as noise might actually present a large spectral variation among them and, thus, the classification of all of them as part of the same group might lead to a misleading interpretation of the spectral characteristics of the dataset.

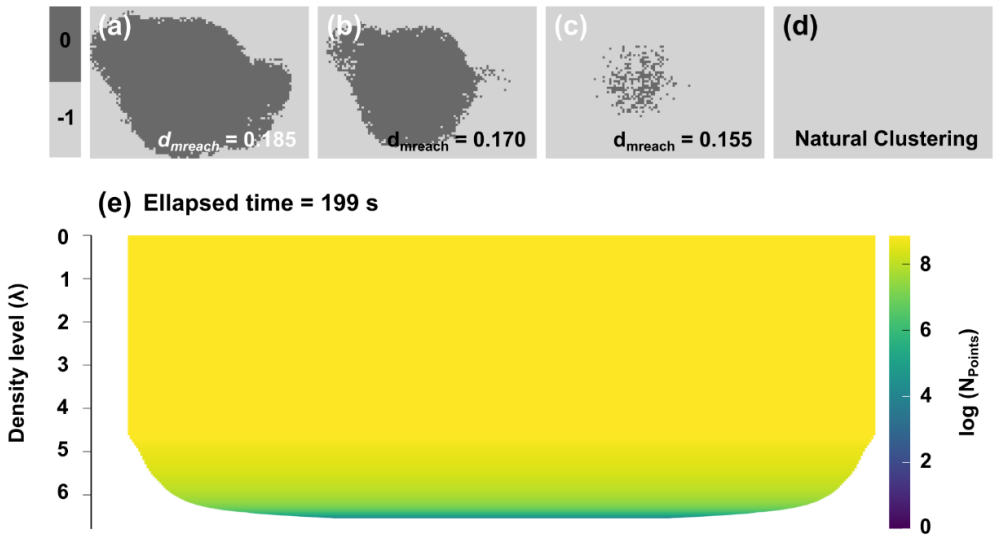


Figure 4.27. (a-c) Flat cuts of mutual reachability distances through the single linkage hierarchical structure built from HDBSCAN for the Fe-Mn NP normalized EELS SI. (d) Natural clustering found by HDBSCAN. (e) Condensed-tree-plot from HDBSCAN.

4.5. Combining DRM and clustering analysis.

The proposition of combining dimensionality reduction techniques and clustering analysis to characterize EELS datasets dates back to the original publication for clustering of EELS SI[76]. There, the use of HA-Ward over the representation matrices of two components resolved by PCA was proven capable of producing a convenient segmentation of an EELS-SI, with regard to some very specific spectral characteristics of interest (i.e., the differences in the manganese edge white lines that pointed out the existence of spatially localized changes in the Mn oxidation state). However, this methodology presents two critical flaws. (1) First, the components are hand-picked by the user from the results of the matrix factorization dimensionality reduction technique. This is a clear invitation to fall in a confirmation bias situation (i.e., the components may be selected according to information about the dataset predating the analysis and, thus, the results may be interpreted likewise). (2) Second, the components resolved by PCA for EELS datasets usually contain a mixture of spectral characteristics of difficult interpretation (produced by a combination of different underlying physical mechanisms). In many cases, these components can even contain non-physical features. These effects are driven by the orthogonality imposition[37] and the statistical nature of the algorithm itself, as described in the introduction of PCA in the previous subsections. Summarizing the potential issues of this methodology: (1) the selection of components is inevitably sorted out by a potentially biased user and (2) the risk of involuntary (systematic) bias arising from the spectral features of the actual resolved components are always present.

The use of PCA for denoising purposes previous to a KM clustering classification has also been tested with successful results[16]. It constitutes a natural evolution from the extended practice of using matrix factorization techniques (most notably PCA) for denoising purposes in EELS data analysis[17,27,77]. Nevertheless, a word of caution is required, as it is also well known how PCA might hide some subtle spectral features inside the noisy

components. Thus, cutting out too many of them in order to clean up the spectra might have some nasty and unpredictable statistical consequences on the recovered spectra[41,43] and further influence the clustering results.

Moving onto the density-based algorithms, the combined use of manifold learning (t-SNE, to get a 2D embedding of the dataset) and OPTICS has been demonstrated as a viable solution for the spectral segmentation of an EELS dataset, tested by separating different plasmonic resonant modes from a low-loss EELS SI[45]. The authors in that publication also included a NMF decomposition step previous to the application of t-SNE, in order to increase the performance of the latter when facing the large dimensionality of the original dataset (EELS SI). They claim that the selection of the whole set of components resolved by the matrix factorization prevents the falling in the aforementioned confirmation bias, as it avoids any type of analysis of the components resolved and completes a fully data-driven methodology. Nevertheless, as already pointed out several times throughout this work, the approximation of a dataset by a low rank model from a matrix factorization technique (among which NMF is included) always implies some degree of statistical assumptions about the dataset. Nevertheless, the use of a hierarchical density-based algorithm on an approximated manifold projection (the t-SNE embedding) is an interesting approach. The low dimensionality of the embedding tries to capture the disposition of datapoints in the original hyperdimensional space and, in parallel, it eliminates the effects of the curse of dimensionality when calculating the cluster hierarchy. However, the use of OPTICS enforces the selection of a flat-cut on the hierarchy (a singular value of the d_{reach}) to extract an actual clustering classification. This clearly undermines the unsupervised nature of the whole process and, up to a certain degree, breaks the illusion of a fully data-driven methodology.

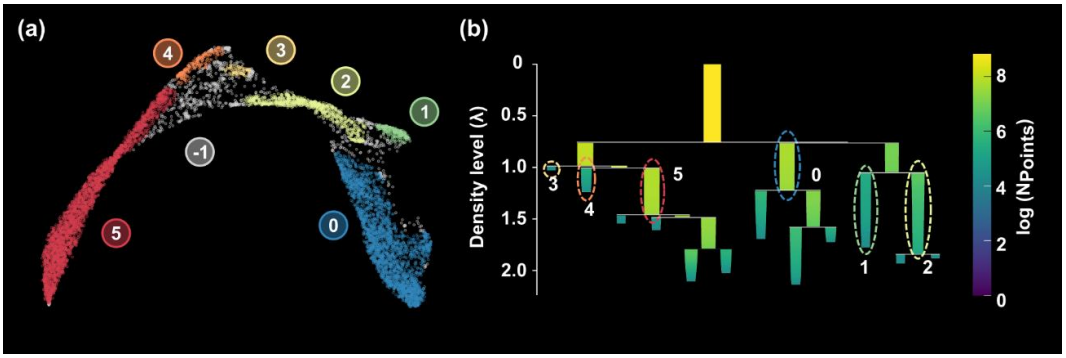


Figure 4.28. (a) UMAP embedding of the Fe-Mn Core shell NP, computed from 45 NN and minimum (spread) distance of 0.05. The datapoints are coloured according to the colour scheme of the natural clusters of the UMAP embedding resolved by HDBSCAN. (b) Condensed tree plot resolved by HDBSCAN, using 50 as the minimum cluster size. The branches for the natural clusters resolved are highlighted with the same colours used in (a).

In the context of this PhD thesis[33], we have gone a step further on the combination of non-linear dimensionality reduction algorithms for the manifold approximation and non-parametric (density-based) hierarchical clustering algorithms for the discovery (and later resolution) of clusters. The novelty comes from the algorithms proposed. (1) Regarding the dimensionality reduction, UMAP is conceptually (and in practice) a superior choice when compared to t-SNE. It has been shown capable to translating both the local and global structure from the original hyperdimensional space of an EELS SI dataset to the lower dimensional embedding. It is also capable of projecting the data to a higher dimension than 3, even if that implies that a single graphical representation is no longer an option. Furthermore, UMAP is capable of dealing with an unspecified large number of original dimensions. (2) Regarding the clustering classification, HDBSCAN outperforms OPTICS in several points, as stated in the previous subsection. Now that a lower dimensional embedding that preserves the hyperdimensional structure of the dataset is in place (via UMAP), the curse of dimensionality no longer applies. The results of following this methodology for the clustering analysis of the raw Fe-Mn oxide core-shell NP EELS SI dataset are shown in **Figure 4.28** and **Figure 4.29**.

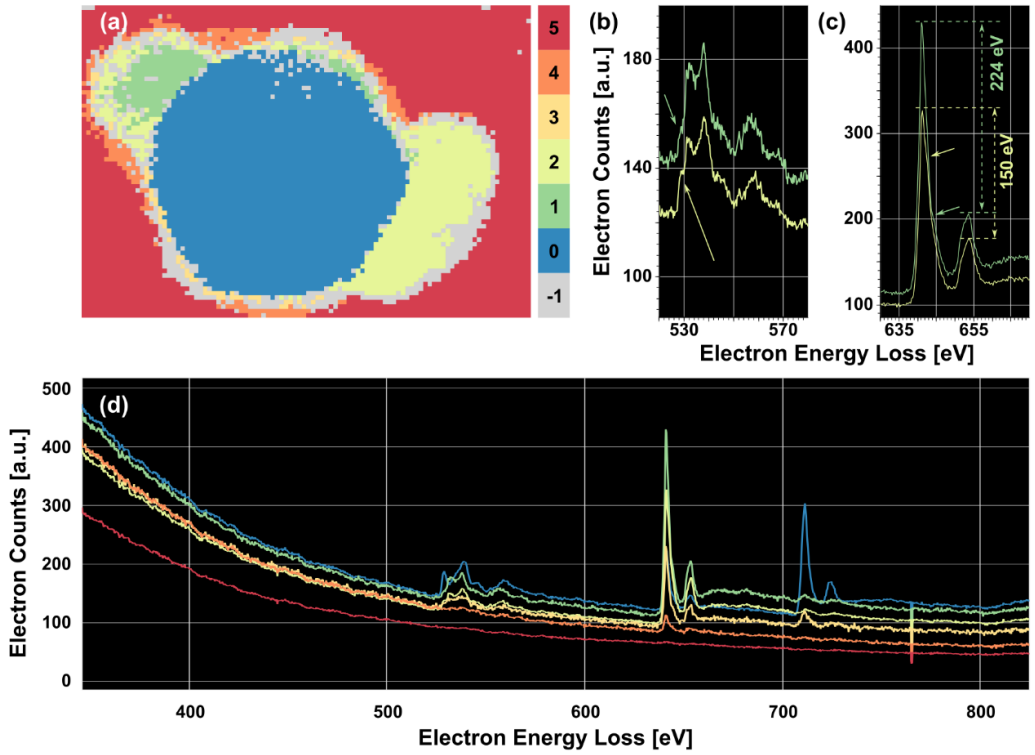


Figure 4.29 (a) Label map for the natural clusters resolved by the HDBSCAN of the UMAP embedding showcased in **Figure 4.28** (a) for the Fe-Mn core-shell NP EELS SI. (b-c) Detail of the average spectra for the pixels resolved as clusters 1 and 2 (manganese oxide shell), zoomed at the oxygen and manganese edges (respectively). (d) Average spectra for each of the natural clusters resolved by HDBSCAN, excluding the pixels labelled as noise (-1).

Given that the final goal of this methodology is to resolve a clustering classification, the parameters used for UMAP gave preference to the separation of the different areas in the 2D embedding projected. The structure in **Figure 4.28** (a) is the result of setting the so-called minimum distance as 0.05. Comparing it to the embedding of the same dataset from **Figure 4.10**(a), the overall structure is still the same but the accumulation areas (blobs of datapoints) are now denser and better defined. The condensed tree plot in (b) contains markings for the natural clusters resolved.

Having the hierarchical relation of clusters is an important achievement, but in EELS data analysis most of the times one is concerned with the average signal per cluster (i.e., the equivalent of a centroid) and the distribution of

clusters themselves over the EELS-SI area. The HDBSCAN label map for the EELS SI is shown in **Figure 4.29** (a). The colour scheme is shared with **Figure 4.28**. A clear separation between the core and shell is visible, and most of the disputed pixels (noise cluster -1, grey) are located at the interface NP-carbon background. Within the manganese oxide shell (see **Figure 4.29** (a) and (d)), two main clusters are resolved (1, green, and 2, yellowish). A detailed zoom of the reference spectra for both clusters is shown in **Figure 4.29** (b-d), focused on the fine structure of the oxygen and manganese edges (respectively). Cluster 2 presents a pre-peak in the oxygen edge, and the WL ratio for the Mn L_{32} edge is much smaller than the ratio for cluster 1. Both are characteristic features of different oxidation states[17]. Cluster 1 would be mostly composed by Mn_3O_4, and cluster 2 by a mixture of Mn_3O_4 and MnO. This qualitative

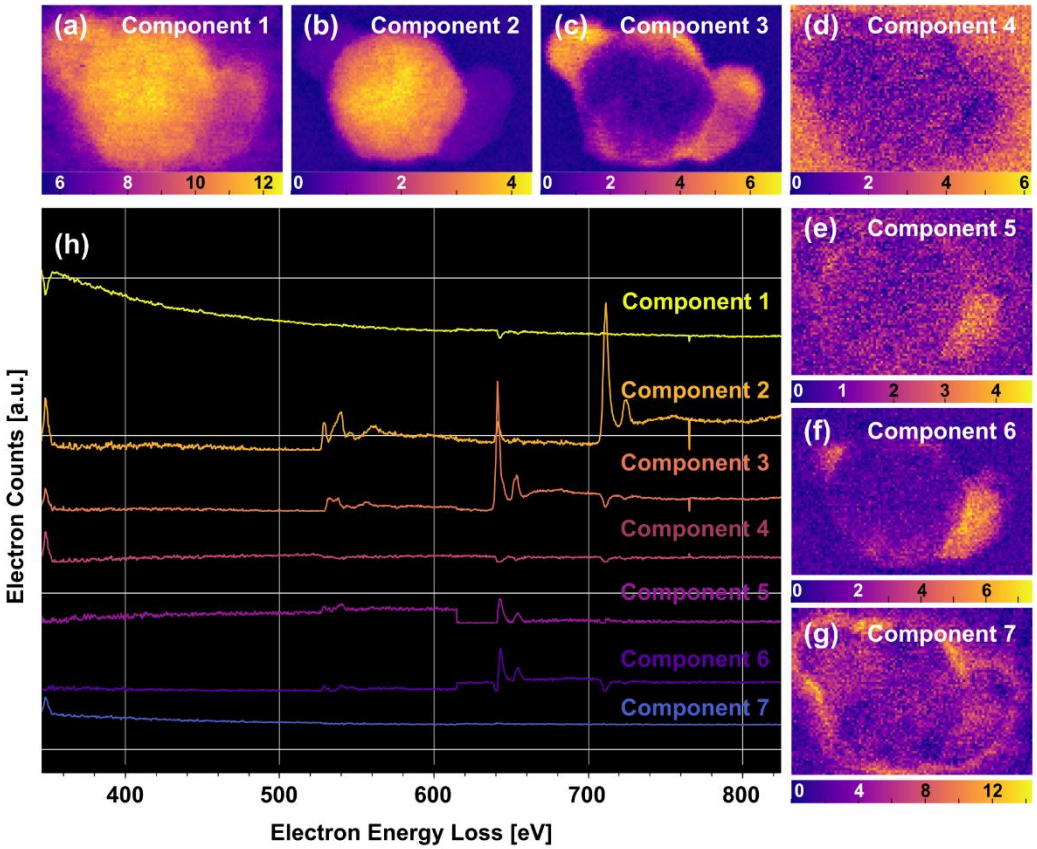


Figure 4.30. (a-g) 7 representation matrices (**L**) resolved by NMF using the Frobenius distance for the minimization algorithm for the Fe-Mn oxide NP EELS SI. (h) Unscaled archetypes for the 7 components resolved by NMF in (a-g).

analysis it is coherent with what has already been observed previously for this exact dataset[17].

Carrying out a matrix factorization decomposition previous to the manifold learning step is no longer a requirement. Nevertheless, some interesting structures in the dataset may arise from such combination of techniques. **Figure 4.30** shows the decomposition results for a NMF of the raw dataset. This time around, the objective function for the minimization problem was the Frobenius distance and, thus, the matrix factorization problem is described by eq.(6). The reason is that the results produced by such factorization method presented a lower degree of sparsity than those results of NMF using the KL-divergence as the objective function for the minimization problem (eq.(11)). This is observable comparing **Figure 4.30** and **Figure 4.6**. Overall, when calculating the UMAP embedding from the NMF components resolved this approach showed a better stability and less tendency towards the collapse to string-like structures of points in the 2D projection. The number of objective dimensions of the NMF decomposition was kept low (7). In order to avoid an excessive repulsion of the local structures calculated from UMAP (ending again in string-like accumulations of points), a fairly large number of NN was included in the calculation[12]. The 2D embedding resolved by UMAP is shown in **Figure 4.31**(a), coloured by the natural clusters resolved from the hierarchical structure calculated by HDBSCAN in (b).

The label map for this clustering classification is shown in **Figure 4.32**(a), and the average EELS signals per cluster are shown in **Figure 4.32** (d). The detailed zooms (**Figure 4.32** (b-c)) on the oxygen and manganese edges still indicate the occurrence of two different oxidation states for the manganese oxide shell, although this time the spatial distribution (i.e., the position of the

[12] Remember from the theoretical introduction of UMAP that the low dimensional embedding is constructed using a force directed layout. The edges between neighbouring datapoints work as attractive forces. Setting a number as high as the 10% of the dataset for the NN calculation ensured that several edges where forcibly drawn between points of different clusters, helping with the prevention of collapse to local structures of the datapoints. This wasn't an issue in the raw data embedding, probably because the higher degree of noise associated to each datapoint ensured an even distribution.

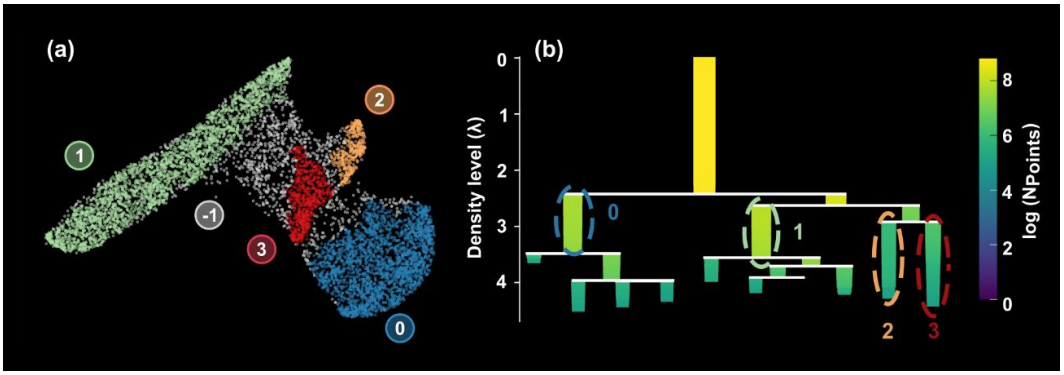


Figure 4.31. (a) UMAP embedding of the 7 components resolved by NMF (**Figure 4.30**) carried over the Fe-Mn Core shell NP, computed from 700 NN and minimum (spread) distance of 0.25. The datapoints are coloured according to the colour scheme of the natural clusters of the UMAP embedding resolved by HDBSCAN. (b) Condensed tree plot resolved by HDBSCAN, using 150 as the minimum cluster size and 30 as the minimum samples value (the minimum core datapoints per cluster). The branches corresponding to the natural clusters resolved are highlighted with the same colours used here in (a).

clusters in the SI) is different (compare **Figure 4.29** (a) and **Figure 4.32** (a)). Also, from **Figure 4.32** (b-c), the oxygen pre-peak is almost non-existent for cluster 3 (red), and cluster 2 (orange) shows a clear a broadening of the MnL_{32} WLs and a displacement towards higher energy losses. Again, this spectral analysis indicates predominance of MnO in the area resolved as cluster 3 and a mixture of MnO and Mn_3O_4 in the area of cluster 2, coinciding with the behaviour previously reported for this sample[17] via MLLS analysis.

This clustering classification from **Figure 4.32** contains a large number of pixels labelled as noise, especially compared to the clustering from **Figure 4.29**. This is not a surprising result, especially considering that the UMAP embedding was carefully adjusted to avoid the collapse of the embedding towards the locally denser regions and that, thus, the datapoints ended more evenly spread over the $\mathbb{R}^2$. The natural clusters are connected by a large number of datapoints that the algorithm of HDBSCAN classifies as noise in order to avoid the merger of locally dense clusters. Given that HDBSCAN allows a soft-clustering analysis to classify the noise datapoint[70,71], this example presents a good opportunity to test this feature. The results are showcased in

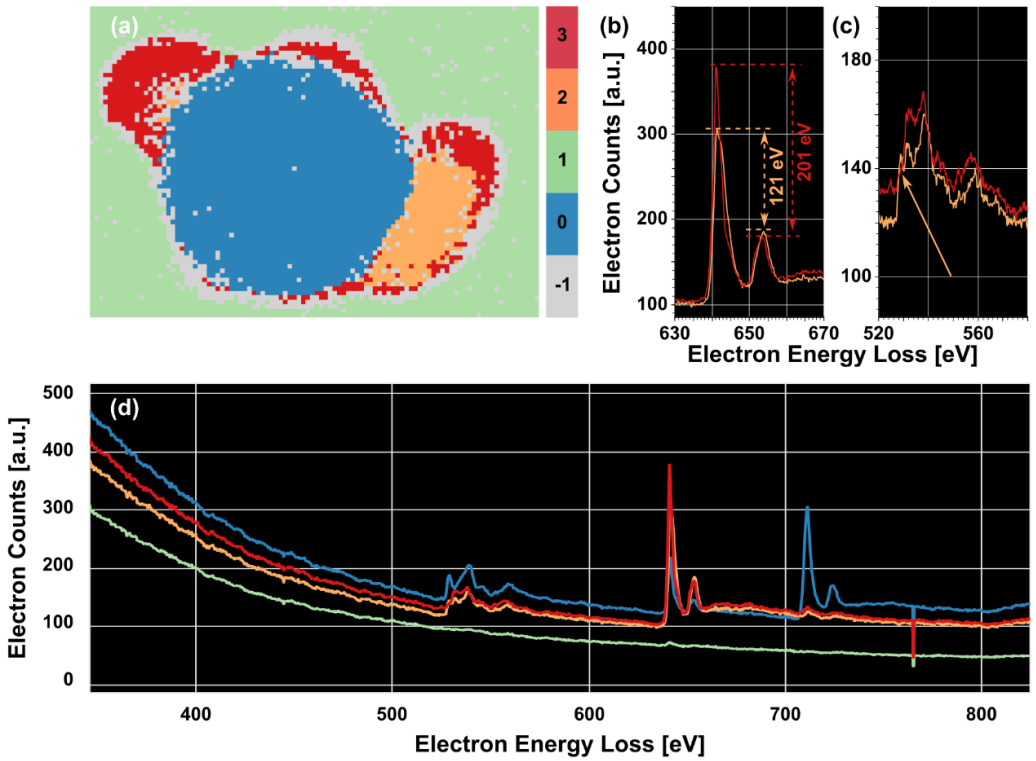


Figure 4.32. (a) Label map for the natural clusters resolved by the HDBSCAN of the UMAP embedding showcased in **Figure 4.31** (a) for the 7NMF components (**Figure 4.30**) of the Fe-Mn core-shell NP EELS SI. (b-c) Detail of the average spectra for the pixels resolved as clusters 2 and 3 (manganese oxide shell), zoomed at the oxygen and manganese edges (respectively). (d) Average spectra for each of the natural clusters resolved by HDBSCAN. Pixels labelled as noise (-1) excluded.

Figure 4.33. The alpha values for the datapoints in **Figure 4.33** (b), corresponding to the soft clustering classification scores[13], are in good agreement with the natural clusters from **Figure 4.33** (a), as the datapoints with alphas closer to 1 are those of the denser regions (core-points) of the natural clusters resolved from the hierarchy of HDBSCAN. The alpha values for the formerly noisy datapoints also provide an idea of the diffuse nature of the boundaries between the natural clusters resolved by HDBSCAN.

[13] The soft-clustering routine assigns to each datapoint a number of scores equal to the number of clusters resolved. These scores are valued between 0 and 1. They represent the proximity of each datapoint to the core-points of each of the clusters resolved. Datapoints closer to the core-points of a given cluster present values closer to 1 for a given score and closer to 0 for the rest, and vice versa.

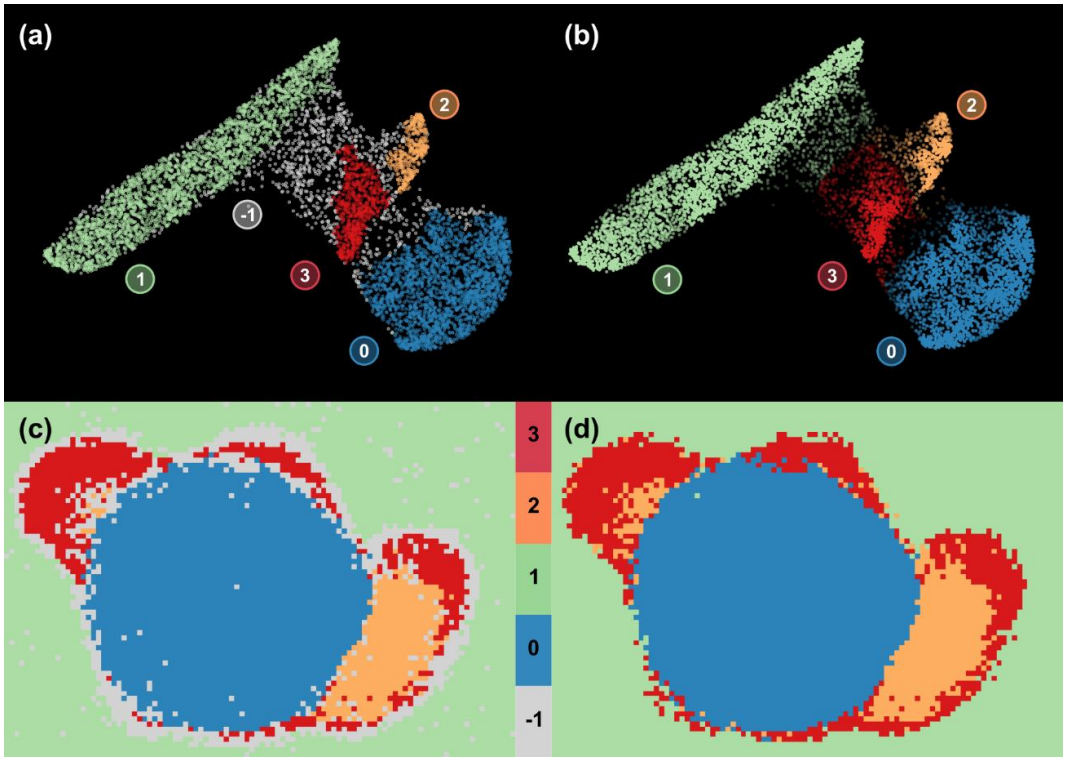


Figure 4.33. (a-b) UMAP embedding for the 7 NMF (Frobenius) components resolved for the Fe-Mn oxide core-shell NP EELS SI. The colour schemes correspond to the natural clusters (a) and the soft-clustering results (b) resolved by HDBSCAN. The alpha values for the colours in (b) are given by the score values of the soft-clustering classification. (c-d) Label maps for the EELS SI using the same colour schemes of (a) and (b), respectively, for the natural clusters and the soft clustering results from HDBSCAN.

One final thought on the use of UMAP and HDBSCAN for clustering. There exists a trade-off between the inclusion of a matrix factorization pre-step or the raw computation of the UMAP embedding from the dataset. For a fully unsupervised routine, that minimizes also the statistical assumptions made about the EELS SI generation processes, the direct calculation of the UMAP embedding from the raw dataset would be the path to follow. It has been demonstrated capable of producing a successful clustering segmentation, by HDBSCAN, that captures most of the spectral features of the challenging noisy EELS SI of the Fe-Mn oxide core-shell NP. However, the inclusion of a NMF pre-step 'fine-tuned' the clusters resolved, capturing to higher degree the important changes in the oxidation state of the manganese in the NP shell

(reported previously elsewhere[17]). This NMF decomposition required, however, a finer control over the embedding parameters, as the projected embedding tended to an uncontrolled collapse towards local structures.

4.6. Performance analysis.

In the previous subsections a series of dimensionality reduction techniques and clustering algorithms were introduced. Each one of them was tested with an experimental EELS SI of an iron-manganese oxide core-shell NP. In each case, a general qualitative analysed was conducted, in an attempt to understand how those results came to be from the actual formulation of the algorithms themselves.

However, the question of which clustering algorithm and which combination of DRM and clustering algorithms produces the overall better results remained unanswered. Carrying out an actual performance analysis on a clustering classification is a complicated task. For once, most of the scoring techniques rely on the knowledge of a ground truth (i.e., knowing, before the classification itself, the actual 'true' grouping of the datapoints). Some other techniques that do not require a ground truth are often limited to a very specific set of clusters. For example, the often-mentioned silhouette analysis[78] only produces an accurate scoring for the parametric family of clustering algorithms (e.g., K-means and HA-Ward)[9].

This section of the chapter can be divided in two main blocks: (1) the quantitative evaluation of the different combinations of DRM and clustering algorithms, and (2) the analysis of the noise effects and outlier detection limits on the combination UMAP-HDBSCAN.

4.6.1. Clustering and DRM combinations. Quantitative analysis.

A detailed quantitative evaluation of the different combinations of DRM and clustering algorithms is presented here. In the general framework of the

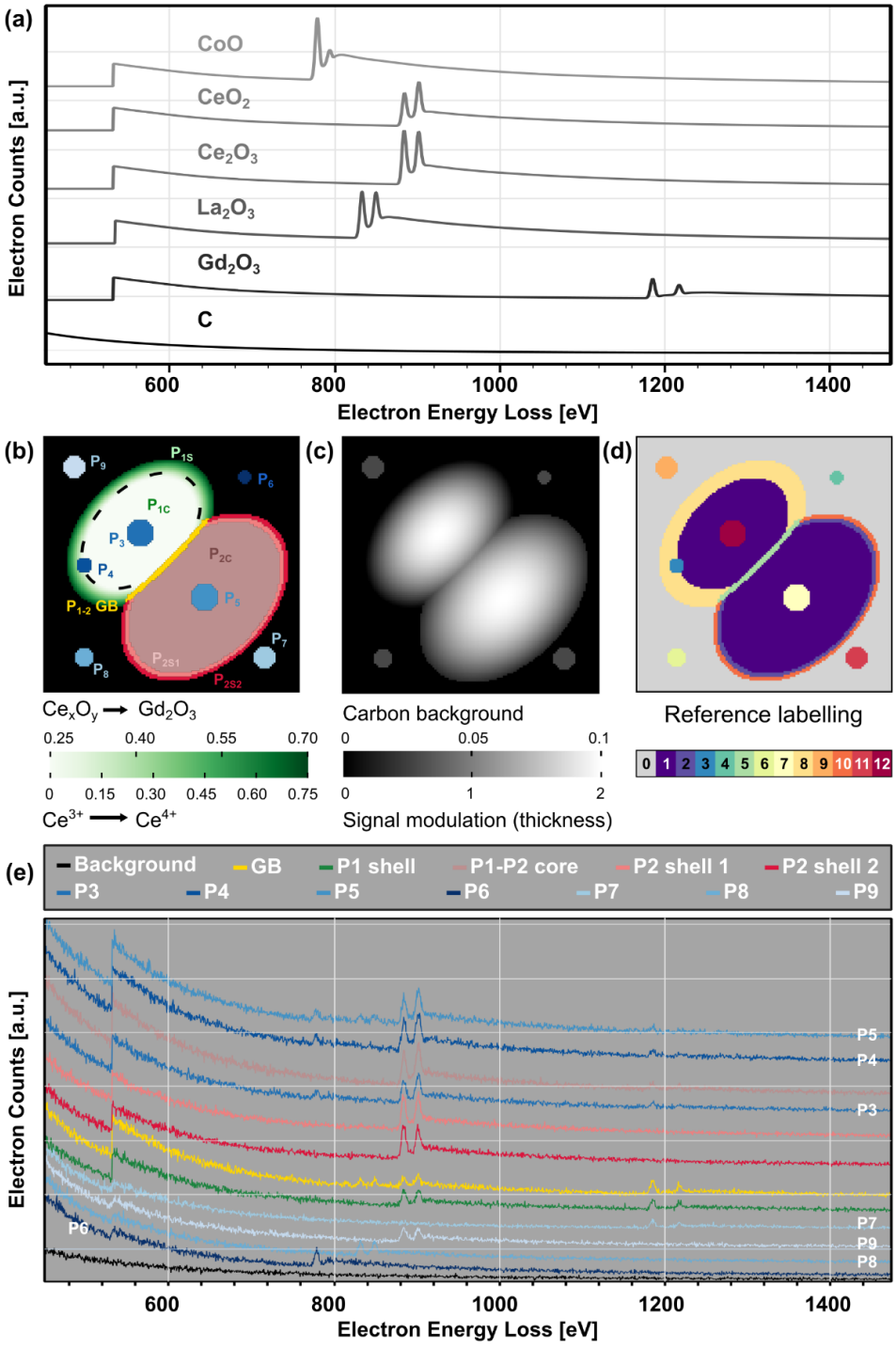


Figure 4.34. (a) Spectral references for the oxides on the mix. (b) Skeleton of the spectral phantom created. Colours determine different spectral regions. (c) Thickness and signal modulation mask. (d) Reference labelling image. (e) Single pixel spectra extracted from each of the different spectral regions in (a).

chapter (and as a natural extension, the PhD thesis) this task represents an important contribution since, for the first time, these combinations are numerically evaluated.

A synthetic dataset was specifically devised for this purpose, with the objective in mind of posing a challenging but representative case of study of the most commonly observed spectral features in an EELS SI. The focus was specifically set on reproducing those characteristics generally associated with core loss spectra acquired for transition metal (TM) and rare earth (RE) oxides. To cite some examples, the dataset contains (1) elemental gradients towards localized regions (grain boundaries or grain surfaces), (2) changes in the oxidation state of certain elements (i.e., chemical shifts and ratio variations of WLs) and (3) smaller regions of different composition scattered throughout the area of the SI.

The actual synthetic dataset, showcased in **Figure 4.34**, was created as a 128x128x2048 EELS SI. The base reference signals, linearly combined to create the spectra for each pixel, are shown in **Figure 4.34** (a). The scheme for the areas filled with different spectral mixtures is shown in **Figure 4.34** (b). Following the scheme in **Figure 4.34** (b) and using the spectral references from **Figure 4.34** (a), the phantom SI consist of:

i. Two big particles (P1 in white/green and P2 in red) with the same core composition (P1C and P2C respectively) of 75% cerium oxide (Ce4+) and 25 % gadolinium oxide,

ii. A shared grain boundary between P1 and P2 (GB, in yellow) with an increased percentage of gadolinium oxide over the cerium oxide and lanthanum oxide (1%) added to the mix.

iii. The P1 shell (P1S), that includes a linear gradient increasing the gadolinium oxide towards the surface and shifting the cerium oxidation state from Ce^{4+} to Ce^{3+}. Hence, a gradient promoting a drift from the

reference signal of CeO_2 towards Ce_2O_3. The P1C - P1S frontier is indicated with a dashed line.

iv. A dual shell structure for P2, where both parts are composed by cerium oxide and lanthanum oxide. The inner shell signal (P2S1) is fully formed by Ce4+ cerium oxide, whereas the outer shell (P2S2) contains a fully reduced cerium oxide signal (Ce3+).

v. Two smaller particles (P3 and P4) added on top of P1, introducing cobalt oxide in the spectral mix, and an even smaller one (P5) on top of P2 adding both cobalt oxide and lanthanum oxide. Notice that the

vi. majority of P4 is on top of the spectral-gradient shell of P1, whereas P3 is entirely over the core region of P1.

vii. Finally, P6, 7, 8, 9, smaller particles over the background carbon grid that have the following single oxide compositions: cobalt oxide (P6), gadolinium oxide (P7), lanthanum oxide (P8) and cerium oxide (P9, fully reduced Ce^{3+}).

These details of the spectral mixtures for the synthetic dataset are listed in **Table 4.4**.

In order to increase the difficulty of classification and mimic the real behaviour of spectra in 3D samples with a finite size affecting the EELS signal intensity, a modulation function was applied to the phantom raw-signal, and an additional non-constant background signal (carbon spectrum) was added, both using the same masking gradients (**Figure 4.34** (c)). Furthermore, Gaussian and Poissonian noise were added to the dataset. A set of various representative spectra for each region described are shown in **Figure 4.34** (e). With all these factors in mind, the expected results for any classifying algorithm should be as sketched in **Figure 4.34** d). These are the reference labelling regions that will be used to evaluate numerically the performance of the different combinations of clustering algorithms and DRM.

Table 4.4 Spectral composition of the different regions of the phantom spectrum image on **Figure 4.34**. The quantities are normalized to 1. P_4 composition are labelled as mixed, as the cerium oxide and gadolinium oxide quantities are described by the gradient shell underneath the particle (0.9 gradient + 0.1 CoO)

	CeO_2	Ce_2O_3	Gd_2O_3	La_2O_3	CoO
P_{1C}	0.75	-	0.25	-	-
P_{1S}	0.75 to 0	0 to 0.25	0.25 to 0.75	-	-
P_{2C}	0.75	-	0.25	-	-
P_{2S1}	0.95	-	-	0.05	-
P_{2S2}	-	0.95	-	0.05	-
$P_{1\text{-}2\ GB}$	0.25	-	0.65	0.1	-
P_3	0.675	-	0.225	-	0.1
P_4	Mixed 0.675	Mixed 0.675	Mixed 0.225	-	0.1
P_5	0.6		0.2	0.1	0.1
P_6	-	-	-	-	1
P_7	-	-	1	-	-
P_8	-	-	-	1	-
P_9	-	1	-	-	-

In that regard, the use of a synthetic dataset was the only way to have access to the Fowlkes-Mallow index (FMI) [79], which is one of the most versatile and accurate scoring techniques, and is generally unaffected by the nature of the algorithm evaluated The index is defined as follows

$$FMI = \frac{TP}{\sqrt{(TP + FP) \cdot (TP + FN)}} \tag{51}$$

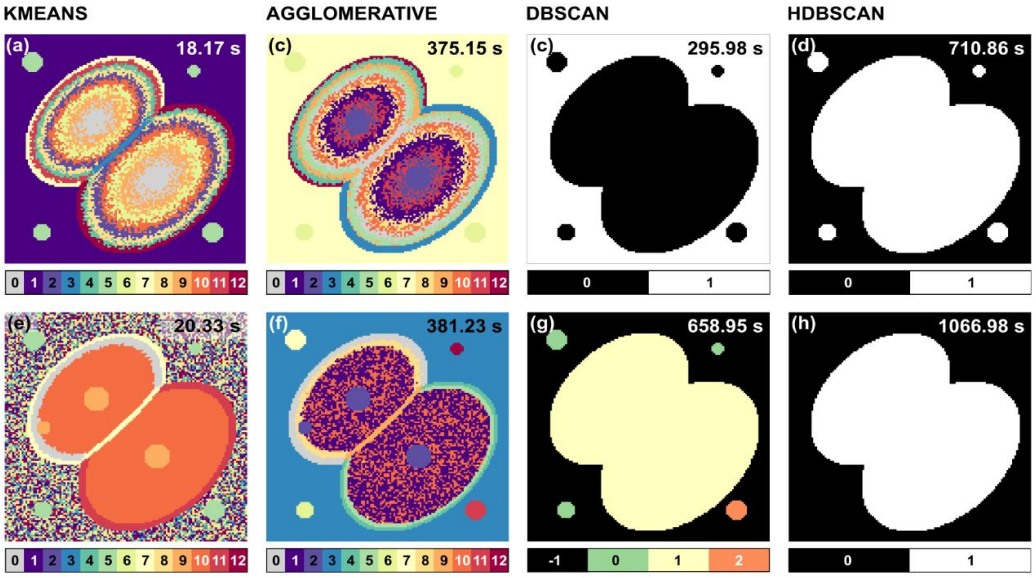


Figure 4.35. (a-d) KM, HA-Ward, DBSCAN and HDBSCAN clustering labels for the raw data of the phantom dataset. (e-h) KM, HA-Ward, DBSCAN and HDBSCAN clustering labels for the L_2-normalized data of the phantom dataset.

where TP are the number of true positives, FP are the number of false positives and FN are the total number of false negatives. Notice, that to get the TP, FP or FN numbers, the ground truth has to be previously known (see in **Figure 4.34** (d)). Hence, all the work done configuring this synthetic dataset, instead of using an experimental one. Also, the problem is now understood as a supervised classification one.

Regarding the clustering algorithms tested on this phantom SI, one from each category described in **Table 4.3** was selected: (1) KM, (2) HA-Ward, (3) DBSCAN and (4) HDBSCAN. This clustering classification was done for the raw and L_2-normalized phantom dataset. Furthermore, the combinations of (1) NMF-clustering, (2) UMAP (raw dataset)-clustering and (3) NMF-UMAP (raw dataset)-clustering, were also included in the analysis.

The results for the clustering classification of the raw dataset and L_2-normalized dataset are shown in **Figure 4.35**. The algorithms applied on the raw dataset (**Figure 4.35** (a-d)) perform in the exact same way as they did for the experimental sample in the previous section, showcased throughout

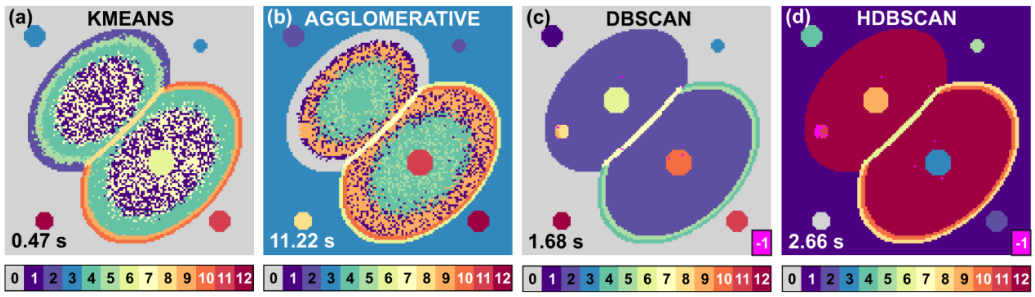


Figure 4.36. Clustering results over the non-negative matrix factorization dimensionality reduced datasets. (a) K-Means labelling. (b) Hierarchical agglomerative clustering label results. (c) DBSCAN resolved labels. (d) HDBSCAN labels. Notice the -1 labels (magenta) for the DBSCAN and HDBSCAN noise datapoints.

Figure 4.14, **Figure 4.19**, **Figure 4.22** and **Figure 4.27**. The parametric algorithms (KM and HA-Ward) classified pixels mainly according to the L_2-norm value of the spectra in each pixel-point. Hence, it can be observed in **Figure 4.35** (a-b) how the cluster labels are organized in concentric rings that follow the bigger particles contours, and the 4 external NPs over the carbon grid are classified as part of the same cluster. On the contrary, the density-based algorithms (**Figure 4.35** (c, d, g, h)) performed really poorly compared to the parametric ones. The cause resides in the curse of dimensionality, already explained when analysing the experimental dataset in the section describing the algorithms.

On the L_2-normalized dataset, some extra information is retrieved from KM and HA-Ward clustering **Figure 4.35** (e-f). They successfully identify $P_{3,4,5}$ (although they are grouped together), they separate the core and shell structure of P_1 and P_2 (although the hierarchical clustering performs better as it separates both shells in P2 as individual entities), they separate P6, 7, 8, 9 from the C-background (although KM groups them together) and they identify the region of GB (although KM includes in the same cluster the GB and the outermost part of P1S). This behaviour is consistent with what was observed in the previous section for the L_2-normalized EELS SI (**Figure 4.16** and **Figure 4.18**)

The results for the NMF-clustering combination are showcased in **Figure 4.36**. This combination was not tested for the experimental dataset in the previous section. Hence, a qualitative analysis is provided now. Twelve dimensions where originally chosen to represent the whole dataset, and the Frobenius distance was chosen as the objective function for the minimization in the NMF decomposition (eq.(6)). For a fully data-driven analysis all the components resolved by the NMF decomposition were included in the clustering analysis. For all the 4 different clustering algorithms, the Euclidean metric was selected. No L_2-normalization was done neither on the raw dataset, nor over the loadings resolved (i.e., the thickness related structure, differences in signal depending on the different regions and the variance information of the NMF components was kept for the clustering algorithms).

From **Figure 4.36** (a-b) it is observable that KM and HA successfully separated the two different shells for P2, hinted the gradient structure of P1S, resolved the GB and P5 and separated P6, 7, 8, 9 from the background (although identifying them solely as 3 clusters instead of 4). Nevertheless, the core regions of P1 and P2 are subdivided into several clusters (driven by noise and thickness modulation), and P3 and P4 are not correctly resolved.

On the other hand, both DBSCAN and HDBSCAN perform well, and are able to identify all the cluster regions excepting P1S. Some of the datapoints in the SI are labelled as noise (cluster -1, in magenta), as the algorithms cannot decide the cluster they belong with the specified parameters. Most of them belong to P4, as its position overlapping the gradient shell makes it a challenging area to resolve.

The combination of manifold approximation algorithms and density based clustering analysis was proven successful in the previous section for an experimental sample, whether the UMAP embedding was calculated from the raw data (**figure 4.28** and **figure 4.29**) and from the NMF decomposition results (**figure 4.31** and **figure 4.32**). For the phantom SI, the results of the UMAP embedding of the raw datasets are showcased in **figure 4.37**.

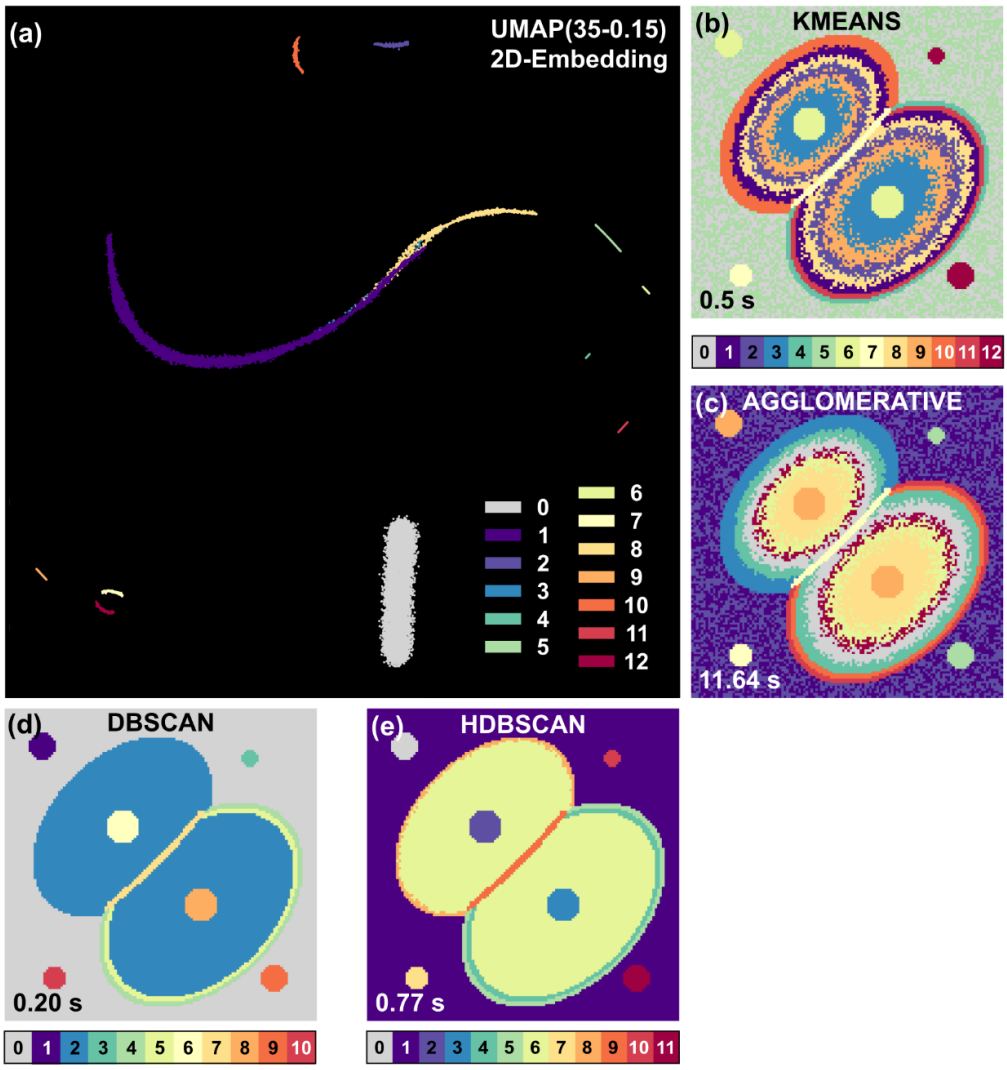


Figure 4.37. (a) 2D embedding resolved by UMAP for the spectral phantom raw data. The colour scheme follows the one for the ground truth shown in **Figure 4.34**(d). (b) K-Means, (c) hierarchical agglomerative, (d) DBSCAN and (e) HDBSCAN clustering labelling.

Analysing the UMAP embedding in **figure 4.37**(a), one can observe that the datapoints for the P1C and P2C (purple, label-1) form a continuous line with the gradient shell P1S (yellow-orange, label-8), and how the datapoints for P4 (blue, label-3), although hardly visible, appear trapped in an intermedium region between them. This is a clear effect caused by the thickness and signal modulation functions introduced in the spectral phantom, that closely links these areas together. The GB points (light green, label-5) follow the same trace, but are separated from the main line. The rest of the regions are also well

separated, and thus one may expect the clustering algorithms to perform well and separate all areas excepting P1S and P4. The results for the density-based algorithms DBSCAN and HDBSCAN confirm this hypothesis (**figure 4.37**(d),(e)), being the later able to even separate a small region of the shell P1S. On the other hand, KM and HA (**figure 4.37**(c)) recover a similar clustering structure to that of the raw-data case. The Gaussian ball assumption and the Euclidean distance metric are to blame once again.

The results of the UMAP embedding for the 12 NMF components of the phantom SI are showcased in **Figure 4.38**. The resulting 2D-embedding is shown in **Figure 4.38** (a). This time, the shell region P1S is clearly separated from the main body P1C, although some of the shell datapoints can be seen clustered with the core region structure. Both BDCAN and HDBSCAN (**Figure 4.38** (d-e)) performed almost perfectly, being the main difference that the datapoints classified as noise by the hierarchical one (-1 label in magenta) are mainly restraint to P4, and the ones in DBSCAN are slightly scatter over the whole SI. As with the case of using the raw-data UMAP embedding, KM and HA (**Figure 4.38** (b-c)) provided worse results than DBSCAN and HDBSCAN. Again, they tend to subdivide the core and background regions forced by the parametric assumptions that they make on the spatial distribution of datapoints.

For the sake on completion, the UMAP embedding resolved for the 12 components of the NMF decomposition (shown already in **Figure 4.38** (a)) is presented 4 times again in **Figure 4.39** coloured by the labels resolved for each of the clustering algorithms under study. The parametric nature of the KM and HA-Ward algorithms is directly observable in **Figure 4.39** (a) and (b), as they assume that the datapoints were drawn from a Gaussian distribution (a 2D ball). Hence, the algorithms try, by all means, to group the datapoints in radially symmetric clusters. The different sizes and shapes of the natural groups of datapoint formed by UMAP are better resolved by the density-based algorithms DBSCAN (**Figure 4.39** (c)) and HDBSCAN (**Figure 4.39** (d)).

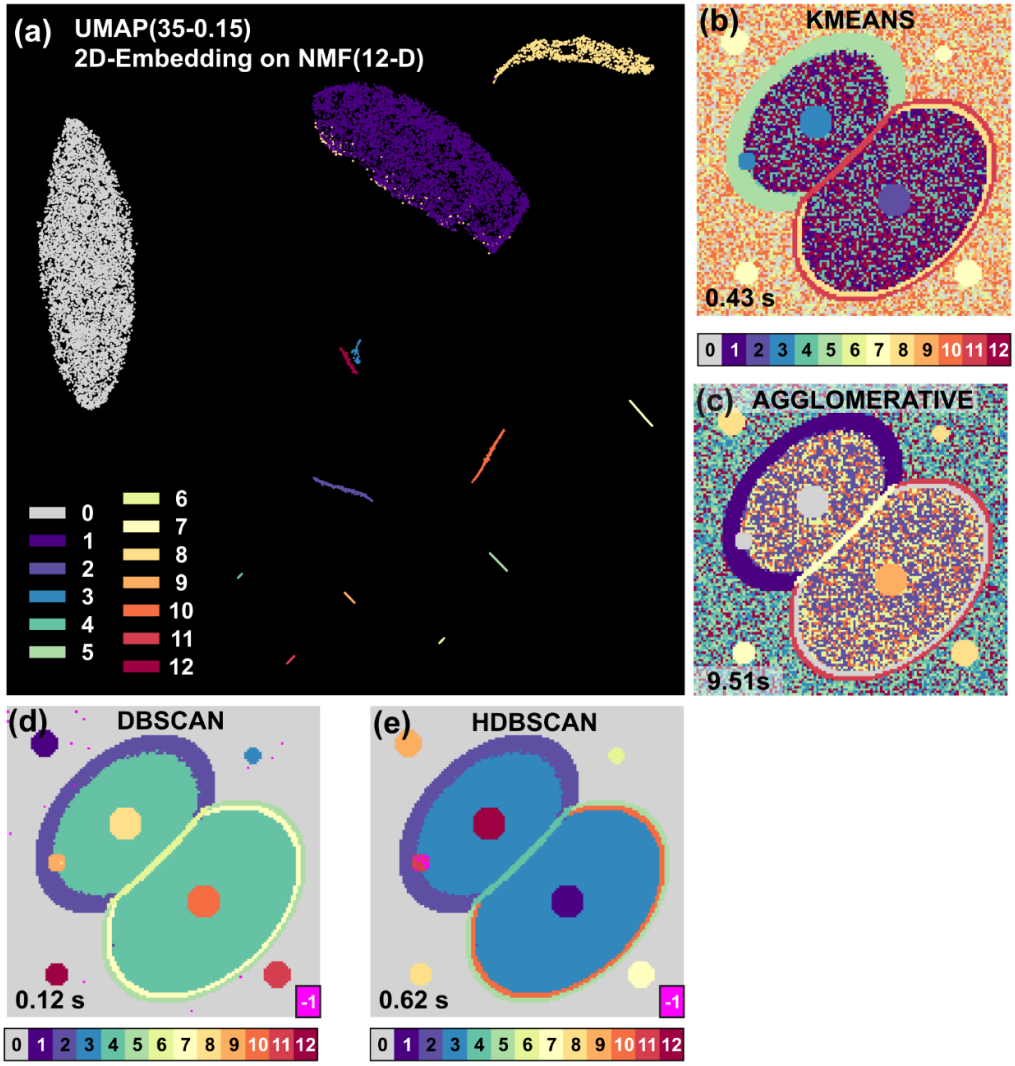


Figure 4.38. (a) 2D embedding resolved by UMAP for the spectral phantom using the resulting NMF dimensionally reduced components. The colour scheme follows the one for the ground truth shown in Figure 4.34(d). (b) K-Means, (c) hierarchical agglomerative, (d) DBSCAN and (e) HDBSCAN clustering labelling.

Finally, given that the ground truth class assignments ('true' labelling) of the samples was known (i.e., the dataset analysed was the synthetic phantom from **Figure 4.34** and the classes are inferred from the spectral mix of each area in the SI), the performance could be evaluated from the Fowlkes-Mallows index (i.e., the clustering performance evaluation is done as in a supervised learning scenario). The results are shown in **Figure 4.40**.

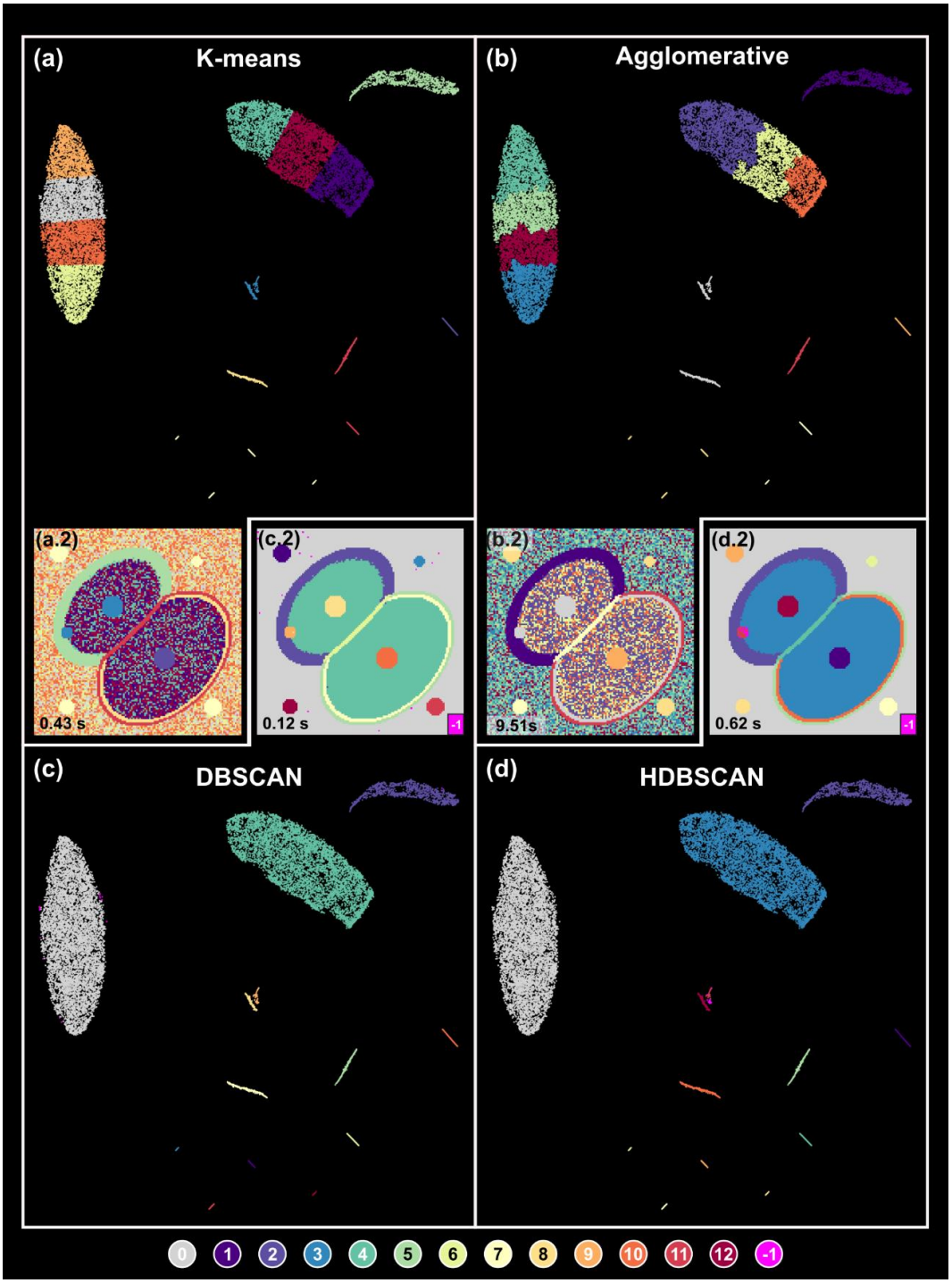


Figure 4.39. UMAP embedding for the 12 NMF components coloured according to the labelling resolved by the (a) KM, (b) HA-Ward, (c) DBSCAN and (d) HDBSCAN clustering algorithms. The resulting label maps are included as insets for the KM (a.2), HA-Ward (b.2), DBSCAN (c.2) and HDBSCAN (d.2) clustering algorithms. The colour scheme is shared by all the clustering results, but the -1 (magenta) tag only applies to the density-based ones (DBSCAN and HDBSCAN).

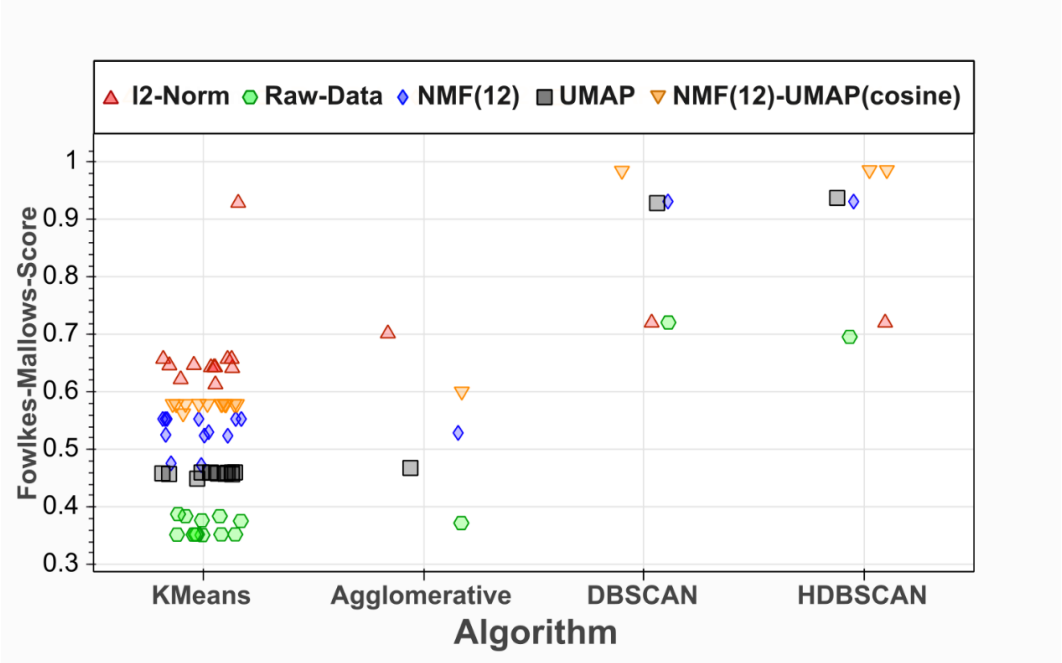


Figure 4.40 Fowlkes-Mallows-Index for the label-scores resolved by the different clustering algorithms, depending on the dimensionality reduction pre-treatments performed.

The scores are grouped by algorithm, and the different markers correspond to different dimensionality reduction techniques. The 'jitter' introduced in the horizontal axis distribution helps with the visualization, as it prevents excessive overlapping of the indexing scores markers.

The scores presented here disregard the labels corresponding to the C-background in the phantom SI. Being the larger group in number of datapoints and having a large spectral and L_2-norm deviation from the rest of the dataset, it was a region easily resolved by every combination of clustering-DRM. Thus, it introduced a positive bias in the scores that outweighed the differences between the ground truth and the clustering labels for the rest of the image (i.e., it lowered the sensitivity to small deviations from the ground truth in the smaller clusters).

KM is the only one showing several scores in each DRM category, the reason being that the sklearn implementation of KM produce undetermined results. As the clustering is randomly initialized, setting the centroids as random

points of the dataset (pixels), different runs are prone to produce different results (convergence to local minima is a well-known effect in KM). Depending on the dataset itself, these types of behaviours may be amplified. As a matter of fact, for the L_2-normalized dataset (red triangles) one of the scores is close to 1. Precisely that score corresponds to the labelling result shown in **Figure 4.35** (e). Apart from the behaviour of that outlier, KM presents the lower performance overall, something that was expected from the qualitative analysis of the labelling maps. HA scores are close to those of KM for every DRM (both algorithms make the same Gaussian ball assumption on the data space).

DBSCAN and HDBSCAN performed almost identically in all cases, being the average score much higher than those of KM and HA. Regarding the scores of the raw-data and L_2-norm data, they are clearly biased by the large number of points of the P1-P2 cores, belonging to the same base cluster (i.e., the performance evaluated from the score indicates a much higher clustering quality than that of the qualitative analysis from **Figure 4.35**, as it happened when including the labelling for the C-background). Besides, any combination of DRM and density-based clustering scores a good value (higher than 0.9 in any all cases).

4.6.2. UMAP-HDBSCAN combination. Noise resilience and outlier detection limits.

The previous sections demonstrated the high potential that the combination of UMAP and HDBSCAN has to produce an accurate segmentation map for EELS datasets. Even in the case of the experimental Fe-Mn oxide demonstrator, with a fairly low signal to noise ratio, the main spectrally relevant structures within the dataset itself were uncovered. However, a systematic analysis of the resilience of this combination of algorithms to noise, and their capacity to deal with extreme fringe cases of small outlier spectral regions (e.g., a small pocket of dopants segregated in a very localized region of an EELS SI) is still missing.

As with any other algorithm (or combination of algorithms in this case) devoted to EELS data analysis, these are important parameters to be discussed.

In order to do so, a new synthetic dataset was devised, given that the previous phantom (**Figure 4.34**) was too complex in its spectral nature. So complex indeed that it hindered the evaluation of noise effects and outlier detection capabilities.

This new (and simpler) synthetic dataset is showcased in (**Figure 4.41**). It only contains a mixture of 3 oxides (CeO_2, La_2O_3, and CoO) and limits the energy range of the spectra to the proximity of the cationic edges (see **Figure 4.41** (b)). The exclusion of the oxygen K edge is intentional, as it helps to mitigate the dominant weight that it may have over the minor spectral features introduced here (e.g., WLs continuous shifts).

The whole SI is (100x256x2048), and is constituted by 3 distinctive spectral regions:

i. The top-half part of the SI is filled with pure cerium oxide. It presents chemical shift for the WLs of Ce towards higher energy loss values from left to right, amounting to a total displacement of 3eV (see the colorbar in at the bottom of **Figure 4.41** (a))

ii. The bottom-half part of the SI presents a continuous compositional variation from pure CeO_2 to pure La_2O_3 (see the colorbar at the left in **Figure 4.41** (a)). It still maintains the same Ce WL chemical shift towards higher energy loss values described for the upper-part region.

iii. A small region of 9 pixels containing a trace quantity of CoO over the CeO_2 background is set over the top-half regions of the SI (from the total number of 25600 pixels, a clear outlier region).

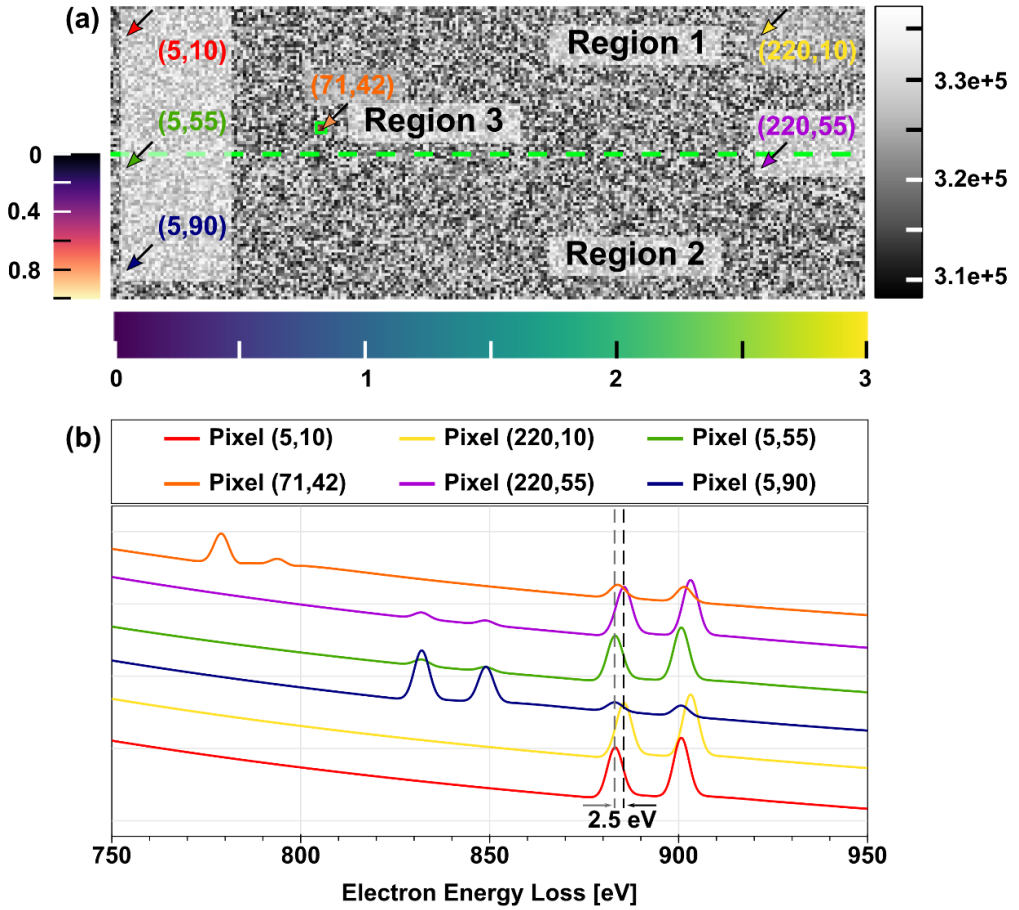


Figure 4.41. (a) Synthetic spectrum image for the new phantom dataset. The colour-bar at the right is the integrated intensity per pixel. The colorbar in the bottom indicates the energy shift of the CeO_2 white lines. The one at the left indicates the compositional gradient exchange between Ce and La oxides form the middle down. (b) Reference spectra for the pixels highlighted in (a).

Permeating the whole SI, a randomly varying C background signal has been added. The chemical shift in CeO_2, the elemental composition changes from pure CeO_2 to pure La_2O_3 and the outlier CoO inclusion in region 3 are easily observed. in the spectra **Figure 4.41** (b), extracted from the random set of pixels selected and marked in **Figure 4.41** (a).

The two different levels of noise tested in this section are showcased in **Figure 4.42**. The spectra shown in this figure correspond to the same set of pixels highlighted **Figure 4.41** (a) in and showcased in **Figure 4.41** (b). The noise introduced on top of the reference signals is pure Gaussian white noise. The

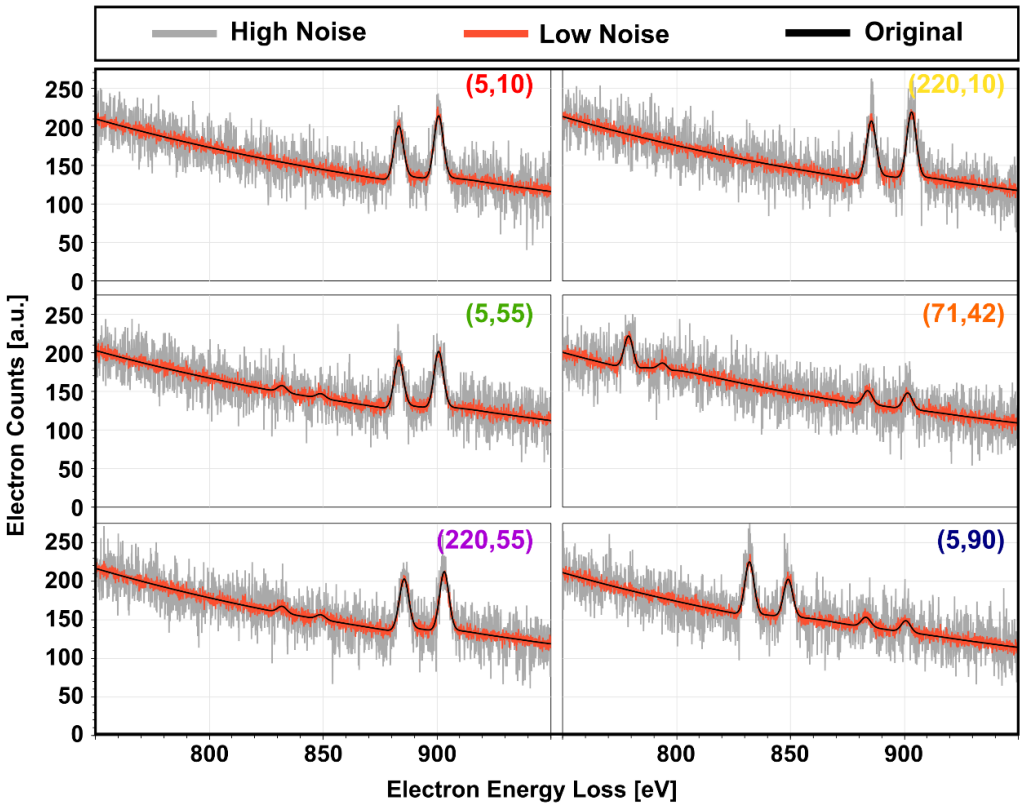


Figure 4.42. Reference spectra (black line) for the pixels highlighted in **Figure 4.41** (a) overlayed with spectra containing 2 different levels of noise: high level (in grey) and low level (in orange).

two different levels amount to 1/6 (low-level) and 1/2 (high-level) of the maximum peak-to-valley distance (in electron counts) measured for every ELNES feature (WL) for the whole SI. Notice how the high-level noise case (grey) almost opaque completely the CoO outlier feature (see pixel (71,42)).

The UMAP projections for the high-level and low-level of noise datasets are showcased in **Figure 4.43**. Both are colour coded according to the vertical dimension (y-axis) of the embedding plane, although this dimension does not convey any direct physical meaning. The reason for the selection of this colour scheme is directly related with the interpretability of the UMAP projection results.

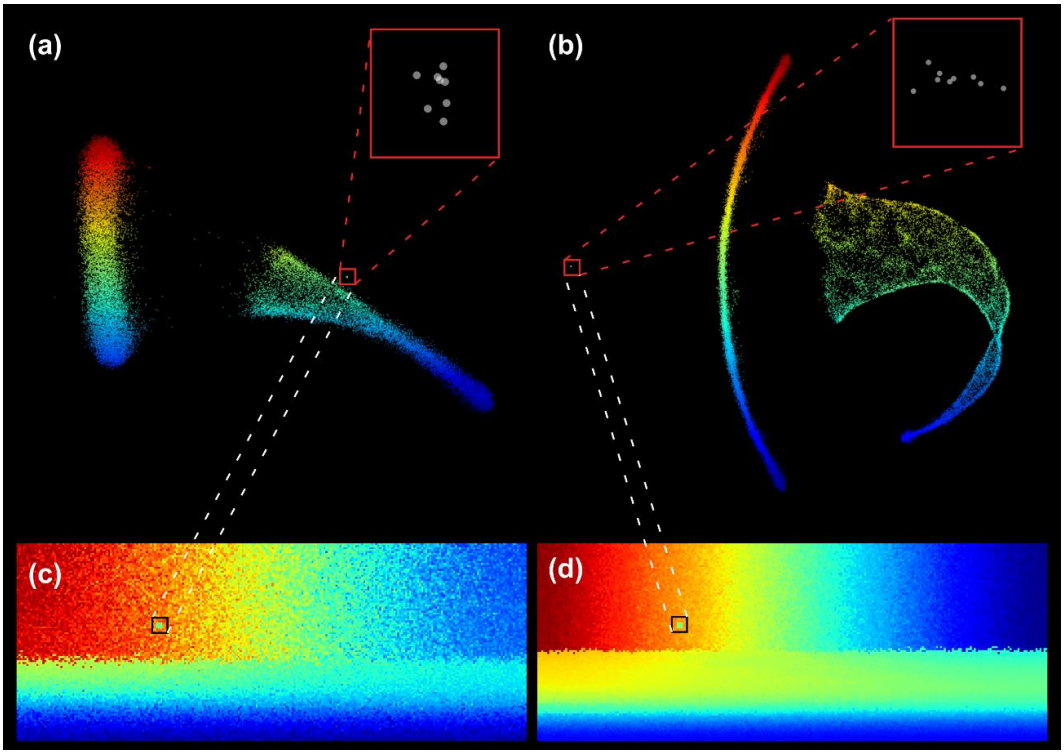


Figure 4.43. UMAP embeddings for the SI introduced in **Figure 4.41**, for two different levels of noise showcased in **Figure 4.42** (a) high-level (1/2) of noise case (b) low-level (1/5) of noise case. The zoomed areas of the insets highlighted in red correspond to the outlier pixels separated by the UMAP projection. The SI areas at the bottom are coloured according to their corresponding UMAP embedding on top. (c) SI area colour-coded as the UMAP embedding in (a), high-level of noise case. (d) SI area colour-coded as the UMAP embedding in (b), low-level of noise case.

As already mentioned in the theoretical description in previous sections, UMAP first computes the hypothetical underlying manifold from the dataset (i.e., EELS SI) itself (i.e., from the spectral composition of each pixel). It later resolves the manifold projection to maintain as much of the local and global structure as possible (i.e., to project the spectral relation between datapoints into the lower dimensional embedding). Given that all the physical information from the EELS SI is encoded into the high-dimensional manifold computed, the final projection space retains it as well. The distribution (or formation) of clusters of datapoints (if any) by UMAP and the inner structure within them is the mechanism by which the physical information of the system is encoded in the lower dimensional embedding. This was already tested to a certain degree

in the supervised dimensionality reduction experiment carried out for the synthetic 'Swiss-roll dataset', showcased in **Figure 4.8** and **Figure 4.9**.

Hence, although it is a challenging task to directly interpretate (or even validate) the projected UMAP embeddings, a promising proxy consist of colour-coding them according to some heuristic criteria (e.g., maximum PCA loading value for each pixel from a previous decomposition, or even something as simple as giving a colour proportional to the random x or y axis position in a hypothetical 2D projection). The translation of this colour-code back to the original SI might help in some cases to identify certain structures in the UMAP projection, and might provide an insight to the physical (spectral) information encoded in it.

Back to the case of **Figure 4.43**, the selection of the y-axis value to colour code the UMAP embeddings is done by convenience, as some of the structures resolved by the manifold are precisely oriented in the y-axis direction.

Notice how the WL gradient included in the dataset (see **Figure 4.41** (a)) is represented by elongated tube-like structures to the left of the projection plane (in both cases **Figure 4.43** (a) and (b)). Meanwhile, the lower part of the SI (that contains both the WLs and elemental composition gradients) produces a clear ribbon-like structure in the lower noise case **Figure 4.43** (b), and an asymmetric opened tube-like structure in the higher noise level dataset **Figure 4.43** (a).

The effects that different noise levels produce over the UMAP projections are also clearly showcased here. In general, a decreased signal to noise ratio will be translated into 'fuzzier' UMAP embedding projections (**Figure 4.43** (a)), as the direct relations between the spectra of the datapoints are partially opaqued (i.e., the measurements used to compute the hyperspectral manifold are influenced by the noise levels). This will result in a certain level of loss of information within the structures resolved (i.e., a loss of physical interpretability of the projections). Hence, the ribbon-like projection with a lot of internal structure shown for the low-noise level case in **Figure 4.43** (b)

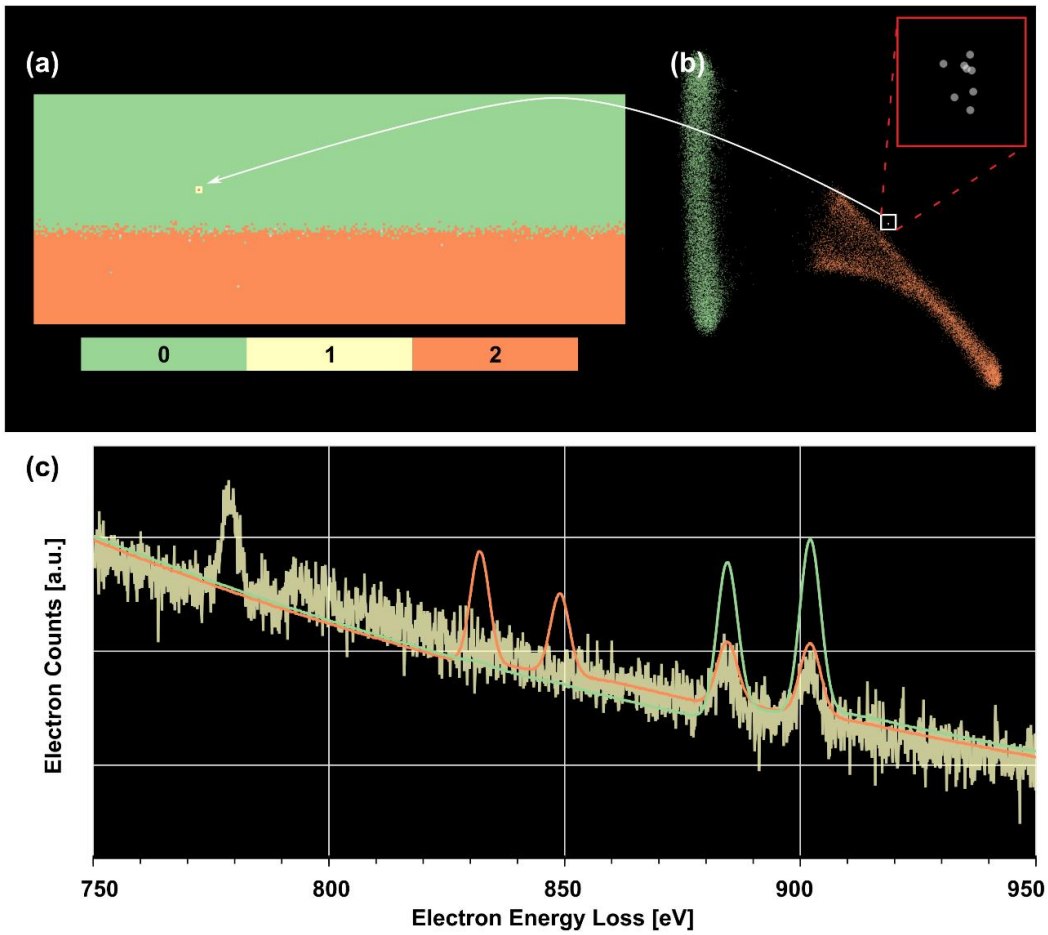


Figure 4.44 (a) HDBSCAN clustering label map of the UMAP embedding for the high-level noise case in **Figure 4.43** (a). (b) UMAP embedding for the high-level noise case coloured according to the clustering label map resolved by HDBSCAN in (a). (c) Average EELS signal values for the clusters resolved in (a).

becomes the fuzzier opened tube-like structure in the high-level noise case in **Figure 4.43** (a). These noise levels are also responsible for the smoother distribution of colour shown in the lower noise-level case (**Figure 4.43** (b) and (d)).

The projected embeddings resolved by UMAP appear to have separated in both noise-level cases the outlier region with CoO. However, the higher noise level case only contains 8 from the original 9 datapoints (red inset of **Figure 4.43** (a)), whereas the lower noise level case contains of them (red inset of **Figure 4.43** (b)). This certainly affects the segmentation maps later produced by HDBSCAN.

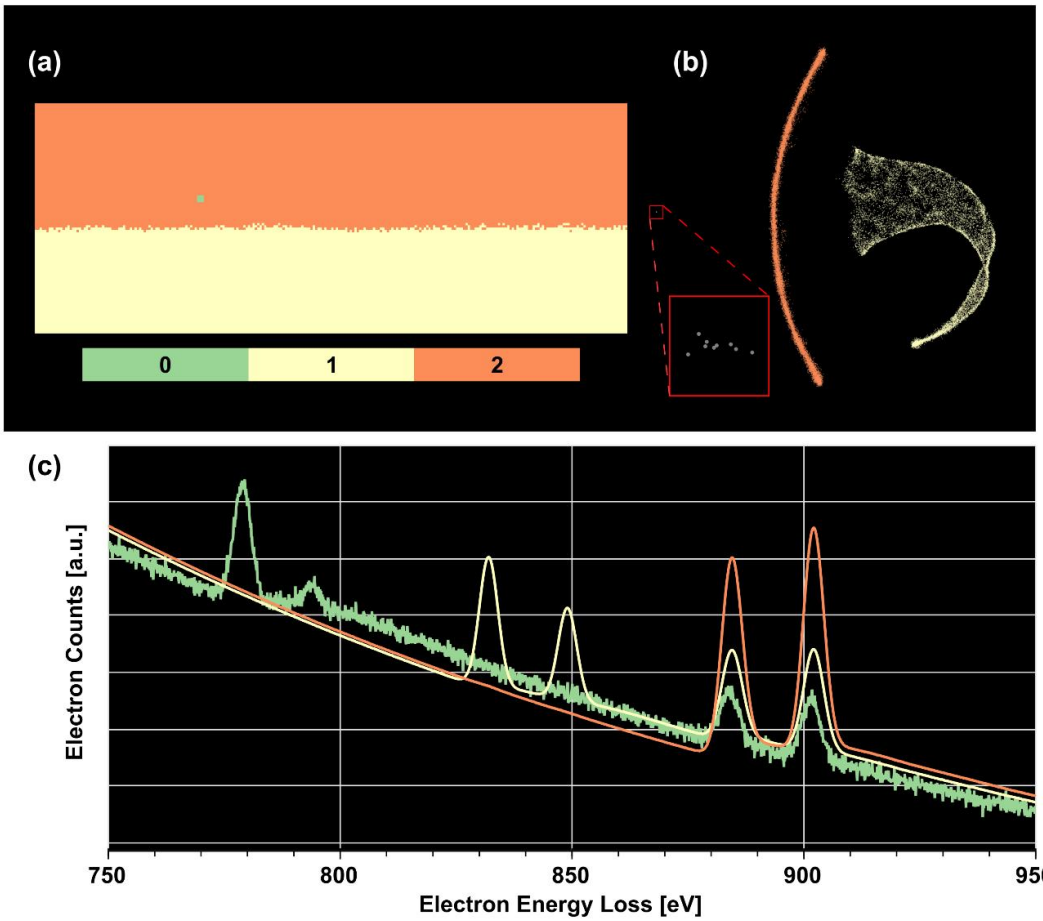


Figure 4.45(a) HDBSCAN clustering label map of the UMAP embedding for the low-level noise case in **Figure 4.43** (b). (b) UMAP embedding for the low-level noise case coloured according to the clustering label map resolved by HDBSCAN in (a). (c) Average EELS signal values for the clusters resolved in (a).

The HDBSCAN clustering results for the high-level noise are showcased in **Figure 4.44**. The label map in **Figure 4.44** (a) shows how the central pixel in the outlier CoO region (mainly belonging to cluster 1, in yellow) is actually ascribed to cluster 2 (orange), which is turn is mainly composed by datapoints from the La - Ce oxide mixture area (region 2). Notice also how the boundary area for the regions 1 and 2 from **Figure 4.41** (a) is displaced towards the bottom part of the SI. This is clearly another effect that the high level of noise has over the UMAP projection embedding resolved. The La oxide inclusion appears to be undetected up to a certain threshold level at which it clearly overcomes the Ce oxide signal. Notice also how the average EELS spectra (i.e.,

cluster reference signals, **Figure 4.44** (c)) for clusters 0 and 2 (regions 1 and 2 from **Figure 4.41** (a)) are perfectly smooth, as the Gaussian white noise is naturally averaged to 0 for the large number of pixels considered. The outlier cluster on the other hand (number 1), only contains 8 pixels and, thus, the Gaussian noise is still visible in the average reference signal. Nonetheless, both Co and Ce fine structures are observable in the EEL spectrum.

The HDBSCAN clustering results for the low-level noise embedding (**Figure 4.43** (a)) are showcased in **Figure 4.45**. This time around, all 9 pixels from the outlier region are separated in their own cluster (cluster 0 in green). Also, the frontier between regions 1 and 2 (clusters 2, orange, and 1, yellow, respectively, comparing **Figure 4.41** (a) and **Figure 4.45** (a)) is closer to the middle line dividing the SI than the frontier in the high-level noise case (**Figure 4.44** (a)). The reason lies again in the detection limit for the La oxide inclusion. As the noise level is decreased, the La WLs are easily distinguishable from the background Gaussian noise signal.

To close this section, the NMF decomposition results for both high and low noise level cases are included (**Figure 4.46** and **Figure 4.47** respectively).

This is done in order to qualitatively compare the performance of the UMAP-HDBSCAN combination (already showcased through **Figure 4.43**, **Figure 4.44** and **Figure 4.45**) with a most commonly extended methodology for analysis based on a matrix factorization algorithm.

Although the high-level noise case (**Figure 4.46**) seems to identify the CoO outlier region in the loading mappings for components 1, 2, 3 and 4, the archetype signals do not show any trace of the CoO fine structure. Thus, one might erroneously interpret that those regions correspond to anomalies in the Ce oxide composition, failing to identify the significance of their presence with the CoO dopant contribution to the signals. Only 6 NMF components are displayed, from the objective number of 12 selected. The absent 6 other components (6 to 11) are not included because they corresponded to random background noise, indistinguishable from that of component number 5.

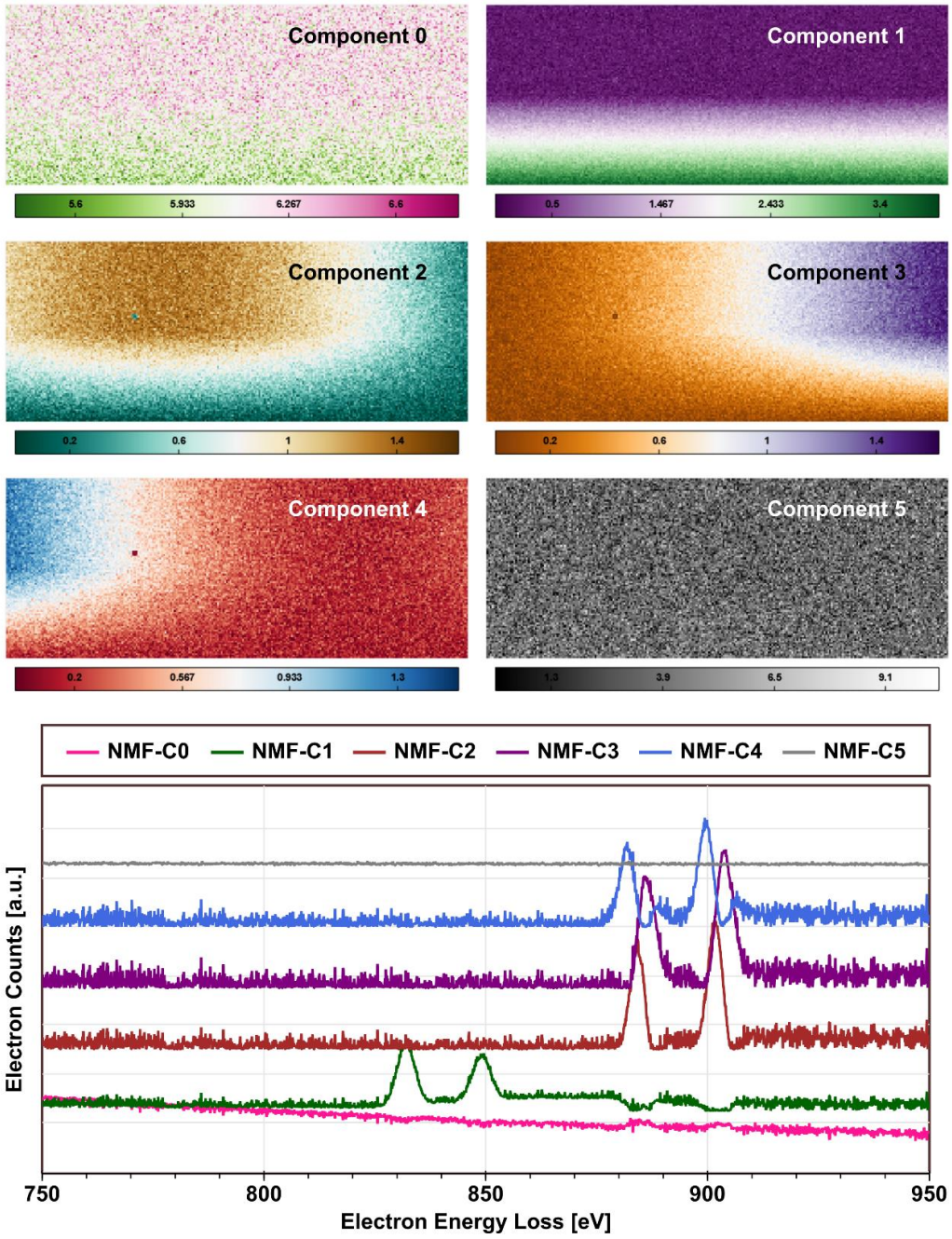


Figure 4.46. Loadings and archetypes for 6 of the 12 components resolved by NMF for the high-level noise dataset from **Figure 4.41** and **Figure 4.42**.

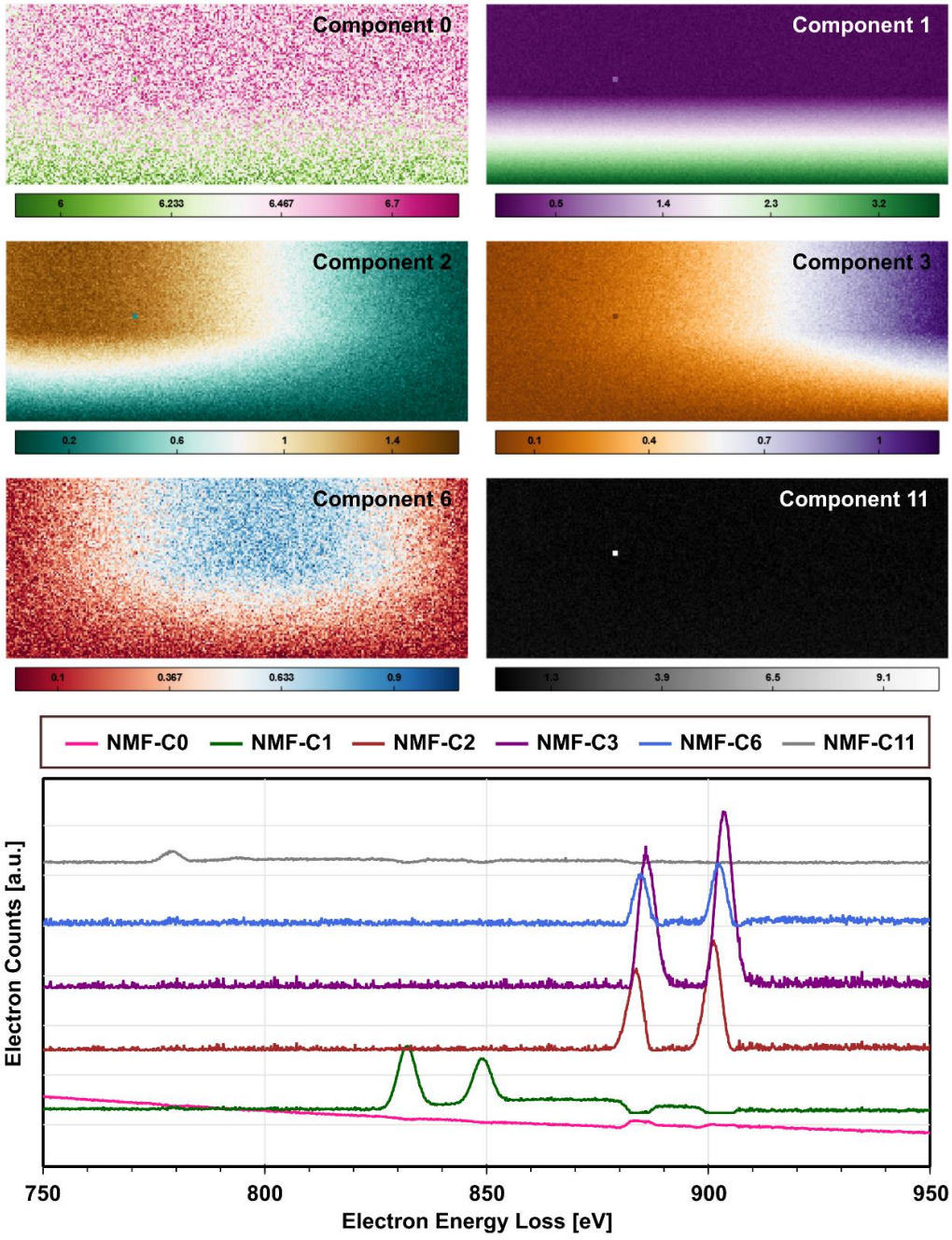


Figure 4.47 Loadings and archetypes for 6 of the 12 components resolved by NMF for the low-level noise dataset from **Figure 4.41** and **Figure 4.42**.

Regarding the low-level noise dataset NMF decomposition (**Figure 4.47**), the CoO is now correctly identified, as one specific archetype (factor) signal (component 11) includes the CoO ELNES features.

In any case, both NMF decompositions exemplify the same 2 underlying problems when using this DRM: (1) the archetypes present non-physical and usually overlapping spectral features, that (2) force the user to provide an interpretation of the results that can be easily biased by the previous knowledge about the sample composition.

Finally, the UMAP-HDBSCAN seem to produce a better-defined spectral segmentation, with an easier interpretation and overall better outlier detection and noise resilience capabilities than the NMF decompositions.

4.7. Conclusions.

This chapter presented an exhaustive revision of the major dimensionality reduction techniques and clustering algorithms utilized throughout this thesis, including the two latest state-of-the-art algorithms introduced to EELS data analysis (one from each main category listed): UMAP and HDBSCAN.

Every algorithm was described from a theoretical point of view, and later showcased by analysing an experimental sample of known composition. Furthermore, a quantitative scoring analysis was carried out with the help of a synthetic EELS dataset, where the ground truth for the classification was known by design. The most common combinations of algorithms (dimensionality reduction and clustering) were tested to explore the better routes towards a good clustering classification.

Overall, this comparative analysis carried out for the different combinations of algorithms highlighted the importance of clearly understanding the background and formulation for each algorithm and the possible statistical assumptions they do about the underlying structure of the analysed data (e.g., the parametric Gaussian ball assumption of KM and HA-Ward). This is

especially true if one expects to extract meaningful information about the spectral structure, avoiding the risk of heavily biased results. Regarding the new additions, UMAP and HDBSCAN outperformed all the other options in their respective classes. Not only they scored better for the synthetic dataset test, but their combination also provides the closer approach to a fully data-driven and unsupervised classification methodology. Also, this combination did a better job of separating outlier regions with a small number of representative spectral cases than the usual linear dimensionality reduction approaches (e.g., NMF), even for cases of severe noise poisoning of the signal.

As a final cautionary note, one must always remember that a clustering classification is an approximation. The goodness of any given clustering classification is always open for debate. For instance, a less accurate but faster solution may be the desired outcome in some specific cases (e.g., analysing large datasets). The key aspects of a truly successful clustering analysis of an EELS dataset are an actual knowledge of each algorithm innerworkings, their shortcomings and the interpretation of the results accordingly. At the end, the obtention of truly quantitative results for an EELS dataset would further require the use of other characterization analysis methods, such as MLLS of NLLS.

The general guidelines given in this work might help to further accelerate and further spread the use of clustering analysis techniques among the TEM-EELS data community.

4.8. References.

[1] C.R. Booth, P.E. Mooney, B.C. Lee, M. Lent, A.J. Gubbens, K2: A Super-Resolution Electron Counting Direct Detection Camera for Cryo-EM, Microsc. Microanal. 18 (2012) 78–79. https://doi.org/10.1017/S1431927612002243.

[2] M. Sun, C.M. Azumaya, E. Tse, D.P. Bulkley, M.B. Harrington, G. Gilbert, A. Frost, D. Southworth, K.A. Verba, Y. Cheng, D.A. Agard, Practical considerations for using K3 cameras in CDS mode for high-resolution and high-throughput single particle cryo-EM, J. Struct. Biol. 213 (2021) 107745. https://doi.org/10.1016/j.jsb.2021.107745.

[3] M. Kuijper, G. van Hoften, B. Janssen, R. Geurink, S. De Carlo, M. Vos, G. van Duinen, B. van Haeringen, M. Storms, FEI's direct electron detector developments: Embarking on a revolution in cryo-TEM, J. Struct. Biol. 192 (2015) 179–187. https://doi.org/10.1016/j.jsb.2015.09.014.

[4] J.L. Hart, A.C. Lang, A.C. Leff, P. Longo, C. Trevor, R.D. Twesten, M.L. Taheri, Direct Detection Electron Energy-Loss Spectroscopy: A Method to Push the Limits of Resolution and Sensitivity, Sci. Rep. 7 (2017) 1–14. https://doi.org/10.1038/s41598-017-07709-4.

[5] P. Longo, J.L. Hart, A.C. Lang, R.D. Twesten, M.L. Taheri, Advantages of Direct Detection and Electron Counting for Electron Energy Loss Spectroscopy Data Acquisition and the Quest of Extremely High-Energy Edges Using Eels, Microsc. Microanal. 23 (2017) 60–61. https://doi.org/10.1017/S1431927617000988.

[6] P.J. Thomas, J. Scott, M. MacKenzie, S. McFadzean, J. Wilbrink, A.J. Craven, Near-simultaneous core- and low-loss EELS spectrum-imaging in the stem using a fast beam switch, Microsc. Microanal. 12 (2006) 1362–1363. https://doi.org/10.1017/S1431927606065512.

[7] P. Longo, R.D. Twesten, Fast STEM Spectrum Imaging Using Simultaneous EELS and EDS, Micros. Today. 21 (2013) 28–33.

https://doi.org/10.1017/s1551929512000909.

[8] F. Provost, T. Fawcett, Data Science and its Relationship to Big Data and Data-Driven Decision Making, Big Data. 1 (2013) 51–59. https://doi.org/10.1089/big.2013.1508.

[9] J. Han, M. Kamber, J. Pei, Data Mining : Concepts and Techniques : Concepts and Techniques (3rd Edition), Elsevier, Amsterdam, 2012.

[10] L. Galluccio, O. Michel, P. Comon, Unsupervised clustering on multi-component datasets: Applications on images and astrophysics data, Eur. Signal Process. Conf. (2008).

[11] E. Becht, L. McInnes, J. Healy, C.A. Dutertre, I.W.H. Kwok, L.G. Ng, F. Ginhoux, E.W. Newell, Dimensionality reduction for visualizing single-cell data using UMAP, Nat. Biotechnol. 37 (2019) 38–47. https://doi.org/10.1038/nbt.4314.

[12] I. Nikishin, R. Dulimov, G. Skryabin, S. Galetsky, E. Tchevkina, D. Bagrov, ScanEV – A neural network-based tool for the automated detection of extracellular vesicles in TEM images, Micron. 145 (2021) 103044. https://doi.org/10.1016/j.micron.2021.103044.

[13] D. Danino, Cryo-TEM of soft molecular assemblies, Curr. Opin. Colloid Interface Sci. 17 (2012) 316–329. https://doi.org/10.1016/j.cocis.2012.10.003.

[14] K. Aboussaïd, S. Bernal, G. Blanco, J.J. Calvino, G.A. Cifredo, M. López-Haro, J.M. Pintado, M.S. El Begrani, O. Stéphan, S. Trasobares, Actual constitution of the mixed oxide promoter in a Rh/Ce 1-xPrxO2-y/Al2O3 catalyst. Evolution throughout the preparation steps, Surf. Interface Anal. 40 (2008) 242–245. https://doi.org/10.1002/sia.2773.

[15] G. Haberfehlner, S.F. Hoefler, T. Rath, G. Trimmel, G. Kothleitner, F. Hofer, Benefits of direct electron detection and PCA for EELS investigation of organic photovoltaics materials, Micron. 140 (2021) 102981. https://doi.org/10.1016/J.MICRON.2020.102981.

[16] F. Baiutti, J. Blanco-Portals, S. Anelli, P. Torruella, M. López-Haro, J.J. Calvino, S. Estradé, M. Torrell, F. Peiró, A. Tarancón, J.B. Portals, S. Anelli, P. Torruella, J.J. Calvino, Tailoring the Transport Properties of Mesoporous Doped Cerium Oxide for Energy Applications, J. Phys. Chem. C. 125 (2021) 16451–16463. https://doi.org/10.1021/acs.jpcc.1c04861.

[17] L. Yedra, E. Xuriguera, M. Estrader, A. López-Ortega, M.D. Baró, J. Nogués, M. Roldan, M. Varela, S. Estradé, F. Peiró, Oxide Wizard: An EELS Application to Characterize the White Lines of Transition Metal Edges, Microsc. Microanal. 20 (2014) 698–705. https://doi.org/10.1017/S1431927614000440.

[18] S. Estradé, L. Yedra, A. López-Ortega, M. Estrader, G. Salazar-Alvarez, M.D. Baró, J. Nogués, F. Peiró, Distinguishing the core from the shell in MnO x/MnO y and FeO x/MnO x core/shell nanoparticles through quantitative electron energy loss spectroscopy (EELS) analysis, Micron. 43 (2012) 30–36. https://doi.org/10.1016/j.micron.2011.04.002.

[19] H. Tan, J. Verbeeck, A. Abakumov, G. Van Tendeloo, Oxidation state and chemical shift investigation in transition metal oxides by EELS, Ultramicroscopy. 116 (2012) 24–33. https://doi.org/10.1016/j.ultramic.2012.03.002.

[20] I. Jolliffe, Principal Component Analysis, Encycl. Stat. Behav. Sci. (2005). https://doi.org/10.1002/0470013192.BSA501.

[21] I.T. Jolliffe, J. Cadima, Principal component analysis: a review and recent developments, Philos. Trans. R. Soc. A Math. Phys. Eng. Sci. 374 (2016). https://doi.org/10.1098/RSTA.2015.0202.

[22] P. Torruella, R. Arenal, F. de la Peña, Z. Saghi, L. Yedra, A. Eljarrat, L. López-Conesa, M. Estrader, A. López-Ortega, G. Salazar-Alvarez, J. Nogués, C. Ducati, P.A. Midgley, F. Peiró, S. Estradé, 3D Visualization of the Iron Oxidation State in FeO/Fe $_3$ O $_4$ Core–Shell Nanocubes from

Electron Energy Loss Tomography, Nano Lett. 16 (2016) 5068–5073. https://doi.org/10.1021/acs.nanolett.6b01922.

[23] F. de la Peña, M.H. Berger, J.F. Hochepied, F. Dynys, O. Stephan, M. Walls, Mapping titanium and tin oxide phases using EELS: An application of independent component analysis, Ultramicroscopy. 111 (2011) 169–176. https://doi.org/10.1016/j.ultramic.2010.10.001.

[24] V.P. Pauca, J. Piper, R.J. Plemmons, Nonnegative matrix factorization for spectral data analysis, Linear Algebra Appl. 416 (2006) 29–47. https://doi.org/10.1016/J.LAA.2005.06.025.

[25] Y.X. Wang, Y.J. Zhang, Nonnegative matrix factorization: A comprehensive review, IEEE Trans. Knowl. Data Eng. 25 (2013) 1336–1353. https://doi.org/10.1109/TKDE.2012.51.

[26] M. Shiga, K. Tatsumi, S. Muto, K. Tsuda, Y. Yamamoto, T. Mori, T. Tanji, Sparse modeling of EELS and EDX spectral imaging data by nonnegative matrix factorization, Ultramicroscopy. 170 (2016) 43–59. https://doi.org/10.1016/J.ULTRAMIC.2016.08.006.

[27] P. Trebbia, N. Bonnet, EELS elemental mapping with unconventional methods I. Theoretical basis: Image analysis with multivariate statistics and entropy concepts, Ultramicroscopy. 34 (1990) 165–178. https://doi.org/10.1016/0304-3991(90)90070-3.

[28] M. Bosman, V.J. Keast, J.L. García-Muñoz, A.J. D'Alfonso, S.D. Findlay, L.J. Allen, Two-dimensional mapping of chemical information at atomic resolution, Phys. Rev. Lett. 99 (2007) 1–4. https://doi.org/10.1103/PhysRevLett.99.086102.

[29] K.J. Dudeck, M. Couillard, S. Lazar, C. Dwyer, G.A. Botton, Quantitative statistical analysis, optimization and noise reduction of atomic resolved electron energy loss spectrum images, Micron. 43 (2012) 57–67. https://doi.org/10.1016/J.MICRON.2011.07.008.

[30] L. der Maaten, G. Hinton, Visualizing data using t-SNE., J. Mach. Learn. Res. 9 (2008).

[31] L. Maaten Van Der, Accelerating t-SNE using tree-based algorithms, J. Mach. Learn. Res. 15 (2014) 3221–3245. http://jmlr.org/papers/v15/vandermaaten14a.html%5Cnfiles/1017/JMLR-van_der_Maaten-2014-Accelerating_t-SNE_using_Tree-Based_Algorithms.pdf%5Cnfiles/1019/JMLR-van_der_Maaten-2014-Accelerating_t-SNE_using_Tree-Based_Algorithms-Supplement.pdf%5Cnhttp://lvdmaa.

[32] L. Mcinnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 2020.

[33] J. Blanco-portals, F. Peiró, S. Estradé, Strategies for EELS Data Analysis . Introducing UMAP and HDBSCAN for Dimensionality Reduction and Clustering, Microsc. Microanal. (2021) 1–14. https://doi.org/10.1017/S1431927621013696.

[34] F. Abramovich, Y. Ritov, Statistical theory: A concise introduction, Stat. Theory A Concise Introd. (2013) 1–225. https://doi.org/10.1201/B14755.

[35] F.J. Anscombe, The Transformation of Poisson, Binomial and Negative-Binomial Data, Biometrika. 35 (1948) 246. https://doi.org/10.2307/2332343.

[36] M. Makitalo, A. Foi, Optimal Inversion of the Generalized Anscombe Transformation for Poisson-Gaussian Noise, IEEE Trans. Image Process. 22 (2013) 91–103. https://doi.org/10.1109/TIP.2012.2202675.

[37] N. Dobigeon, N. Brun, Spectral mixture analysis of EELS spectrum-images, Ultramicroscopy. 120 (2012) 25–34. https://doi.org/10.1016/J.ULTRAMIC.2012.05.006.

[38] J. Mairal, F. Bach, J. Ponce, G. Sapiro, Online learning for matrix factorization and sparse coding, J. Mach. Learn. Res. 11 (2010).

[39] C. Boutsidis, E. Gallopoulos, SVD based initialization: A head start for nonnegative matrix factorization, Pattern Recognit. 41 (2008) 1350–1362. https://doi.org/10.1016/J.PATCOG.2007.09.010.

[40] K.J. Batenburg, S. Bals, J. Sijbers, C. Kübel, P.A. Midgley, J.C. Hernandez, U. Kaiser, E.R. Encina, E.A. Coronado, G. Van Tendeloo, 3D imaging of nanomaterials by discrete tomography, Ultramicroscopy. 109 (2009) 730–740. https://doi.org/10.1016/J.ULTRAMIC.2009.01.009.

[41] P. Potapov, On the loss of information in PCA of spectrum-images, Ultramicroscopy. 182 (2017) 191–194. https://doi.org/10.1016/j.ultramic.2017.06.023.

[42] J. Spiegelberg, J. Rusz, Can we use PCA to detect small signals in noisy data?, Ultramicroscopy. 172 (2017) 40–46. https://doi.org/10.1016/J.ULTRAMIC.2016.10.008.

[43] S. Lichtert, J. Verbeeck, Statistical consequences of applying a PCA noise filter on EELS spectrum images, Ultramicroscopy. 125 (2013) 35–42. https://doi.org/10.1016/j.ultramic.2012.10.001.

[44] L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, (2018). https://arxiv.org/abs/1802.03426v3.

[45] J. Ryu, H. Kim, R.M. Kim, S. Kim, J. Jo, S. Lee, K.T. Nam, Y.-C. Joo, G.-C. Yi, J. Lee, M. Kim, Dimensionality reduction and unsupervised clustering for EELS-SI, Ultramicroscopy. (2021) 113314. https://doi.org/10.1016/J.ULTRAMIC.2021.113314.

[46] G. Hinton, S. Roweis, Stochastic neighbor embedding, Adv. Neural Inf. Process. Syst. (2003).

[47] J. Cook, I. Sutskever, A. Mnih, G. Hinton, Visualizing similarity data with

a mixture of maps, in: Artif. Intell. Stat., 2007, 67–74.

[48] K. Pal, M. Sharma, Performance evaluation of non-linear techniques UMAP and t-SNE for data in higher dimensional topological space, Proc. 4th Int. Conf. IoT Soc. Mobile, Anal. Cloud, ISMAC 2020. (2020) 1106–1110. https://doi.org/10.1109/I-SMAC49090.2020.9243502.

[49] L. McInnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, (2018). http://arxiv.org/abs/1802.03426.

[50] O. Maimon, L. Rokach, Data Mining and Knowledge Discovery Handbook, Springer-Verlag, 2005. https://doi.org/10.1007/b107408.

[51] B. Li, L. Han, Distance Weighted Cosine Similarity Measure for Text Classification, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 8206 LNCS (2013) 611–618. https://doi.org/10.1007/978-3-642-41278-3_74.

[52] R.F.F. Egerton, Electron Energy-Loss Spectroscopy in the Electron Microscope, Springer Science, Boston, MA, 2011. https://doi.org/10.1007/978-1-4419-9583-4.

[53] D.B. Williams, C.B. Carter, The Transmission Electron Microscope, in: Transm. Electron Microsc., Springer US, Boston, MA, 1996: pp. 3–17. https://doi.org/10.1007/978-1-4757-2519-3_1.

[54] A. Adhikari, J. Adhikari, Advances in Knowledge Discovery in Databases, Springer International Publishing, 2015. https://doi.org/10.1007/978-3-319-13212-9.

[55] AnkerstMihael, B. M., KriegelHans-Peter, SanderJörg, OPTICS, ACM SIGMOD Rec. 28 (1999) 49–60. https://doi.org/10.1145/304181.304187.

[56] V. Estivill-Castro, J. Yang, Fast and Robust General Purpose Clustering Algorithms, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes

Artif. Intell. Lect. Notes Bioinformatics). 1886 (2000) 208–218. https://doi.org/10.1007/3-540-44533-1_24.

[57] J.D. Banfield, A.E. Raftery, Model-Based Gaussian and Non-Gaussian Clustering, Biometrics. 49 (1993) 803. https://doi.org/10.2307/2532201.

[58] S.P. Lloyd, Least Squares Quantization in PCM, IEEE Trans. Inf. Theory. 28 (1982) 129–137. https://doi.org/10.1109/TIT.1982.1056489.

[59] M.J. Sabin, R.M. Gray, Global Convergence and Empirical Consistency of the Generalized Lloyd Algorithm, IEEE Trans. Inf. Theory. 32 (1986) 148–155. https://doi.org/10.1109/TIT.1986.1057168.

[60] V.R. Patel, R.G. Mehta, Performance analysis of MK-means clustering algorithm with normalization approach, Proc. 2011 World Congr. Inf. Commun. Technol. WICT 2011. (2011) 974–979. https://doi.org/10.1109/WICT.2011.6141380.

[61] R.-F. Cai, M.-T. Chang, S.-C. Lo, C.-C. Chen, Novel spectral unmixing approach for electron energy-loss spectroscopy, New J. Phys. 22 (2020) 033029. https://doi.org/10.1088/1367-2630/AB7A89.

[62] D. Arthur, S. Vassilvitskii, k-means++: The Advantages of Careful Seeding, Stanford (2006).

[63] J.H. Ward, Hierarchical Grouping to Optimize an Objective Function, J. Am. Stat. Assoc. 58 (1963) 236–244. https://doi.org/10.1080/01621459.1963.10500845.

[64] B.W. Silverman, Density Estimation for Statistics and Data Analysis, Density Estim. Stat. Data Anal. (2018) 1–175. https://doi.org/10.1201/9781315140919.

[65] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, kdd 96 (1996) 226–231.

[66] E. Schubert, J. Sander, M. Ester, H.P. Kriegel, X. Xu, DBSCAN revisited, revisited: Why and how you should (still) use DBSCAN, ACM Trans. Database Syst. 42 (2017). https://doi.org/10.1145/3068335.

[67] J. Sander, M. Ester, H.P.P. Kriegel, X. Xu, Density-based clustering in spatial databases: The algorithm GDBSCAN and its applications, Data Min. Knowl. …. 194 (1998) 169–194. https://doi.org/10.1023/A:1009745219419.

[68] M.E. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, A. Zimek, Can Shared-Neighbor Distances Defeat the Curse of Dimensionality?, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 6187 LNCS (2010) 482–500. https://doi.org/10.1007/978-3-642-13818-8_34.

[69] A. Zimek, E. Schubert, H.-P. Kriegel, A survey on unsupervised outlier detection in high-dimensional numerical data, Stat. Anal. Data Min. ASA Data Sci. J. 5 (2012) 363–387. https://doi.org/10.1002/SAM.11161.

[70] R.J.G.B. Campello, D. Moulavi, A. Zimek, J. Sander, Hierarchical density estimates for data clustering, visualization, and outlier detection, ACM Trans. Knowl. Discov. Data. 10 (2015) 1–51. https://doi.org/10.1145/2733381.

[71] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering, J. Open Source Softw. 2 (2017) 205. https://doi.org/10.21105/joss.00205.

[72] R.J.G.B. Campello, D. Moulavi, J. Sander, Density-Based Clustering Based on Hierarchical Density Estimates, Lect. Notes Comput. Sci. (Including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics). 7819 LNAI (2013) 160–172. https://doi.org/10.1007/978-3-642-37456-2_14.

[73] L. Lelis, J. Sander, Semi-supervised density-based clustering, Proc. - IEEE Int. Conf. Data Mining, ICDM. (2009) 842–847. https://doi.org/10.1109/ICDM.2009.143.

[74] J.A. Hartigan, Clustering algorithms, John Wiley & Sons, 1975.

[75] A.K. Jain, R.C. Dubes, Clustering Methodologies in Exploratory Data Analysis, Advances in computers 19 (1980) 113-228. https://doi.org/10.1016/S0065-2458(08)60034-0.

[76] P. Torruella, M. Estrader, A. López-Ortega, M.D. Baró, M. Varela, F. Peiró, S. Estradé, Clustering analysis strategies for electron energy loss spectroscopy (EELS), Ultramicroscopy. 185 (2018) 42–48. https://doi.org/10.1016/J.ULTRAMIC.2017.11.010.

[77] M. Bosman, M. Watanabe, D.T.L. Alexander, V.J. Keast, Mapping chemical and bonding information using multivariate analysis of electron energy-loss spectrum images, Ultramicroscopy. 106 (2006) 1024–1032. https://doi.org/10.1016/j.ultramic.2006.04.016.

[78] P.J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, J. Comput. Appl. Math. 20 (1987) 53–65. https://doi.org/10.1016/0377-0427(87)90125-7.

[79] E.B. Fowlkes, C.L. Mallows, A method for comparing two hierarchical clusterings, J. Am. Stat. Assoc. 78 (1983) 553–569. https://doi.org/10.1080/01621459.1983.10478008.

[80] R. Kass, P. Vos, Geometrical foundations of asymptotic inference, John Wiley & Sons, 2011. https://doi.org/10.1002/9781118165980.

[81] N. Halko, P.-G. Martinsson, J.A. Tropp, Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions, SIAM Rev. 53 (2009) 217–288. https://arxiv.org/abs/0909.4061v2.

[82] A. Cichocki, A.H. Phan, Fast local algorithms for large scale nonnegative matrix and tensor factorizations, IEICE Trans. Fundam. Electron. Commun. Comput. Sci. E92-A (2009) 708–721. https://doi.org/10.1587/TRANSFUN.E92.A.708.

[83] C. Févotte, J. Idier, Algorithms for nonnegative matrix factorization with the β-divergence, Neural Comput. 23 (2011) 2421–2456. https://doi.org/10.1162/NECO_A_00168.

Chapter 5

ELNES analysis combining clustering and NLLS. The WhatEELS software solution.

5.1. Introduction.

When it comes to sample analysis via EELS spectroscopy, a direct quantification measurement based on the Egerton method will often suffice to characterize the material to the desired extent[1–4]. Usually, well-known dimensionality reduction algorithms as principal component analysis (PCA)[5] or independent component analysis (ICA)[6], are used to isolate and identify specific regions with distinctive spectral features induced by local differences in composition or chemical state (e.g., high dopant concentration areas or changes in the oxidation states).PCA can be used in combination with the aforementioned Egerton method[7], as a noise-reduction pre-step. Nevertheless, PCA does not come without problems, namely user bias on the spectral base selection and interpretation and problems of non-compliant data samples[8–10]. Hence, further efforts are constantly being made to integrate new techniques into the standard EELS data-treatment arsenal.

In recent times, unsupervised clustering classifying algorithms (mainly K-means and hierarchical clustering) have been presented as fast methods to achieve a qualitative segmentation of the ever-increasing in size EELS datasets[11]. One major drawback to be noticed is the lack of a robust way to easily evaluate the accuracy of the clusters resolved, other than a later inspection by the user (i.e., always prone to some sort of bias). Also, if the sample presents areas with significant differences in thickness, the algorithms

may lose effectivity (i.e., the total number of raw electrons counts, dominated by thickness effects, will outweigh any minor spectral variation associated to the presence of trace elements or low dopant concentration areas). Nonetheless, it is a fast way to inspect large datasets and get an initial assessment of the presence of different areas in the sample segmented by spectral features.

The correct use of clustering algorithms has played a key role in this thesis. As a matter of fact, the specifics on how to properly use clustering algorithms for EELS data analysis, and which algorithms should be given preference in specific cases are carefully explained in chapter 2 (***Machine Learning for EELS data analysis: Clustering and dimensionality reduction methods***) of this PhD manuscript.

Moving on to other analysis techniques more inclined towards quantitative analysis, multiple linear least squares (MLLS) fitting creates a series of weighted maps for user-selected reference signals, that may be externally provided or extracted from the same dataset under analysis. It is a trusted method in the EELS community to get quantitative measurements [12–15], but its accuracy may suffer when facing dataset with marginal regions with small differences in compositions, given its dependency on the manually set references. The combined use of clustering and MLLS has been recently proposed as a way to overcome the bias when selecting these references, by using the centroids as the signals to be fitted[16].

Finally, the non-linear least squares (NLLS) method applied to EELS[17–19] consists of fitting a variety of individual curves (components) to get the best approximation of the given raw data. Usually, a combination of simple peak-like curves for the specific near edge features (e.g., Gaussians, Lorentzian and Pseudo-Voigt), and arctangent, power-law decays or cross-section curves (calculated by Hartree-Slater approximations or from hydrogenic models) for the continuum excitations, are carefully selected for each edge on the spectra. The posterior analysis of the parameters of these curves is what makes NLLS a great choice to study particular ELNES characteristics[20], such as local

variations in the atomic oxidation states, given that the most common reported methods of characterization in the literature are the white-line (WL) ratio variation measurements[21–23], the relative position of WLs and their distance to the onset of the oxygen edge[24] (computed by integrating the peak-like curves or measuring their centres after the fitting). Also, NLLS can be a valuable tool for elemental quantification in cases of heavily overlapping edges (e.g., samples of mixed cerium-praseodymium oxides[25]), as the fitting process is able to separate the electron counts belonging to different elements. One major drawback is that the NLLS data fitting process is usually slow, given the large number of parameters involved with each added component. Furthermore, the increment in convergence times is not linear with the number of components (i.e., adding a single extra curve to the fitting may dramatically increase the time it takes to finalize the fitting process, or even cause divergent results).

At the beginning of this thesis, several software tools were already available for EELS data analysis. Most of them incorporated in some degree a software solution for some of the aforementioned analysis and fitting techniques. The two more popular were Digital Micrograph (**DM**, currently on its third version, it is a proprietary tool from GATAN) and **HyperSpy**[26] (a free-software, python-based multidisciplinary spectroscopic tool). The main problem of **DM** was that most of the EELS and ELNES analysis tools, although powerful and reliable, were not distributed freely. **HyperSpy** is still, arguably, the most popular free-software tool, continuously supported and updated by a sizable community of users. Nevertheless, when it comes to the specific task of EELS spectral fitting it lacked the speed for an efficient workflow, and it presented a convoluted system to carry out NLLS fittings. It also lacked the solutions for clustering analysis of EELS datasets. But the main problem with **HyperSpy** is that it requires a minimum knowledge of Python language to be successfully used and for the posterior analysis of the results, which often drives away potential users.

As a direct result of the different shortcomings of the available tools, a key aspect of the work carried out during this doctoral thesis was the design from scratch of a new tool, specifically focused on ELNES analysis[27]. **WhatEELS** was born as a direct successor of the **DM** applet **Oxide Wizard**[20], but with the intention of providing an autonomous platform on top of Python (i.e., a completely user free software solution with unlimited expansion capabilities).

Although its backend is coded in Python, it was made along with an interactive and modular graphical interface with the intention of facilitating its introduction to users with a limited knowledge of programming languages. The software combines the segmentation capabilities of clustering algorithms and the detailed structure description accessed by NLLS fitting routines. The combined use of clustering analysis and NLLS was expected to improve convergence times in problems with several regions of different elemental compositions (e.g., multi-layered structures with different compounds per layer), and also help the algorithm to converge in pixels classified as spectral outliers (i.e., pixels presenting strong and unique variations in their spectral characteristics in problems with complex multi-component spectra). The main advantage of combining clustering and NLLS is the fine control it grants to the user over the models fitted, as any given region in the dataset may be tweaked and analysed independently at any time.

The chapter is structured as follows:

- ***Elemental quantification in WhatEELS.***

  This first section consists of a brief background introduction to some of the theoretical basics of EELS spectra formation and the effects of different experimental parameters. This section is required to understand some of the computations that **WhatEELS** carries out, especially for the partial cross sections computations and ionization cross sections used as the excitation to the continuum functions and for the elemental quantification.

- ***WhatEELS. Software architecture.***

  All the modules are described in detail in this second section. Furthermore, some quick-guides and flow charts are included to understand how the software would be used during an analysis process.

- ***Iron- Manganese oxide nanoparticles.***

  The third section consist of an example of use of **WhatEELS**. A real case scenario is proposed through the analysis of a Fe-Mn oxide core-shell NP of known composition. The particular combination of elemental composition and changes in the cationic oxidation states present in this sample granted the opportunity to test all the analysis tools included in **WhatEELS**. At the same time, its known composition provides a sort of 'ground truth knowledge' to cross check the results extracted, as it was previously extensively characterized [4,20] and has also been used in chapter 2 to illustrate the clustering methods.

- ***The panel structure.***

  This last section contains images of the current state of the panels that compose **WhatEELS**. They are included as large panels and are directly linked to the functionalities and panels described in the software architecture section

5.2. Elemental quantification in WhatEELS.

One of the functionalities included in **WhatEELS** is the possibility to carry out EELS elemental quantification using the Egerton's method[7]. This quantification method relies on the previous knowledge of the scattering cross sections for the elements analysed. Furthermore, the NLLS fitting routine includes functions to model the excitation to the continuum (i.e., the function describing the excitation of an atomic inner-subshell electron to the continuum of states). These functions are based on theoretical atomic calculations for generalized oscillator strength (GOS, description below), and

are read from a tabulated database. Finally, some corrections are put in place for the scattering cross sections calculated including the effects of the finite nature of the convergence and collection angles in an EELS experiment.

Therefore, a minimum understanding of the physical process behind the formation of energy loss spectra and the elemental quantification computations is required, in order to fathom the inner workings of **WhatEELS**.

5.2.1. Inelastic scattering of atomic inner-shell electrons and Bethe's theory.

The following lines constitute a brief summary of the Bethe theory for the electron inelastic scattering process. They are not intended as a complete mathematical and physical argumentation, as such is not the objective of this work. They rather give as a small piece of information that will be helpful later to understand a key aspect of the **WhatEELS** innerworkings: the angular corrections for the elemental quantification. The reader is encouraged dive deep into a much more accurate and detailed description in the Egerton's reference text[1] (chapters 3 and 4 of the 3rd edition).

For simplicity, let us consider that the edges of the core-loss region of the spectra can be totally ascribed to inelastic scattering events with electrons in the atomic inner-shells[14]. The incident electron (from the TEM beam) would then transfer a certain quantity of energy and momentum to the interacting inner-shell atomic electron. The quantity that describes an electron scattering event for a given solid angle and an energy loss value of E_n is the differential cross section.

In Bethe theory, the differential cross-section for an inner-shell electron transitioning from an initial state 0 to an energy state n can be written as

[14] In reality, the core-loss edges are not only shaped by the inner-shell scattering events. For instance, the plasmon region in the low-loss energy region of the spectra may contribute to the electron counts in the core-loss region, whenever the sample is thick enough and, thus, multiple scattering events are common. The plasmon region is dominated by the inelastic scattering of electrons with the valence band of the sample. In such cases, deconvolution of the spectra with the low-loss spectrum region may be required.

$$\frac{d\sigma_n}{d\Omega} = \left(\frac{4\gamma^2}{a_0^2 q^4}\right)\frac{k_1}{k_0}|\varepsilon_n(\vec{q})|^2 \quad (1)$$

where $\vec{q}$ is the transferred momentum, k_1 and k_0 are the wavevectors magnitudes for the scattered electron after and before the event respectively, a_0 is Bohr's radius and γ is the relativistic gamma factor for the incident electron. $\varepsilon_n(\vec{q})$ is the so-called inelastic form factor, and is calculated from the wavefunctions of the excited inner-shell electron at the initial and final states (i.e., this is the term that gives each edge its distinctive shape). These wavefunctions can be calculated by several methods.

In **WhatEELS**, they are extracted from the Hartree-Slater approximations, but the use of the hydrogenic model can also be found in the literature.

For convenience, the scattering cross section is sometimes presented in terms of the quantity called ***generalized oscillator strength*** **(GOS).**

$$f_n(\vec{q}) = \left(\frac{E_n}{R}\right)\frac{|\varepsilon_n(\vec{q})|^2}{(qa_0)^2} \quad (2)$$

where R = 13.6 eV is the Rydberg energy (constant), and E_n is the energy transferred to the inner-shell electron (i.e., the energy loss value for the transition).

Both formulations for the differential cross section and GOS in eq.(1) and (2) treat the energy dependence in a discrete way (i.e., a single electron excited from an initial quantum state to a single higher energy one). However, there are many instances where these functions are required to vary continuously with the energy loss values (e.g., the description of the excitation to the continuum of states). The continuous formulation with respect to the energy excitation (or equivalently, the energy loss value for the incident electron) for the differential cross section is

$$\frac{d^2\sigma}{d\Omega dE} = \left(\frac{4\gamma^2 R}{E q^2}\right)\frac{k_1}{k_0}\frac{df}{dE}(q,E) \quad (3)$$

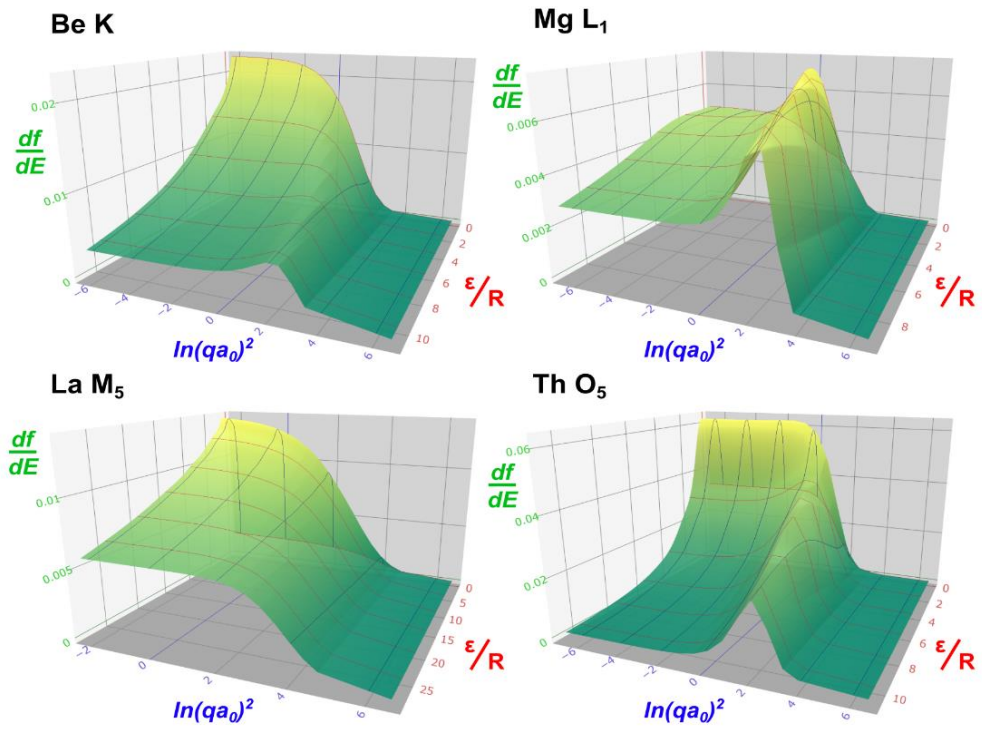


Figure 5.1. Bethe surfaces for the energy dependent GOS functions of an inner shell electron been excited from the (a) K shell in a beryllium (Be) atom, (b) the L (L_1) shell in a magnesium (Mg) atom, (c) the M (M_5) shell in a lanthanum (La) atom and (d) from the O (O_5) shell in a thorium (Th) atom.

where the (differential) GOS function contains both the angular and energy dependence now.

Oftentimes, the continuum GOS df/dE is portrayed as a 2D surface (in a 3D plot), known as Bethe's surface (some examples included in **Figure 5.1**). The epsilon ($\boldsymbol{\varepsilon}$) in the x axis (in red, the energy axis) is the energy measured from the onset energy loss (i.e., $\varepsilon = E_{loss} - E_{onset}$), divided by Rydberg's constant to obtain a dimensionless quantity. The y axis (in blue, the momentum / angular dependence) is expressed in terms of the natural logarithm of the transferred momentum (q) times Bohr's radius. This dimensionless quantity is introduced to facilitate the numerical calculations later.

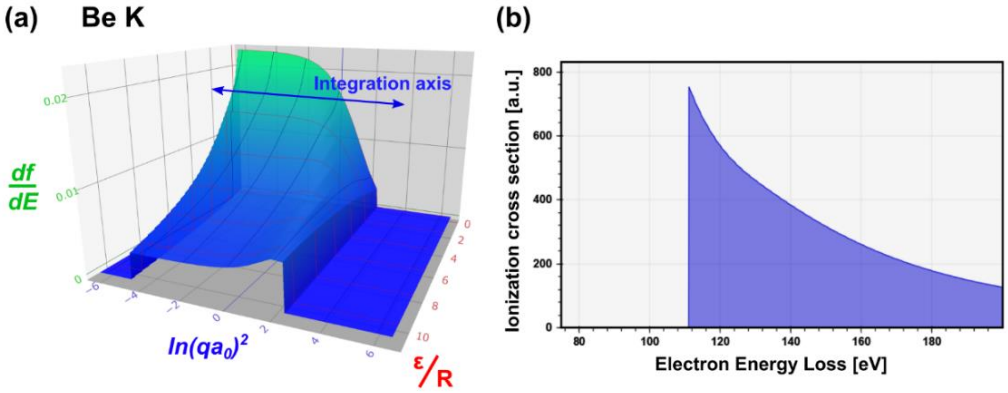


Figure 5.2. (a) Bethe surface cut by the limits imposed in the y-axis by equations (5) and (6) (i.e., transferred momentum limits). (b) Ionization cross section for the excitation of an electron in a beryllium atom from the K shell, obtained by integrating the surface in (a).

The ionization cross section is defined by integrating $d^2\sigma \,/\, d\Omega dE$ up to a certain angle (i.e., the collection angle β) defined by the experimental constraints (valid for scattering angles below 1 rad). Usually, it is expressed as follows

$$\frac{d\sigma}{dE} \sim (4\pi a_0^2)\frac{R^2}{ET}\int \frac{df(q,E)}{dE}\, d[\ln(qa_0)^2] \tag{4}$$

where the integration is done over the logarithmic axis for the momentum transfer (angular) variable. The transition from eq.(3) to eq.(4) incorporates the results of the conservation of energy. As it stands, resolving the integral in eq.(4) is equivalent to integrate the area under the E-constant curves of the Bethe surface (see **Figure 5.2**) for every energy value (i.e., integrate the surface along the y-axis). The integration limits in eq.(4) are defined by the conservation of momentum as

$$(qa_0)^2{}_{min} = \frac{E^2}{4RT} \tag{5}$$

$$(qa_0)^2{}_{max} = (qa_0)^2{}_{min} + 4\gamma^2\frac{T}{R}\sin^2\left(\frac{\beta}{2}\right) \tag{6}$$

Notice from eq.(6) the inclusion of the beta (collection) angle in the calculations. The ionization cross section defined by eq.(4) is the function that models the ionization to the continuum of states.

5.2.2. Elemental quantification and the effects of a finite convergence angle.

Considering (by now) the plural scattering events negligible, each edge k (inner-shell excitation) contributes to the EEL spectrum intensity a quantity equal to

$$J_k^1(\beta, E) = N I_0 \frac{d\sigma}{dE} \tag{7}$$

which is called the single-scattering intensity for the k inner-shell excitation. N is the areal density (number of atoms per unit of area) and I_0 is the zero-loss intensity (no plural scattering included, i.e., no plasmon contribution to the spectrum).

The partial cross section is defined from the integration of J_k^1 in a finite energy range Δ. Considering now (finally) the effects of the plasmon excitations on the core-loss energy loss spectrum one would get[15]

$$I_k(\beta, \Delta) \sim N\, I(\beta, \Delta)\sigma(\beta, \Delta) = N\, I(\beta, \Delta) \int_{E_K}^{E_k+\Delta} \frac{d\sigma}{dE}\, dE \tag{8}$$

Hence, the elemental quantification from a single spectrum with several edges (contributions of different inner-shell excitations) would be calculated by pairs as

$$\frac{N_a}{N_b} = \frac{I_{ka}(\beta, \Delta)}{I_{jb}(\beta, \Delta)} \cdot \frac{\sigma_{jb}(\beta, \Delta)}{\sigma_{ka}(\beta, \Delta)} \tag{9}$$

For two elements a and b and their respective edges k and j.

[15] This last equation leaps a couple of steps. The key change is the substitution of I_0 (i.e., zero-loss peak integrated intensity) by $I(\beta, \Delta)$ (i.e., the integrated intensity of the low-loss spectra in a energy range of Δ). For more information see [1]

Normally, the introduction of the elemental quantification in EELS analysis would be concluded at this point. However, there is still a final step missing from these calculations: the inclusion of the effects of a finite convergence angle.

Two ways of calculating these corrections are presented in the literature. The first one, and most commonly described[1,28], introduces a series of F_1 and F_2 factors in the integral calculations for the partial cross sections. Sometimes, these factors are approximated by the calculation of an effective β^* angle that is directly introduced in the ionization cross section calculations.

However, the most popular software solution for EELS quantification - the proprietary software from GATAN (**DM**) - states in its documentation that the corrections are introduced in the quantification by the analytical calculation of a cross-correlation factor between the objective aperture and detector (collection aperture) functions[29]. Hence, the effects of finite convergence angle plus its interaction with a finite collection angle are translated into the quantification calculations by the aperture function in the computation of a cross-correlation factor.

This factor can be described by the simple geometrical relation between the incident electron ($\vec{\alpha}$) and the scattered electron ($\vec{\beta}$) vector electron angles. Graphically, this can be represented as the shared area of two circles with radius described by the convergence and collection angles and with centres at a distance equal to the scattering angle (see **Figure 5.3**). Thus, in this text the correction factor will indistinctively be called geometric-factor or cross-correlation-factor.

The formula for the factor function with respect to the scattering angle derived in the original publication[29] is

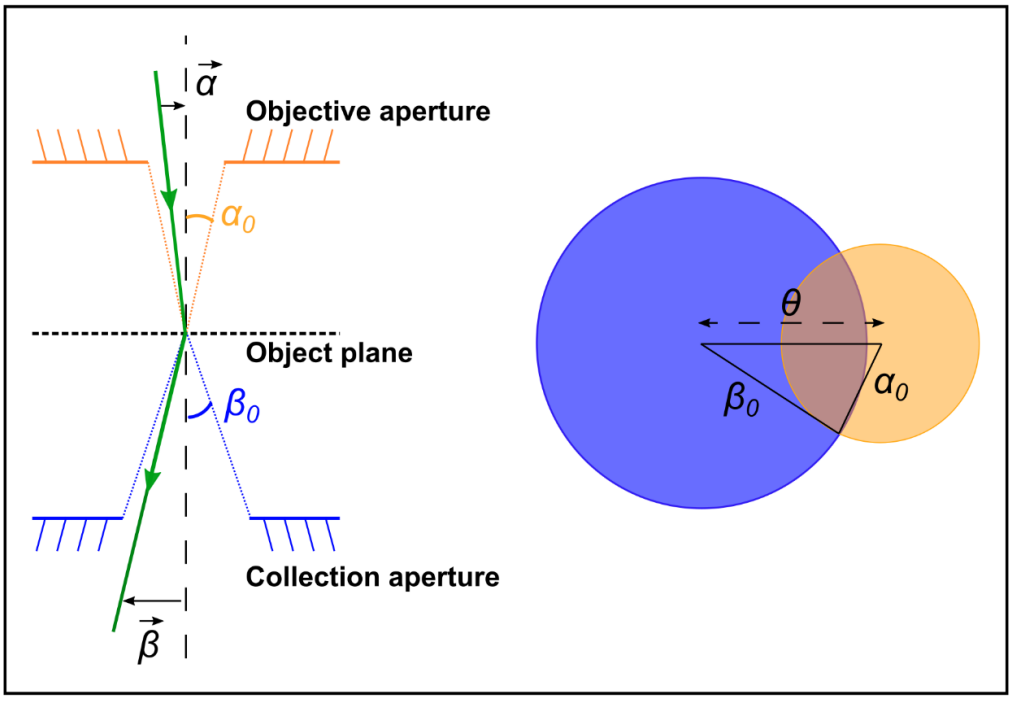


Figure 5.3. Schematic ray diagram of a STEM experiment and geometrical relation that describes the cross-correlation factor between the objective aperture and the detector function (collection aperture in the EELS spectrometer). The angular relation for the scattered angle would be $\vec{\boldsymbol{\theta}} = \vec{\boldsymbol{\beta}} - \vec{\boldsymbol{\alpha}}$.

$$F(\alpha_0, \beta_0; \theta) = \begin{cases} (\theta_\angle/\alpha_0)^2, & if\ 0 \le \theta \le |\alpha_0 - \beta_0| \\ \pi^{-1} \cdot (\arccos(x) + (\beta_0^2/\alpha_0^2)\arccos(y)) - \\ \pi^{-1}(1/2\alpha_0^2)\sqrt{4\alpha_0^2\beta_0^2 - (\alpha_0^2 + \beta_0^2 - \theta^2)^2}, & if\ |\alpha_0 - \beta_0| < \theta < \alpha_0 + \beta_0 \\ 0, & otherwise \end{cases} \tag{10}$$

where α_0 and β_0 are the convergence and collection angles, θ is the scattering angle (which in **WhatEELS** is translated back into the $\log(qa_0)^2$ form), and $\theta_\angle$, x and y are defined as

$$x = \frac{\alpha_0^2 + \beta_0^2 - \theta^2}{2\alpha_0\theta} \tag{11}$$

$$y = \frac{\beta_0^2 + \theta^2 - \alpha_0^2}{2\beta_0\theta} \tag{12}$$

$$\theta_{\angle} = \min\{\alpha_0, \beta_0\} \tag{13}$$

Once this factor is calculated for every scattering angle contained in the GOS databases for a fixed of given values of convergence and collection angles, it is directly introduced in the calculation of the ionization cross section. The integration limits are modified so the upper integration limit depends on $\alpha_0 + \beta_0$.

Initially this formulation was derived for a bright field imaging experiment, where the limiting β_0 angle was referred to the detector radius. The translation into EELS analysis is straight forward, changing the detector size by the collection aperture in the spectrometer.

Using the same machinery for the corrections that **DM** has been using for the past decades is of paramount importance, in order to be able to compare future quantification results with the ones available in the literature. To test the accuracy of the calculations carried out by **WhatEELS**, a series of partial cross sections for different electronic inner-shell excitations were calculated

including the geometric correction. The **WhatEELS** calculations for the partial cross-sections without the geometric correction in place and the digital micrograph calculations (with error bars included) were calculated as well for the same edges, for comparative purposes.

These results are presented in **Figure 5.4**. All the partial cross sections were calculated for a constant energy range of 40 eV, measured from the onset of each edge included and for an incident electron energy of 200 keV. Two different representative cases are shown in the figure. The first one (**Figure 5.4** (a)) corresponds to a convergence angle (15 mrad) smaller than the

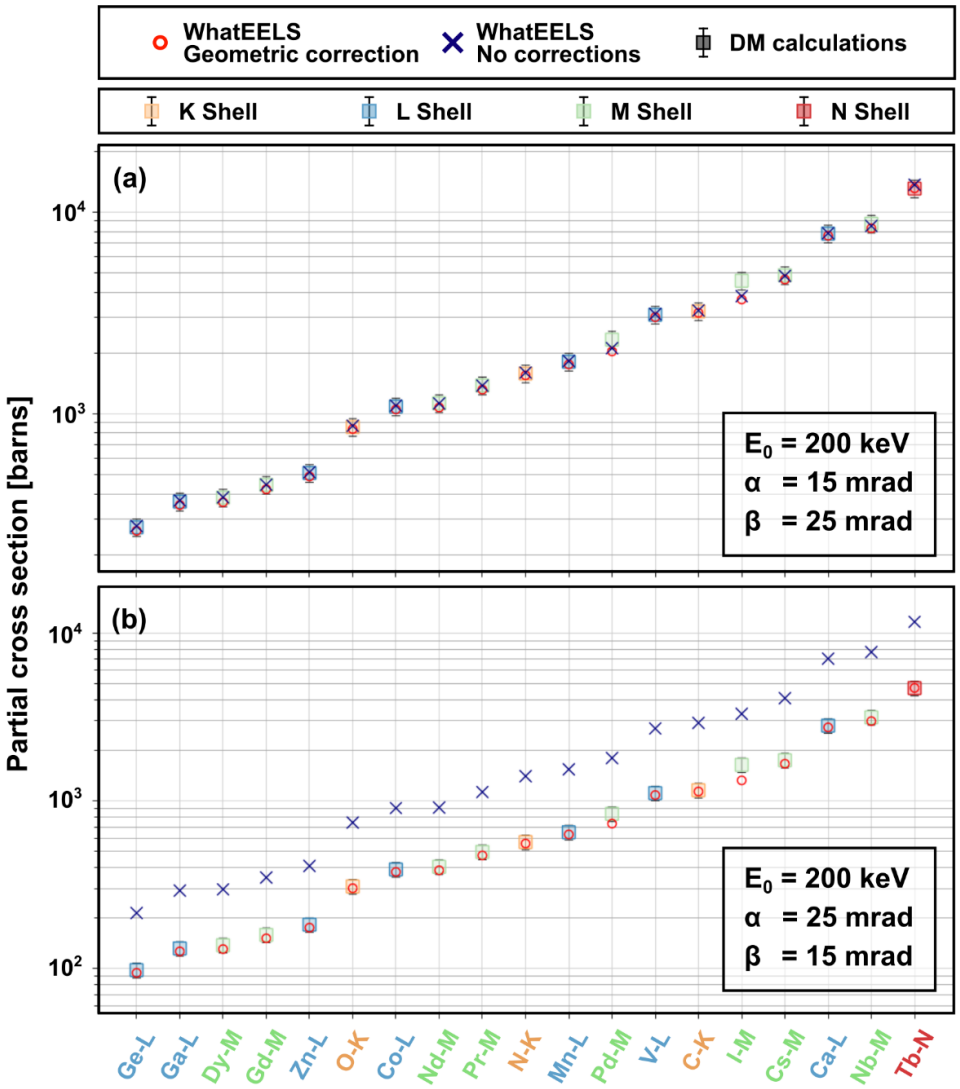


Figure 5.4. Partial cross sections, integrated for an energy range of 40eV from the onset and for a variety of inner-shell electronic excitations of different elements. (a) Integration done for α = 15 mrad, β = 25 mrad and E_0 =200 keV. (b) Integration done for α = 25 mrad, β = 15 mrad and E_0 =200 keV. Both panels share the same colour code for the type of inner shell excited and the markers. The y-axis is logarithmic (**DM** stands for Digital Micrograph).

collection one (25 mrad). The partial cross-sections calculated by **WhatEELS**, both with and without geometric corrections, fall (generally) within the accuracy limits of the ones calculated by **DM** (down to the point of being almost indistinguishable). However, as shown in (**Figure 5.4** (b))when the

convergence angle is larger (25 mrad) than the collection one (15 mrad), the partial cross-sections without the geometric correction in place calculated by **WhatEELS** present a positive bias. Notice that the scale is logarithmic, hence the separation is orders of magnitude higher with each incremental growth of the partial cross section. Regarding the results including the geometric factor, the calculations carried out by **WhatEELS** are almost identical to the ones of **DM**. The minimal fluctuations observed can be probably ascribed to the differences in the numerical evaluation (the numerical integration of the tabulated surfaces). In any case, they are well within the error limits given by **DM** for each of the partial cross sections calculated.

5.3. WhatEELS. Software architecture.

In this section, the architecture of the **WhatEELS** software is introduced. Each one of the available modules is described. Also, a series of panels for the current visual aspect of the software are included. Some quick-guide boxes containing a step-by-step manual of the most commonly used functionalities of each module are also included, so any users can explore the software capabilities on their own.

5.3.1. Overview.

The interactive shell of **WhatEELS** is based on **Panel**[30], and **Holoviews**[31]. The current graphical backend is **Bokeh**[32], as it gives access to the plasticity of the JavaScript graphical interfaces and widgets for interactive customization through Python code. It is also interchangeable with **Plotly**[33] and **Matplotlib**[34], provided some minor changes in the source code. Below the surface, the NLLS fitting is based on a library called **lmfit**[35], that expands the **SciPy**[36] fitting capabilities. The clustering and pre-processing tools use the **Scikit-learn**[37,38] (**sklearn**) library. Finally, the current file loading system relies on **HyperSpy**[26], as it is the most complete tool available to get the information in the standardized dm3 and dm4 formats extracted from the TEM.

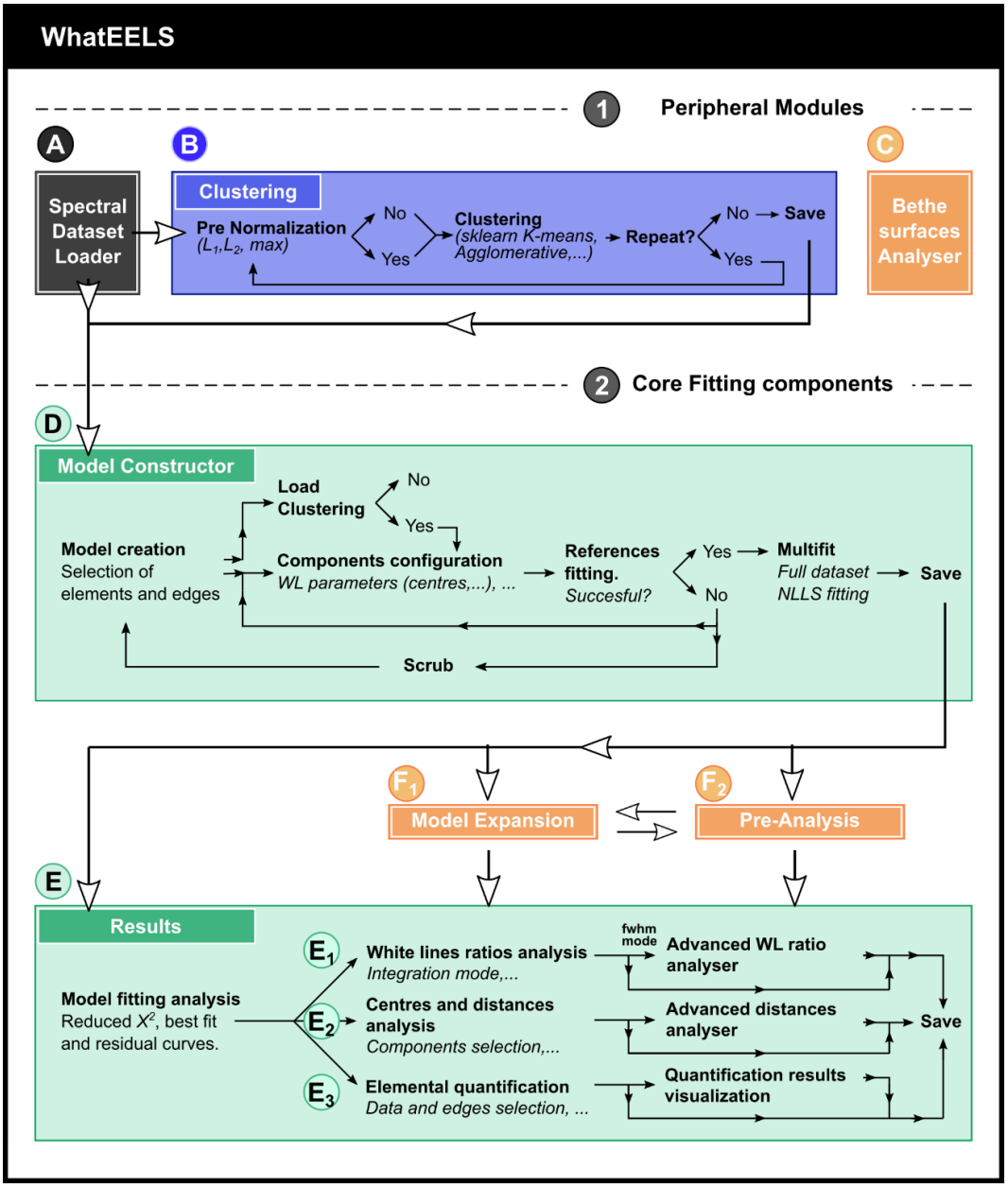


Figure 5.5 Flow-chart for **WhatEELS**. Each coloured box represents a separated tool. The arrows indicate the chronological progression in a standard EELS data analysis process. The numerical tags for the blocks identification and the alphabetical tags (capital letter) for the specific tools (modules) in each block are consistent throughout the whole chapter.

A workflow chart for a standard EELS analysis procedure using **WhatEELS** is shown in **Figure 5.5**. The software tool is comprised of two main blocks. **-1-** The '**peripheral modules**' block currently includes the tools that can be utilized independently from the NLLS fitting routines which are: **-A-** the

spectral dataset loader, **-B-** the *clustering* analysis tool and **-C-** the *Bethe-surfaces analyser* tool. The *dataset loader* **-A-**, although included as part of the peripheral modules, is always the departing point for any EELS dataset analysis utilizing **WhatEELS**. Notice that the **-C-** *Bethe-surfaces analyser* (top-right corner) is isolated from the rest of the tools, as it can be used separately at any time and does not demand a loaded dataset. **-2-** The '**core fitting components**' block contains the specific tools that carry out the fitting of the spectral datasets and provide an integrated solution for data analysis and results visualization: **-D-** the *model constructor* and the **-E-** *results* analysis tool, as well as the **-F_1-** *pre-analysis* and **-F_2-** *model expansion* tools. These last tools **-F-** are still under development, and are not required to complete a NLLS fitting and analysis.

Notice that the chart is colour-coded. The tools framed in orange add optional functionalities that are not required for the completion of a NLLS fitting and analysis. The tools framed in green are the strictly necessary ones to complete the NLLS fitting and the results analysis using **WhatEELS**, with the exception of the **-A-** *spectral dataset loader* that is coloured in black. The **-B-** *clustering* analysis tool is coloured in blue as, although it belongs to the 'peripheral modules' section, it plays a central role in the new methodology combining clustering and NLLS introduced alongside this software tool. Finally, the arrows indicate the standard chronological order of tasks to complete the type of analysis described in this work.

WhatEELS was conceptualized as future-proof software, with a modular architecture flexible enough to allow the inclusion of new analysis techniques. For example, one could easily add extra tools to the **-1-** peripheral block to include a support vector machine classifier[39], a matrix factorization dimensionality reduction routine or even a manifold learning tool to be combined with the clustering analysis classification tool[40].

Furthermore, a similar approach can be taken to include an MLLS fitting routine as a new **-2-** core components block. This would enable, for example,

the use of the centroids from the unsupervised **-B-** *clustering* classification as endmembers for the MLLS fitting[16].

5.3.2. Peripheral modules -1-.

-A- Loading panel and main HUB.

Figure 5.6 shows the tools from the initialization panel. This is the first panel loaded when launching **WhatEELS** (to see the actual distribution of tools in the app **Figure 5.30**). The main HUB (controls) tools **Figure 5.6** (a) are launched together with the **-A-**[16] *spectra loading pane l* **Figure 5.6** (b).

The left side column shown in **Figure 5.6** (a) contains the button -a- to boot/reboot the modules and the tab-selector of the **-1-** peripheral and **-2-** core-components (NLLS). The messages display panel -r- is always visible. The different modules for each tab **-1-**/**-2-** are showcased in **Figure 5.7**.

Regarding the spectra loading panel (**Figure 5.6** (b)), the different text-bars (at the top) show the current directory address -f-, the folders inside the current directory -g- and the files inside the current folder -i-. The directory address shown in the 'current-directory' bar -f- is used by the rest of modules as the reference point to load and save any information, including spectra and results. By going back -b- and forth -c- on the directory tree, or advancing inside the folder in the folder-bar -g- (pressing the down-arrow button -h-), the address in the 'current-directory' bar -f- is modified. To recover the default directory, one can always press the reset button -d-. Furthermore, copy-pasting any direction directly on the 'current directory bar' -f- is allowed. If a valid address is given, the app will set it as the current directory after pressing the 'go' button -e-.

The files-bar -i- will only show files with the following extensions: .msa, .hspy, .hdf5, .dm4 and .dm3. Those are the only valid extensions for spectral datasets,

[16] In the rest of this chapter, the capital letters tags (e.g., **-A-** or **-D-**) correspond to the different modules available and the lowercase letters tags (e.g., -a_1- or -h-) are specific buttons and tools. Notice that some of them are colour-coded, and this code is maintained throughout the chapter as well.

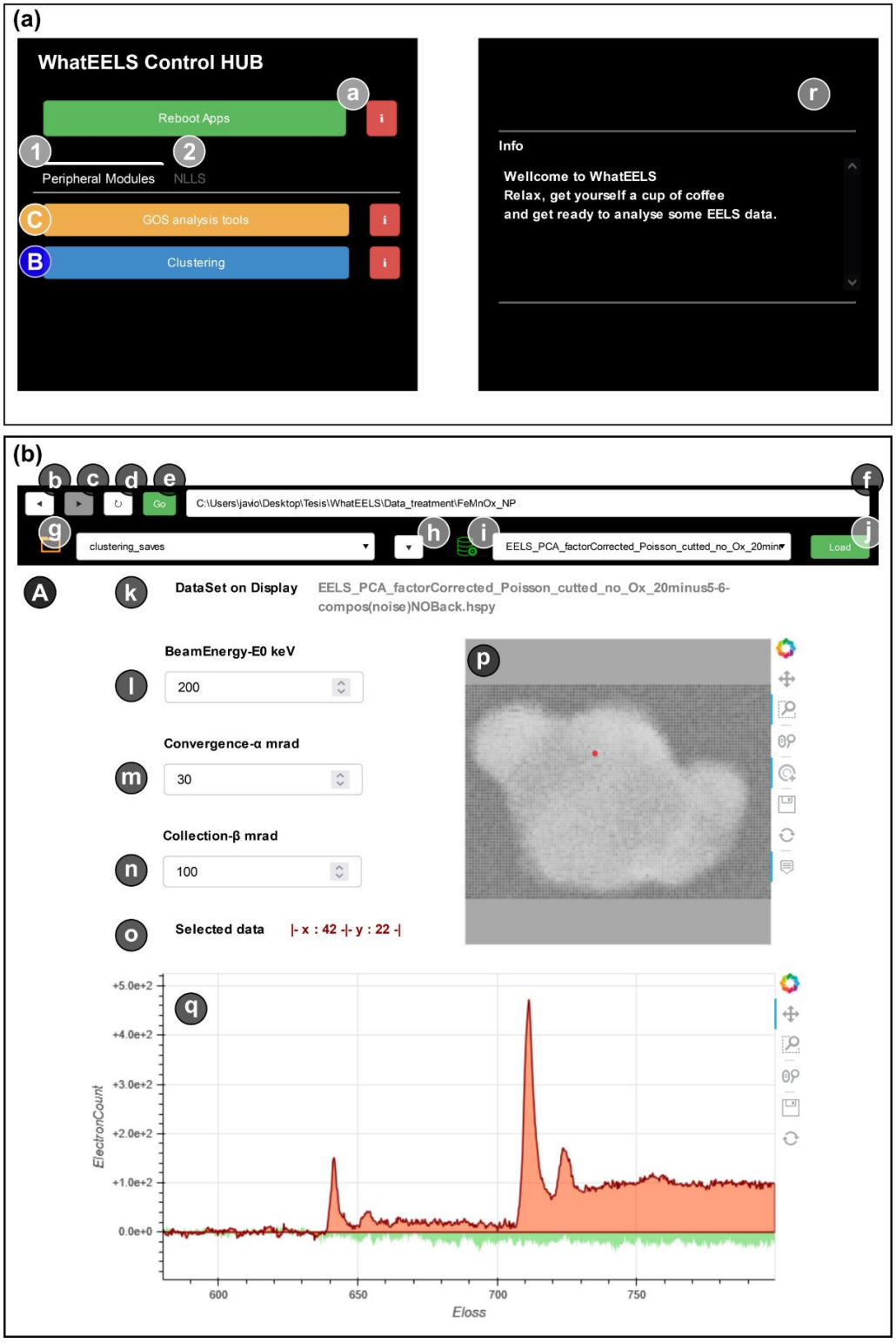


Figure 5.6 (a) **WhatEELS** main controls HUB and (b) loading panel. The capital letters tags correspond to the different modules available. The lowercase letters tags (-a- to -r-) are specific buttons and tools, corresponding in this case mostly to the **-A-** *dataset loader module.* The -r- tag corresponds to the message panel. It shows the info (red [i] buttons) and error messages.

and the most common ones in TEM software tools. By pressing the 'load' button -j-, the dataset information will be read (using **HyperSpy**), including the metadata for the beam energy -l-, convergence -m- and collection -n- angles extraction (if available). If any of these parameters are not readable from the metadata, they are set to 0. In that case, **WhatEELS** would not launch any application except the GOS analysis tools -C-, as they are needed for the calculations. Trying to boot the rest of the modules via the 'boot/reboot' button -a-, without valid (non-zero) values for the spectral

parameters (-l-m-n-) will prompt an error message in -r-. Once loaded -j-, the name of the current dataset is shown in -k-. Also, an interactive visual representation of the EELS dataset (spectrum image, spectrum line or single spectrum) will be displayed -p-. The spectrum of any given pixel (or line) can

be displayed in -q- by 'clicking' or hoovering over the area of the SI -p-. The pixel coordinates are shown in -o- for the current spectrum displayed in -q-.

A step-by-step guide to run the fitting and clustering modules is provided (**Quick-panel 1**). to allow the reader of this thesis to experiment with the software tool.

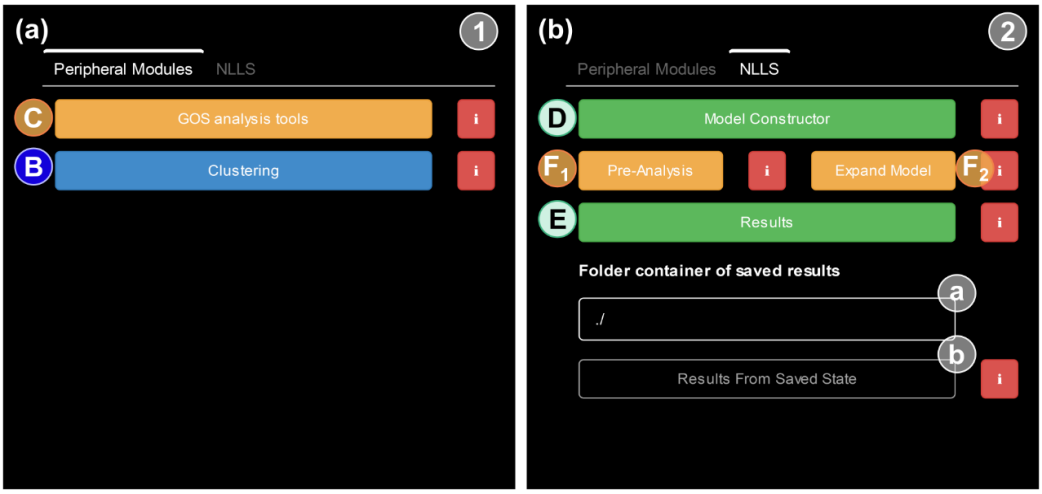


Figure 5.7 Launching controls for the different modules of **WhatEELS**. (a) Buttons to launch the -B- clustering and -C- GOS modules. (b) Button to launch the core-components modules (-D-E-F_1-F_2-) for the NLLS fitting and analysis.

Quick-panel 1. Quick instructions to navigate the initialization HUB. The letter tags and numbers are consistent with those displayed in **Figure 5.6**.

Loading a dataset

i. Navigate in the directories in -f- and -g-, by using the buttons -b, c ,d, e- and -h- and load the desired dataset -i- pressing the -j- button.

ii. Check the experimental parameters -l, m, n-, and change or complete them if any is missing.

iii. The dataset -k- is now available for an initial inspection -p- and -q-.

iv. Boot the apps in -a-, and navigate the tabs **-1-** and **-2-** to access different modules available.

Once the dataset is loaded, the following options are available

Clustering analysis

v. Select tab **-1-** for the peripheral modules, and press **-B-** to launch the clustering panel (**Figure 5.8**)

GOS Analyser (*The GOS analyser -C-, that is always available*).

vi. Select tab **-1-** for the peripheral modules, and press **-C-** to launch the GOS analyser tool (**Figure 5.9**). The previous loading of a dataset is not required for this module

Model construction and NLLS fitting

vii. Select tab **-2-** and press the **-D-** model constructor button (**Figure 5.10**)

Results analyser

There are 2 ways to access this panel.

viii. After completing a multifit in the model constructor (**Figure 5.10**), press the button **-E-** (**Figure 5.7** (b)) to launch the results panel (**Figure 5.13**)

ix. From a previously completed fitting. Paste the directory address for the results stored previously in the text-tab -a- (**Figure 5.7** (b)) and press the results-from-saved -b- button (**Figure 5.7** (b)).

If any of the above steps fail, the following will most likely solve the problem:

Check the experimental parameters. If any of them is 0, the majority of the panels will not be launched.

-B- Clustering module.

Once the dataset is loaded in the main HUB, one has access to the clustering segmentation tool panel (see **Figure 5.8**). The use of this module is not mandatory to achieve a successful NLLS fitting afterwards, but the combined use of clustering and NLLS is one of the main novelties of **WhatEELS** and is heavily recommended.

The different clustering algorithms -c- (implementations extracted from sklearn) and the customization options for their hyperparameters -d- can be accessed interactively from the module (**Figure 5.8** (a)). At the moment, only the parametric clustering algorithms (K-means and hierarchical agglomerative clustering) are available (see chapter 3 for more information about this clustering algorithms and their application to EELS data analysis). The button to carry out a clustering analysis -a- (Go-clustering!) is only unlocked when a number of objective clusters is specified for any of the two available algorithms. Once the clustering calculations are completed, the results can be locally stored by pressing the store-run -b- button. Also, at any time, a pre-processing step can be added -e- to the clustering classification. The L2 and L1 normalizations are available, as well as an option called max normalization. They have the following mathematical formulation

$$L_1{:}\,norm(\vec{x}) = \|\vec{x}\|_1 = \sum_{i=1}^{m} |x_i| \tag{14}$$

$$L_2{:}\,norm(\vec{x}) = \|\vec{x}\|_2 = \left(\sum_{i=1}^{m} x_i^2\right)^{\frac{1}{2}} \tag{15}$$

$$max{:}\,norm(\vec{x}) = \max(x_i)_{\forall\, i=1,\dots,m} \tag{16}$$

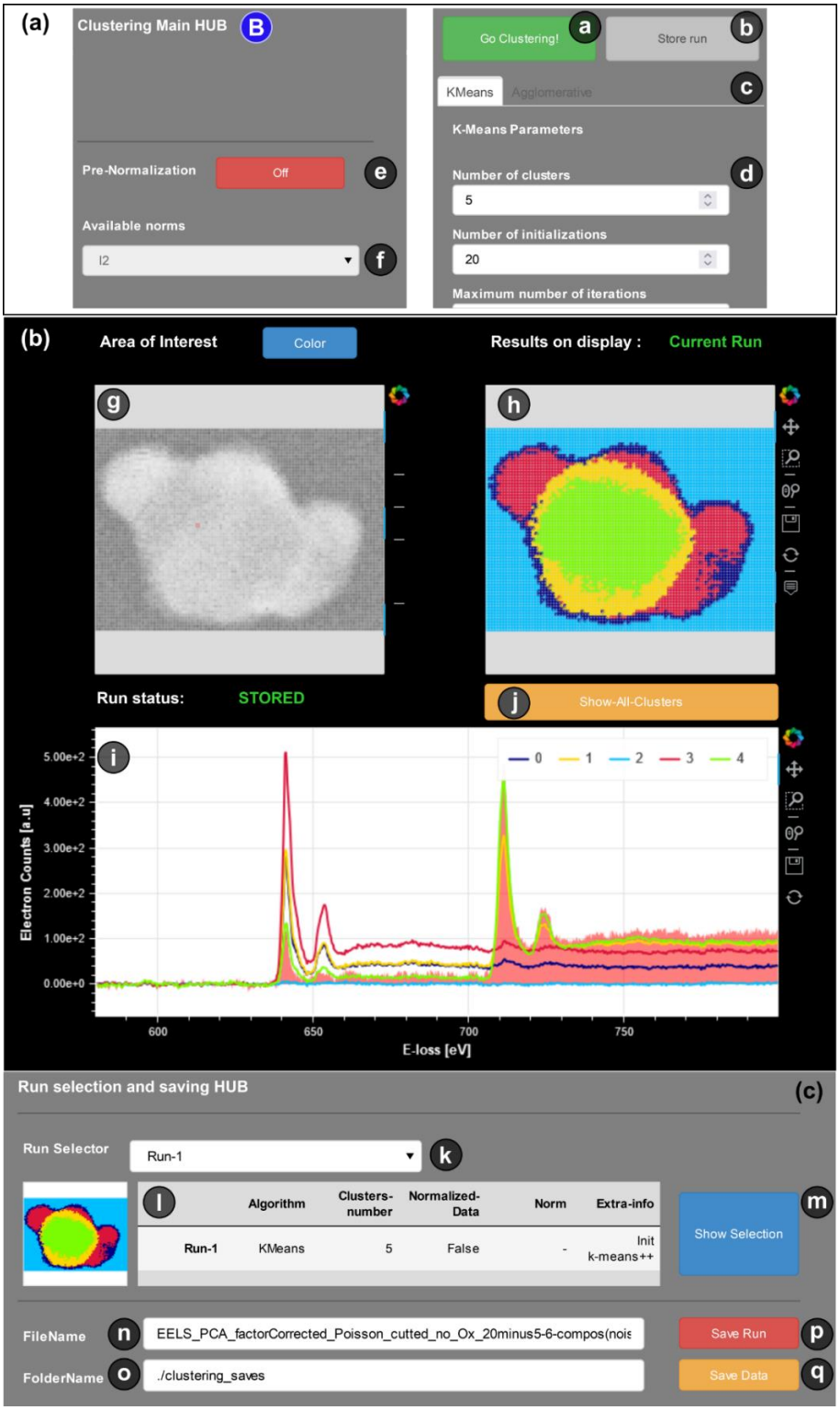


Figure 5.8. Tools of the module for the clustering analysis **-B-** in **WhatEELS**. (a) Customization options for the sklearn clustering algorithms and dataset normalization pre-processing. (b) Visualization tools for the clustering analysis module. (c) Controls to store the results.

for any spectrum from the EELS SI expressed as a vector $\vec{x} = \{x_1, \dots, x_m\}$. Each spectrum for the normalized spectrum image would be then $\vec{y} = \frac{\vec{x}}{norm(\vec{x})}$. The effects of applying a pre-normalization to an EELS clustering analysis are explored in chapter 4 as well.

The visualization tools (-g-/-h-/-I-/-j-) for the clustering **-B-** panel are shown in **Figure 5.8** (b). After a clustering classification is completed, the labelling segmentation results are shown as a colour map -h-. Any single pixel spectrum from -g- can be displayed superimposed with the centroid reference in -i-, by tapping the image -g- and double tapping the desired clustering reference area in -h-. All the centroids can be displayed at the same time pressing the button -j-.

The storage tools (-k-/-l-/-m-/-n-/-o-/-p-/-q-) are shown in **Figure 5.8** (c). The selector -k- allows to switch between different clustering runs previously

Quick-panel 2. Steps to achieve and store a clustering segmentation.

Segmentation results – K-means clustering

i. Select the objective number of clusters -d- on the K-means -c- parameters column.

ii. In this case, leave the normalization option deactivated. Including a normalization pre-step when using a SI with the background removed produces artefactual results.

iii. Visualize the resulting -h- clustering map and the reference spectra -i-

Notice that the labelling numbers and colours are assigned randomly to each cluster. Different runs with the same parameters may change the labelling.

iv. Store the current result -b-. The information in the storing area -k- and -l- will be updated.

v. Save the clustering results -p- in any given folder -o- (we recommend the name ./Clustering_saves, as the program will later look for this folder to automatically load clustering results).

stored -b-. The information and a small label map are displayed in -l- for the selected 'run'. At any given moment, any stored run -k- can be loaded to be visualized by pressing -m-. The completion bars -n- and -o- allow to set the directory and file-name to save (-p-) the clustering run permanently (i.e., for a posterior incorporation to the fitting routines). Also, the images in the display (-g-/-h./-i-) and the raw data can be saved at any time by pressing -q- and launching the saving panel (**Figure 5.23**) .

A quick guide is once again included (

Quick-panel 2). In this case, it shows how to attain a k-means segmentation and store the results for a later use. The steps are also valid for a hierarchical agglomerative clustering analysis.

-C- GOS module.

The GOS module is a special one in **WhatEELS** in the sense that it is the only tool that can be launched (aside from the dataset loader, launched at the beginning) without loading any specific dataset. It works as a visualizer of the Bethe surfaces (the 3D representation of the GOS functions[1]) and allows the users to test on them the effects of any combination of values for the experimental parameters: incident electron energy (E_0), convergence angle (α) and collection angle (β).

The GOS estimates for a series of scattering angles and energy loss values are read from the **DM** database. This repository contains the GOS for many inner-shell excitations of most of the natural-occurring atomic elements of the periodic table (i.e., up to the uranium), calculated from the Hartree Slater approximations of the atomic electrons wavefunctions. They are stored as a table of discrete values, and the continuous GOS surface is extracted by a cubic spline[36] extrapolation (see the introduction 1.1 for more information on the GOS function).

These GOS functions are also used by **WhatEELS** to compute the partial cross-sections for the elemental quantification using EELS spectra and they

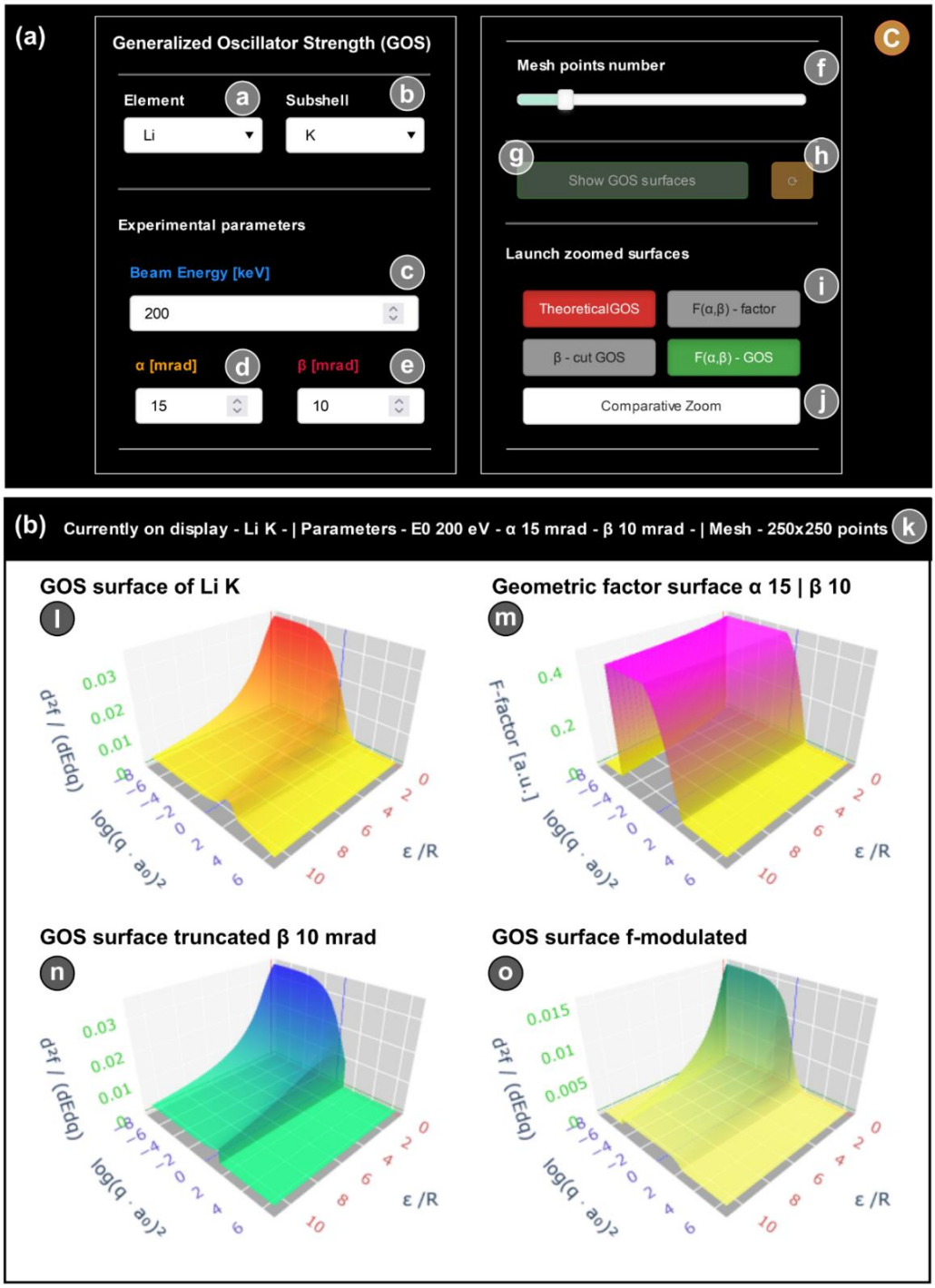


Figure 5.9 Panel for the GOS analyser module of **WhatEELS**. (a) Controls for the visualization. (b) Current visualization parameters selected and surfaces calculated.

are the basis for the functions used as the excitation to the continuum of states in the NLLS fitting routines (the ionization cross sections).

Nevertheless, these processes are automatically carried out by the model creation routine, and do not require the manual use of this module.

An image of the tools from this panel is shown in **Figure 5.9**. The controls in **Figure 5.9** (a) (i.e., tags -a-/to/-h-) allow the modification of the experimental parameters and trigger the visualization of the different GOS surfaces. **Figure 5.9** (b) shows the visualizations under the current selected parameters (title bar -k-). From top to bottom and left to right, the visualization panel includes -l- the tabulated GOS surface (read from the files), -m- the geometric factor surface calculated for every angle and energy value of the original GOS surface, -n- the GOS surface truncated by the $(qa_0)^2$ minimum and maximum values (conservation of momentum from eq.(5) and (6)) and -o- the GOS surface corrected by the geometric factor from -m-. Finally, the buttons -i- and -j- allow

Quick-panel 3. Quick guide to use the GOS module of **WhatEELS**, showcased in **Figure 5.9**

Visualizing the GOS surface for a set of experimental parameters

i. Select the element -a- and subshell -b- GOS to be shown.

ii. Set the experimental parameters to be explored (-c-, -d- and -e-).

iii. Once all the experimental parameters have values different form 0, the Show GOS surfaces button -g- is unlocked. Pressing it will start the calculations and prompt the update of the visualization panels (-l-,-m--n- and -o-). These parameters will be displayed in the header -k-.

iv. Changing the parameters without changing the element or subshell will unlock the refresh button -h-. This will produce new surfaces for the updated values of the parameters (repeat this process at will).

v. To visualize the resulting surfaces with a higher degree of detail, an extra panel may be launched from -j-. To unlock this option, select one or two of the surfaces (buttons -i-) to be visualized in the new panel.

the user to launch a new panel with one or two of the generated surfaces with a higher level of detail.

5.3.3. Core fitting components -2-.

-D- The model constructor module.

This is the main component of **WhatEELS**, the module in charge of the model creation, the adjustment of each component parameters and the multiple-pixel NLLS fitting (multifit) process. It also allows the integration of the clustering segmentation from the clustering module **-C-** into the model fitting routine.

An overview of the panel distribution is given in **Figure 5.10**. It is divided in 3 sections. (a) The model creation space and the clustering loading hub, (b) the model parameters configurator and multifit hub, and (c) the visualization and reference fitting tools.

The first task in any given model fitting routine is to create the model. To do so, at least one element must be included -b-. Whenever a subshell is activated in -b-, information about the onset energy of the edge -c- is displayed below. To add the selected component to the model, the button 'Add element' -e- must be pressed with at least one subshell active for the selected element. The 'soft-edge' subsection (-d-) allows the selection of the actual GOS surface integrated (q-axis) to get the ionization cross section function (1 of the 3 available options presented in the GOS analyser tool **-C-**, **Figure 5.9**). This function is later used to model the excitation to the continuum of states for a given inner-shell electron (edge).

Setting the green button in -d- as 'on' introduces an extra step in the model creation routine. The ionization cross section function (i.e., the spline interpolated from the integrated surface selected) is convoluted with a gaussian function. The fwhm of this gaussian is equal to the numerical value

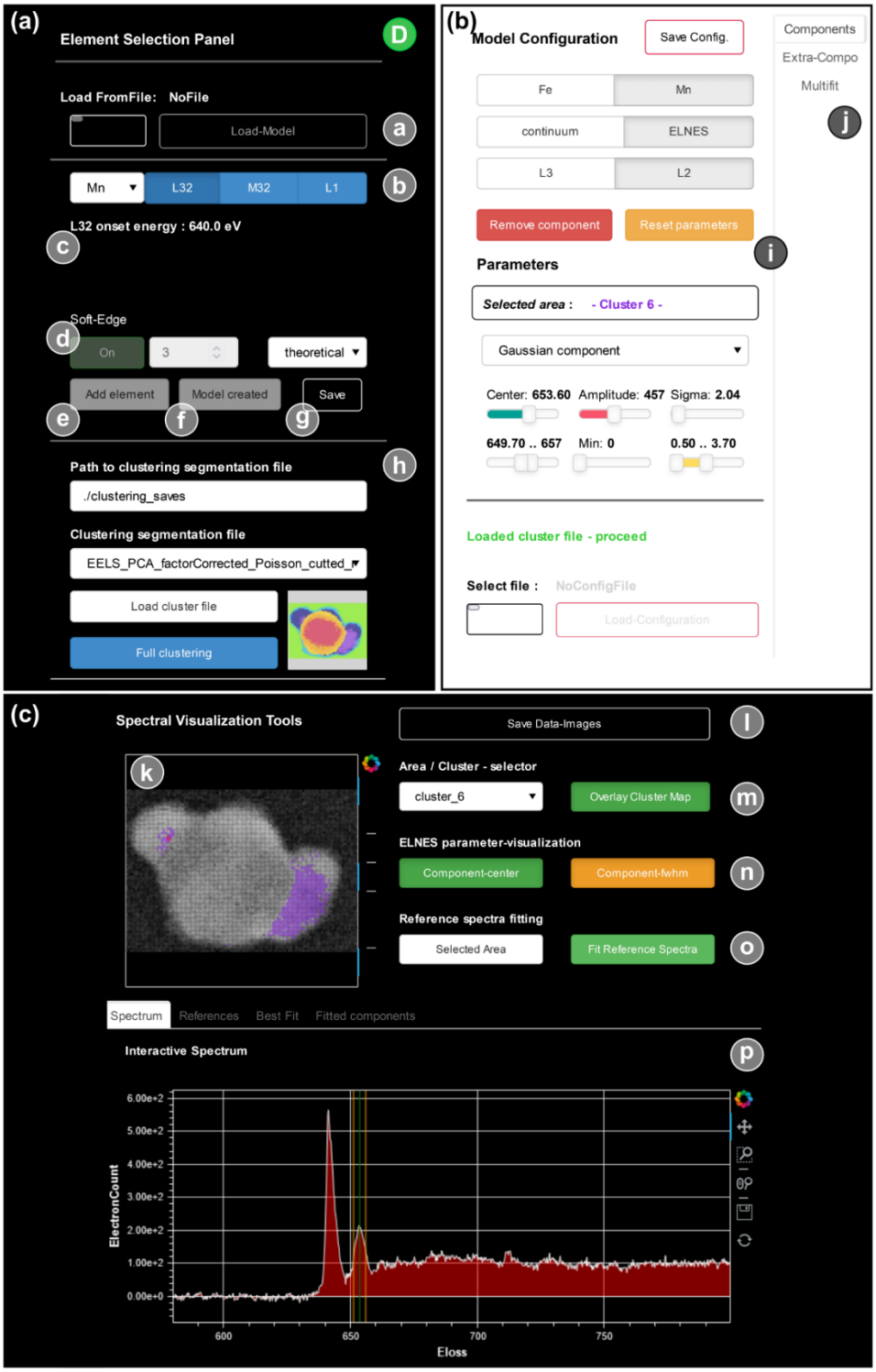


Figure 5.10. Tools from the model constructor panel for **WhatEELS**. In this case, a sample is already loaded, and a model has been created. Also, the results of a previous clustering analysis have been loaded into the module. (a) Model creation controls and clustering loading hub. (b) Parameters configurator and multifit HUB. (c) Visualization and reference fitting tools.

displayed to the right of the 'on' button. This pre-processing step can bypass some bad behaviours of the ionization cross section calculated from the GOS surfaces, such as sharp steps and peaks on the onset of the edges. Furthermore, it is also coherent with the experimental effects of having a finite energy resolution[41]. As such, the use of the same fwhm as the hypothetically measured zero loss peak (acquired under the same conditions and for the same sample) is recommended.

With at least one component added to the staging state, the button to create the model is unlocked -f-. The software will automatically create the fine structure for the added edges and will compute the continuum functions as well. In the case of M_{54},M_{32}, L_{32} and N_{54} subshells, two peak-like components (gaussian curves by default) are automatically added to simulate the white lines in the edge onset. In a L_1 subshell a single one is added.

The created model can be saved -g- for later use. Saving a specific model may help both with the reproducibility of the results and with the analysis of different datasets of the same sample, acquired under the same experimental conditions. The previously described model creation steps can be, therefore, bypassed by loading an already saved model -a- at the beginning.

After the model creation, the possibility to add a previously computed clustering segmentation **-C-** becomes available -h-. Adding a clustering segmentation to the NLLS fitting procedure is the signature characteristic of **WhatEELS.** It helps with the assignment of meaningful values for the component parameters for different areas of the dataset, separated by different spectral characteristics and identified by an unsupervised algorithm.

Precisely, the configuration panel **Figure 5.10** (b) is unlocked also once the model has been created. A detailed view of the two main tabs in this section of the panel is shown in **Figure 5.11**. The model configurator (**Figure 5.11** (a)) gives access to the components of the created model -b-, for each of the elements and subshells added. The parameters and boundaries for each component (ELNES or continuum) can be adjusted individually for each region

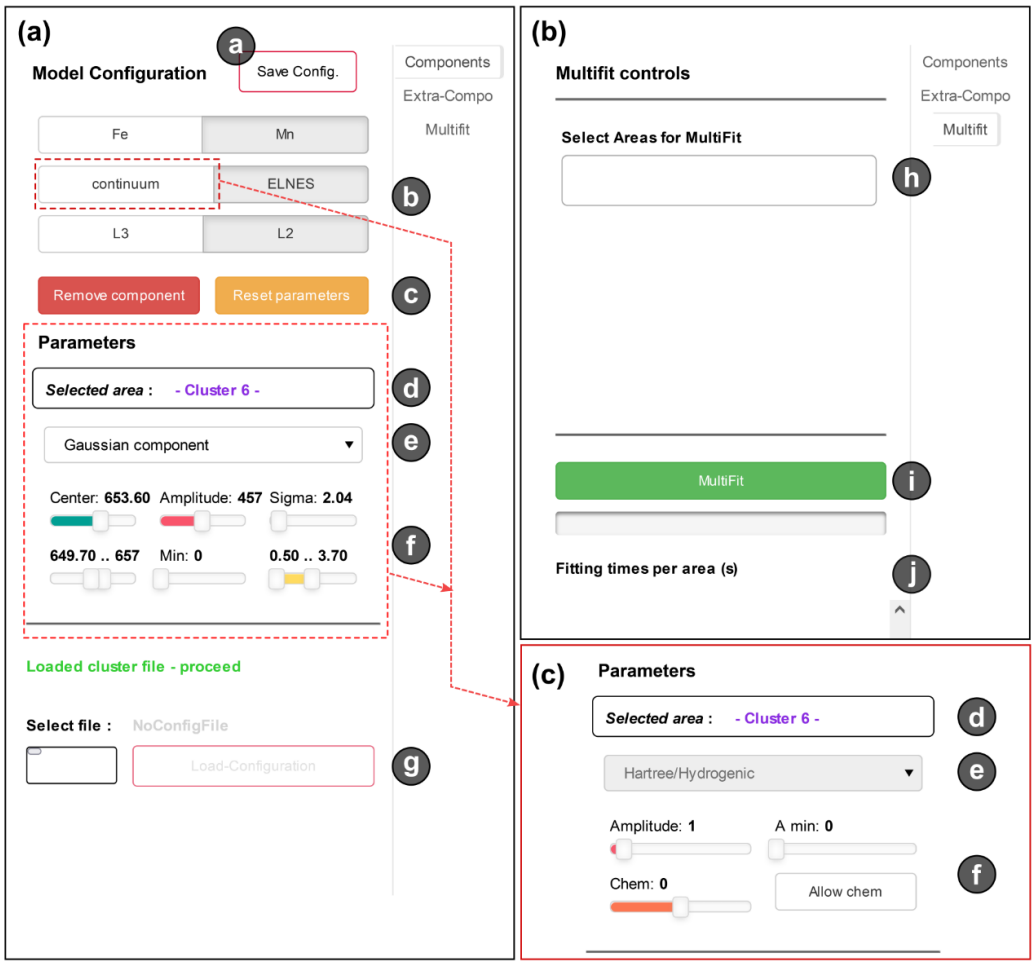


Figure 5.11. Parameters adjustment (a) and multifit column (b) from the model constructor panel in **Figure 5.10**. (c) ***Parameters*** Subsection of panel (a) when a continuum component is selected.

of the dataset. In case of having loaded a clustering segmentation reference, the current region being modified is indicated in -d-. If no clustering is loaded, the *default* area refers to the whole dataset, and the reference spectrum is the average spectrum of the whole set. Initially, the peak-like curves added to model the ELNES are automatically set as gaussian curves

$$f_{Gaussian}(E;c,A,\sigma) = \frac{A}{\sigma\sqrt{2\pi}}\exp\left[-\frac{(E-c)^2}{2\sigma^2}\right] \qquad (17)$$

where c stands for centre, and A is the amplitude. When the model is created (**Figure 5.10** -f-), the initial c value for each of the peak-like structure added is set as the edge onset value, read from the GOS database. The values for A and

sigma are approximated from the average spectra of the whole dataset initially. Later, if a clustering segmentation is loaded, all these initial values (c, A and sigma) are recomputed for each labelled region using the average EELS signal per cluster (centroid).

Three other options are made available for the ELNES peak-like structures through the **Figure 5.11** (a) -e- selector: Lorentzian[42], PseudoVoigt[43] or SplitLorentzian curves,

$$f_{Lorentzian}(E;c,A,\sigma)=\frac{A}{\pi}\left[\frac{\sigma}{(E-c)^2+\sigma^2}\right] \tag{17}$$

sharing the same variables of the gaussian curve, but being this time around a heavy tailed distribution.

$$f_{SplitLorentzian}(E;c,A,\sigma_l,\sigma_r)= \\ \frac{2A}{\pi(\sigma_l+\sigma_r)}\left[\frac{\sigma_l}{(E-c)^2+\sigma_l^2}\cdot H(c-E)+\frac{\sigma_r}{(E-c)^2+\sigma_r^2}\cdot H(c-E)\right] \tag{18}$$

where $H(c-E)$ is the Heaviside step function

$$H(E)=\begin{cases}1, & if\ x>0\\ 0, & if\ x\le 0\end{cases} \tag{18.2}$$

Notice that two sigma values are given for the SplitLorentzian, which is basically an asymmetrical version of the Lorentzian curve. The subindex *l* stands for left, and the subindex *r* stands for right. In the current implementation of **WhatEELS** only a single sigma parameter is accessed, which acts as the initial sigma value for both 'sigmas' (left and right) when fitting the model,

$$f_{PseudoVoigt}(E;c,A,\sigma,\alpha)=\frac{(1-\alpha)A}{\sigma_g\sqrt{2\pi}}\exp\left[-\frac{(E-c)^2}{2\sigma_g^2}\right]+\frac{\alpha A}{\pi}\left[\frac{\sigma}{(E-c)^2+\sigma^2}\right] \tag{19}$$

where alpha is the proportionality parameter for the mixture of Gaussian/Lorentzian curves that composes the PseudoVoigt model. Again, two sigmas are provided here. Only a single sigma parameter is accessed manually

in the model configuration panel, and it will act as the initial value for both (σ_g, σ).

If the component accessed -b- is the excitation to the continuum of states, the available customization options are fewer that those of the ELNES components (see **Figure 5.11** (c) -f-). Recall that these functions are the cubic spline models of the discrete values obtained from the numerically integrated GOS in the momentum axis (see section 1.1). As such, the general formulation would be the following

$$f_{continuum}(E; A, chem) = A \cdot f_{spline}(E + chem) \qquad (20)$$

where A is a multiplicator (the amplitude parameter) and chem stands for the 'chemical displacement', which would be a translation in the E axis of the initial function. By default, the chem parameter is set as 0 and locked in place. It is possible to unlock it by activating the allow-chem button (not recommended).

The model configurator tool also allows the user to remove any given component from any specific reference area (default or cluster) of the dataset, or to reset all the parameters from said area to their initial values -c-. The removal of components for specific edges is still an unstable feature, so muting (setting to 0) the amplitude of the unwanted component is the recommended path.

All these values for the components parameters can be saved in a file at any moment -a-, for their later use in a similar dataset or for reproducibility reasons. Likewise, a set of values for the model parameters may be loaded -g- from a saved file at any moment.

Before launching the multifit (**Figure 5.11** (b) -i-) for the whole dataset or smaller subsections described by a clustering classification, the reference spectra for each of the areas of interest have to be previously fitted, the reason being that every pixel for each area in the multifit will be initialized with the parameters picked from the fitted values of its corresponding reference signal.

The fitting process for the reference spectra is controlled from the visualization panel (**Figure 5.10** (c)). This subfigure is at the same time subdivided in 2 regions. The upper part contains the actual controls for the fitting process, and the lower part contains a series of tabs for visualization of the results.

The dataset under analysis (in this case an EELS-SI, **Figure 5.10** (b) -k-) is shown as an interactive image. The spectrum of any pixel can be inspected in the '*Spectrum'* tab -p- below by selecting the pixel directly from the image. To visualize the reference areas for each of the clusters loaded (if any), the Overlay Cluster-Map option -m- has to be activated. Changing the reference areas (clusters) in the area-selector -m- would change the coloured area over the image -k-. It would also change the parameter values displayed in -i-, to accommodate the ones corresponding to the area selected. In order to facilitate the configuration of the initial values of the parameters (only centres and sigmas) for the ELNES components, the ELNES parameter-visualization options -n- have to be activated. Then, a green line representing the central position of the component and two yellow ones marking the sigma separation are drawn over the Spectrum representation -p-. Once the parameters are adjusted for the area selected, the reference spectrum can be fitted by activating the Fit-reference-spectra -o- button. The whole set of possible reference spectra can also be fitted at once, by changing the status -o- of the '*selected-area'* button to '*all-areas*'.

The rest of the possible visualization graphs (the different tabs in -p-) are showcased in **Figure 5.12**. The resulting fitted spectra can be visualized as whole (i.e., the best- fit for the model, (b)) or as the separated fitted components of the model (c).

This process of fitting the references can be done iteratively, to fine tune the initial parameters that will be used for each cluster area in the multifit routine.

Finally, one may proceed to perform the multifit (**Figure 5.11** (b) -i-) for such prepared areas. It is not mandatory to have fitted all the available areas, but

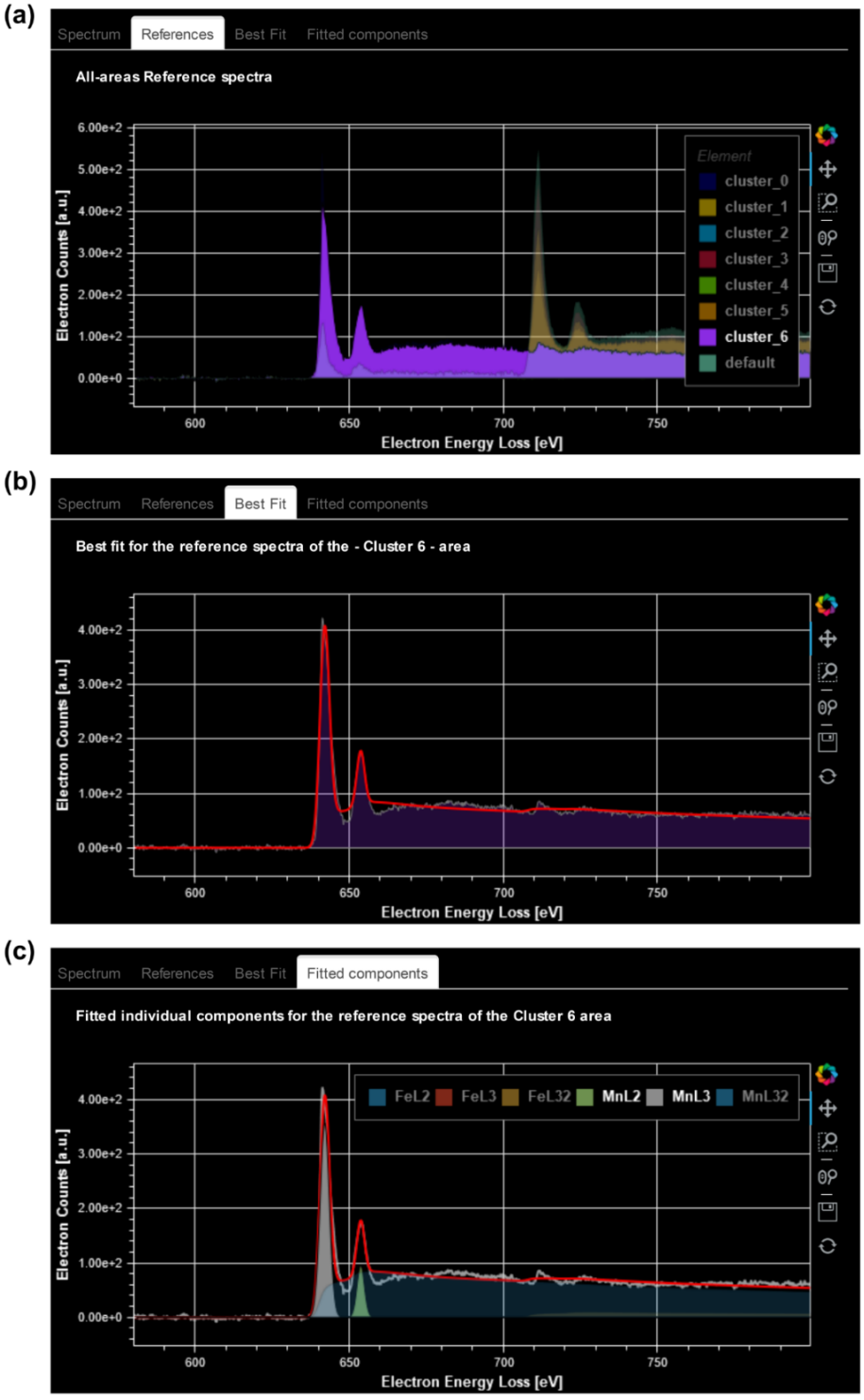


Figure 5.12. (a) Reference spectra tab. When a clustering is loaded all the reference spectra are added to this graph. (b) Best-fit model for the reference area. (c) Fitted-components tab contents after performing the fitting of the reference spectra of cluster 6.

Quick-panel 4. Quick steps to get a model fitted by NLLS, using the information of clustering segmentation previously calculated.

Model construction and dataset multifit

i. **Add the specific edges for each element (Figure 5.10** -b- **)** in the sample to the model creator list **(Figure 5.11** (a) -b- **)**. Selecting the subshell of a given element will indicate the expected onset energy loss **(Figure 5.10** -c- **)**, read from the database. After adding all the elements, create the model **(Figure 5.10** -f- **)** and the rest of the configuration tools will be unlocked.

ii. **Load the clustering segmentation maps (Figure 5.10** -h- **)** Although not mandatory, it is a highly recommended action.

iii. **Configure the parameters for the model components and their boundaries (Figure 5.11** (a) and (b) -d- , -e-, and -f- **)**. To visualize the actual selections, activate the show-centre and show-sigma buttons in the visualization panel **(Figure 5.10** -n- **)**. Also, beware of the area selected **(Figure 5.10** -m- **)**. The initial parameters values are independent for each segmented area (cluster). Although repetitive in some cases, this process ensures a faster convergence later on if the parameters values and boundaries are set with caution for each area.

iv. Evaluate the results after **fitting the reference spectra (Figure 5.10** -o-**)**. Each area/cluster reference spectrum may be fitted independently (if we are only interested in modifying the parameters in a specific area of the sample). Conversely, they can be fitted all at once by clicking the SelectedArea button and set it as AllAreas. An accurate visual match between reference spectrum and proposed model **Figure 5.12** (b)-(c) is good enough at this stage to evaluate if further modifications are needed or if, on the contrary, extra modifications are needed on the initial parameters and boundaries of each component.

v. **Multifit** - Launch the model NLLS fitting routine for the whole dataset (default), or for any given number of segmented areas via clustering that have a fitted model for its reference spectrum **Figure 5.11** (b) -h- and -i-.

only those with a reference spectrum fitted may be selected for the multifit (**Figure 5.11** (b) -h-). The progress bar will start over every time that a new cluster is begun. The fitting times of each individual cluster will be iteratively added to the area below the progress bar (**Figure 5.11** (b) -j-).

This procedure of selecting specific areas to be fitted at any given time is particularly interesting when one is only interested in particular areas of a given sample (e.g., the grain boundaries, or interphases between different layers).

Finally, once the multifit process is completed, the results analysis tool becomes available (**Figure 5.7** (b) and **Figure 5.13** -E-).

-E- The results analyser module.

The results analyser -E- can be launched from the main HUB (**Figure 5.6** and **Figure 5.7** (b)) once a multifit process has been completed -D- (**Figure 5.10** and **Figure 5.11** (b)). An image of the tools from the panel is shown in **Figure 5.13**. It is divided in 2 sections.

Figure 5.13 (a) Contains the visualization controls that allow the user to overlay -d- the clustering segmentation image with the $r \cdot \chi^2$ mapping in -h-, and to change -c- the colour-theme. It is also the placeholder for the buttons that launch the specific analysis tools: -E_1- the white-lines ratio analyser, -E_2- the components centres and features distances analyser and the -E_3- EELS quantification tool. Any of these tools can be launched using -e- the multifit results from the first run (directly after the model multifit, **Figure 5.11** (b) -i-), or using the secondary modified fit (when available, in a future iteration of the software). This panel gives access to the -C- GOS analyser tool -f- once again. However, only the Bethe surfaces for the edges included in the model under analysis and with fixed experimental parameters (the ones used for the model creation) can be accessed. The surfaces to be displayed can be selected -g- between: (1) the tabulated GOS or theoretical surface, (2) the F(geometric)-

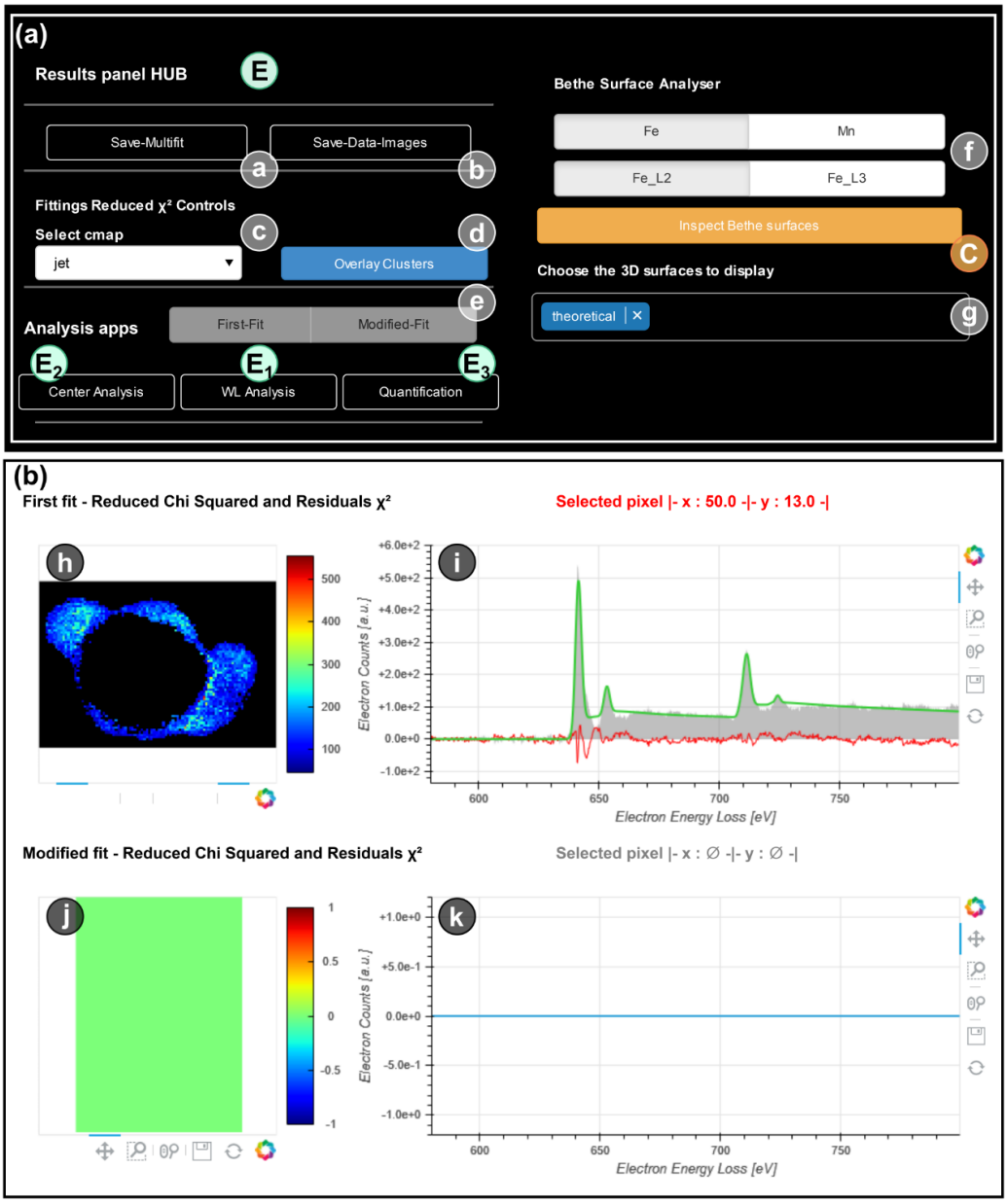


Figure 5.13. Results analyser panel for **WhatEELS**. The specialized tools for the components centre, white lines ratios and quantification analysis are accessed from this panel. (a) Main controls for the themes in the visualization and launchers of the specific results analysis modules. (b) Visualization area for the $r \cdot \chi^2$ and the best fit and residual curves per datapoint.

factor surface representation, (3) the GOS surface hard-cut by the experimental angular constraints and (4) the GOS surface corrected by the geometric f-factor.

Figure 5.13 (b) contains the visualization heatmap -h- for the reduced chi squared[17] ($r \cdot \chi^2$) coefficients for each of the datapoints fitted (i.e., pixels with a fitted model) and the visualization graphs -i- that overlay the original EELS signal, the fitted model and the residual for the model (i.e., the difference between the data and the model fitted) for each pixel (interactively updated by selecting different pixels on the heatmap -h-). The same visualization structure is repeated below (-j-/-k-). These areas are a placeholder for a potential second fit (using the optional modules **-F1-** and **-F2-**, **Figure 5.7** (b))

with modified components, to allow a comparative analysis of the two proposed models. At the moment of writing this text, this feature is included to future-proof the results module **-E-** , as the pre-analysis **-F1-** and the expansion-model **-F2-** modules are not fully operational yet.

Finally, the multifit result **-D-** can be saved from this panel (-a-). This would enable the initialization of the results panel from a saved state (**Figure 5.7** (b) -a-/-b-).

[17] The $r \cdot \chi^2$ or χ^2_v coefficient is defined as chi square coefficient χ^2 per degree of freedom v.

$$\chi^2_v = \frac{\chi^2}{v} = \frac{1}{n-m}\sum_i \frac{(O_i - M_i)^2}{\sigma_i^2}$$

being O_i the observations (in this case, $\vec{x}_i$ the spectral data), M_i the fitted model (i.e., $\vec{x}'_i$), σ_i the variance of the data, n the total number of observations (the number of channels in the spectra) and m de degrees of freedom (centres, amplitudes and sigmas per peak-like component and amplitudes for the excitation to the continuum functions).

Usually, a model will be deemed successful (accurate) if the χ^2_v test returns a value under 1. However, that is rarely the case for NLLS fittings in EELS data analysis, where a large number of components are usually considered, and the norm is facing noisy datasets. Hence, the actual objective of a χ^2_v test for NLLS fittings in EELS data analysis is to assess the homogeneity of the values retrieved, which would indicate that every area of the dataset is fitted to the same degree of fidelity.

-E$_1$- The WL ratio analyser.

This module allows the user to investigate the white lines ratios for the edges added to the model in **-D-** after the multifit. In the literature, measuring the changes in WL ratios has been one of the most successful methods to resolve possible changes in the oxidation state of materials through EELS analysis[4,22,44]. This methodology has been especially useful for the cases of transition metal (TM) and rare earth (RE) oxides (as those are the materials that generally show WLs features in the ELNES regions of the spectra).For example, cerium oxide will present an abrupt change in the M_5/M_4 ratio for the Ce WLs whenever the cerium oxidation state transitions from Ce^{4+} to Ce^{3+} in a reduction process of the CeO_2 to Ce_2O_3. [45]

The tools from the panel launched are shown in **Figure 5.14**, divided in 3 different sections. **Figure 5.14** (a) shows the dataset visualizer -a-/-b-that allows the user to access the spectral information. It works in tandem with the -visualization controls -n- in **Figure 5.14** (c), which also contains the WL ratio map -m- visualizer once it has been calculated. This control panel from **Figure 5.14** (c) -n- allows the users to invert the WL ratio shown -m-. It can also activate the overlay of centres and integration windows over the spectrum visualizer in **Figure 5.14** (a)-b- (only when a manual integration width has been selected for each pixel **Figure 5.14** (b)-f-/-h-). Finally, it also contains the controls for the colorbar limits and theme for the ratio heatmap -m-, and the launcher for the advanced analyser tool -o- (the latter, only when the advanced-fwhm -e-/-f- option has been selected and computed).

Figure 5.14 (b) shows the customization options for the WLs ratio calculations grouped together. These WL ratio calculations can be carried out for the fitted edges -d- in the model that contain at least 2 ELNES components. This is a straightforward requirement, as calculating a ratio implies that a minimum of two fine structure components are available. As a single edge may contain more than 2 ELNES components (in future versions, extra components can be added in **-C-**), 2 of them must be selected -g- to carry out the calculations.

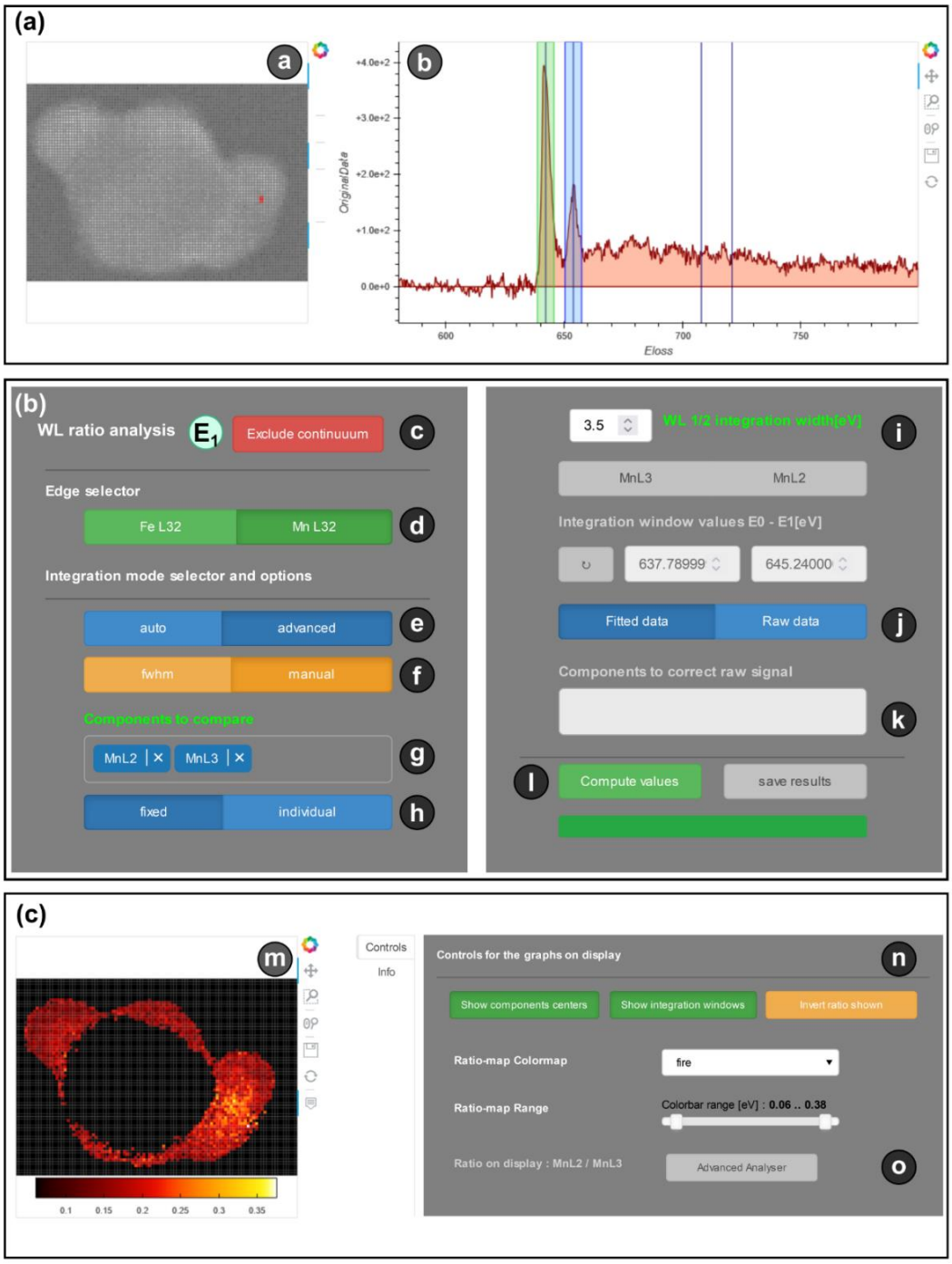


Figure 5.14. Panel for the white line ratio analysis in **WhatEELS**. (a) Dataset visualization and visual guidelines for the integration windows used to compute the WL ratio. (b) Controls for the edge and WLs selection, and integration mode configuration. (c) WL ratio heatmap visualizer -m- and -visualization controls -n-. It also gives access to the advanced analyser -o- whenever the advanced/fwhm -e-/-f- option is selected for the ratio computation -l-.

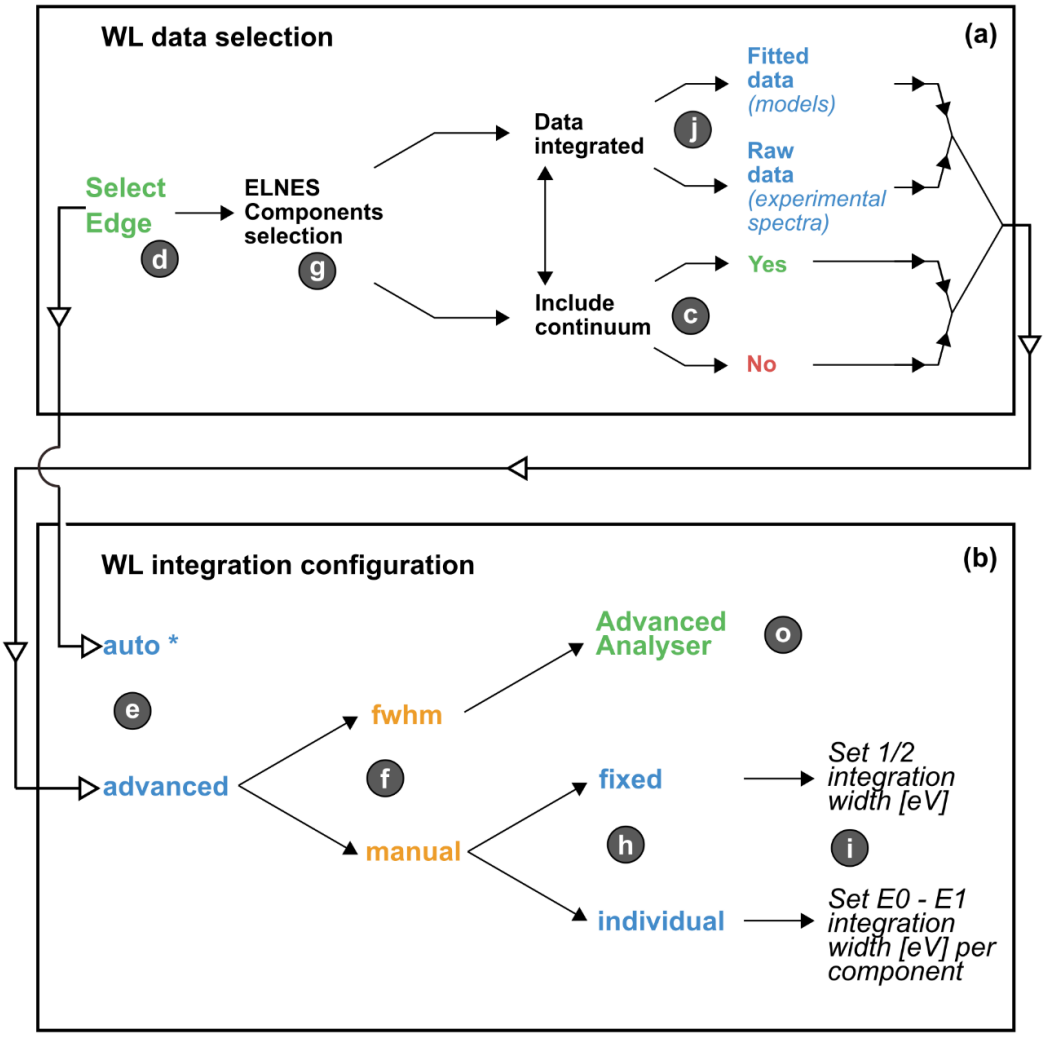


Figure 5.15. Flow chart for the WL ratio analyser (shown in **Figure 5.14**). (a) Selector options for the dataset used to compute the ratios. (b) Selector options for the integration width configuration.

The WL ratio can be automatically calculated -e- for any valid edge. The integration parameters would then be internally selected by the software, extracted partially from the fitted models. However, some advanced -e- options for the integration are available. They can be grouped in two blocks (**Figure 5.15**).

(a) The data selection for the integration of WLs. With the -d- edge selected, two handpicked ELNES components from the model can be chosen -g- (whenever a non-automatic ratio calculation -e- is activated). Furthermore,

the integration can be done -j- for the raw spectral data and the model fitted data. Finally, the fitted excitation to the continuum function for the selected edge can be subtracted from both the raw and/or modelled data before carrying out the integration.

(b) The WL integration configuration. If the advanced mode for the integration of WL is selected -e-, two types of integration options become available -f-.

(b1) The fwhm option will compute a set of ratios for every datapoint. Each computed ratio (for each datapoint) uses for each pixel the maximum among the model-resolved fwhm of each of the components selected -g- as the base for the width of the integration window. This base is multiplied by a set of factors ranging from 0.25 to 6 to get the whole set of ratios per datapoint. A representative example from the set of ratios calculated is displayed in -m-, corresponding to a multiplying factor of 3. When this path is chosen, the advanced analyser tool -o- is unlocked, giving access to the analysis of all the ratios computed.

(b2) The manual -f- option allows the user to set a fixed -h- value for the width of the integration windows (for every component considered), or to set manually an individual -h- width value for each of the components selected (setting the left and right limits of energy loss values for the integration window -i-).

The advanced analyser is shown in **Figure 5.16**. The spectra from original dataset can be interactively visualized -a-/-b- overlayed with the data used for the WL ratio calculations (activating -c-). To access this spectrum-per-pixel information make sure that the highlighted tool -a_1- is activated.

Each time that a specific pixel is selected -a-/-a_1-, the scatter plot that shows the stored ratios-per-pixel-per-integration-width -f- will be automatically changed as well. Each of the points in said graph represents the ratio from a different integration width for the selected pixel -a-. The x-axis -f- represents the multiplier value for the modelled fwhm. Larger values of the multiplier

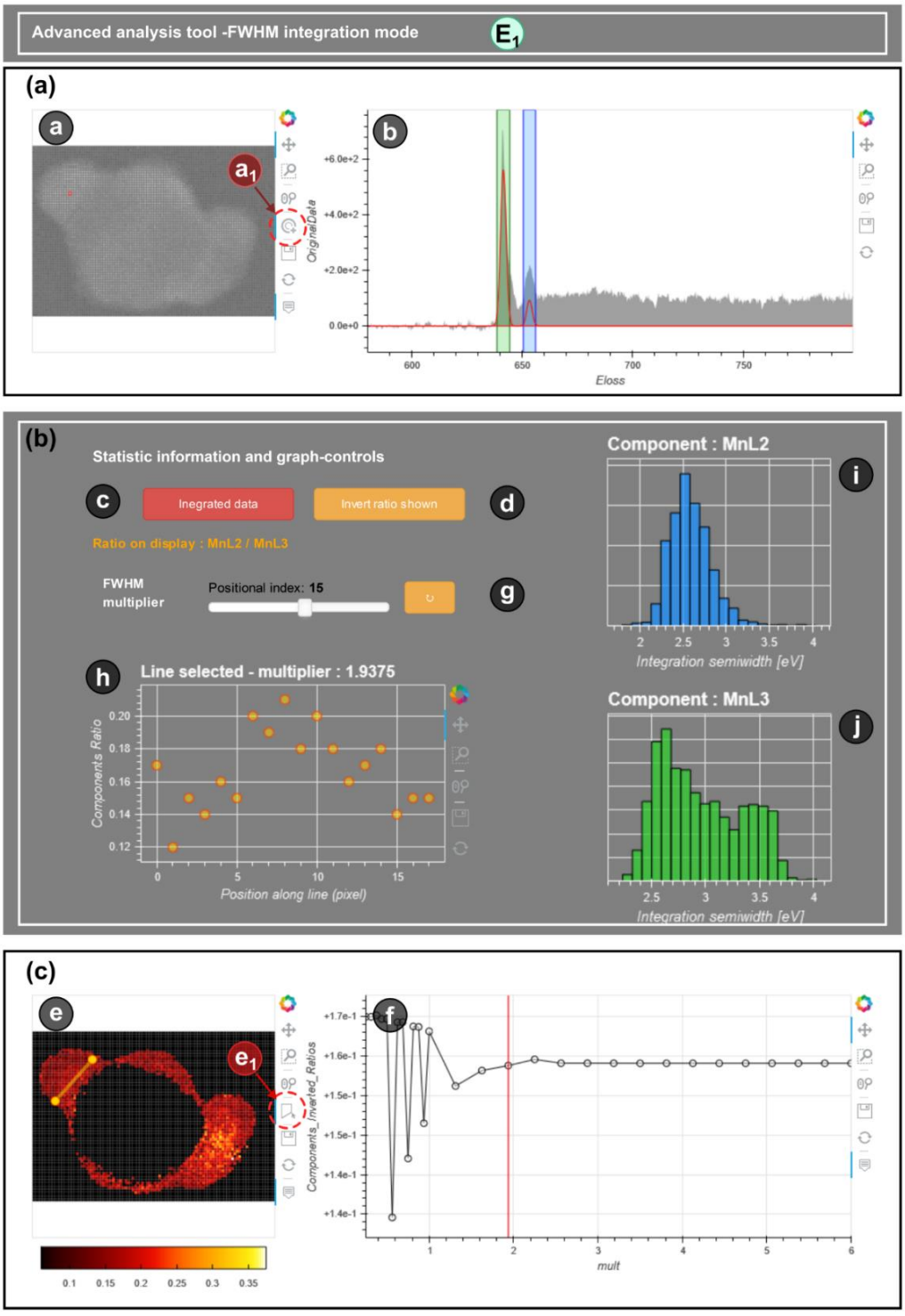


Figure 5.16. Advanced WL ratio analyser. This panel can be accessed when the advanced-fwhm pathway was followed in the main WL-ratio analyser -E1- panel (**Figure 5.14**). (a) Integrated data and windows visualizers. (b) Controls for the multipliers, 1D lines graph representation and integration width statistics. (c) Ratio-by-multipliers heatmap and scatter plot (per pixel) visualizers.

(*mult.*) correspond, thus, to larger integration windows widths. These multipliers can be accessed individually by the -g- slider widget. Whenever the position of this index -g- is changed, the vertical red line overlayed on top of the ratios scatter-plot -f- will be displaced along the abscissa axis to mark the specific ratio selected.

The WL ratio heatmap -e- will also change automatically whenever the multiplier selector is changed. Furthermore, the visual representation of the integration windows for each of the components selected (green and blue windows in -b-) will be automatically updated.

Quick-panel 5. Quick guide to get a WL ratio analysis (**Figure 5.14**, **Figure 5.15** and **Figure 5.16**)

WL ratio – Manual setup for the calculations

i. Select the edge **Figure 5.14** (b) -d- and ELNES components (WL) to be integrated.

ii. Select the data to be included in the integration (see **Figure 5.15** (a))

iii. Set the desired mode of integration.

The recommended path is to begin with an advanced/fwhm (Figure 5.14(b) and Figure 5.15(b)). Later, the WL analysis can be repeated for a single width.

iv. Launch the advanced analyser tool, and inspect the statistical distribution of widths and the WL ratio maps produced for each of the multipliers in the integration (**Figure 5.16**).

v. Repeat the WL analysis for a single width value. Try setting this value as the one that presented the overall better stability (a minimum) in the scattering multiplier-v-ratio plot per pixel (**Figure 5.16** (c) -f-).

These steps can be repeated as many times as needed, checking the different results that different dataset integrations (**Figure 5.15** (a)) produce.

The heatmap for the WL ratio results -e- allows the user to draw a path-line over the SI area (activating the highlighted tool -e_1-). The objective of such action is to extract a 1D plot of the pixel-WL-ratios in areas of interest (e.g., crossing a known gran boundary). This plot will be automatically updated in the panel (**Figure 5.16** (b) -h-).

This line is obviously dependent on the current multiplier selected. Changing the multiplier -g- will also change the line selected in -h- (it requires a manual refresh pressing the button in -g-).

Finally, the advanced WL analyser panel contains two histograms -i-/-j- to provide some of the analytical statistics of integrations widths for a given multiplier. Recall that each of the pixels presents a distinctive fwhm as base reference, depending on the values set by the fitted model. Hence, a quantity of integration widths equal to the number of datapoints times the number of multipliers is stored in memory. When the multiplier is adjusted -g- , the histograms -i-/-j- are automatically updated. Knowing the statistical distribution of widths and their relationship with the ratio map -e- may be helpful to select (later) a constant integration width in the main panel.

-E_2- The centres position and feature distances analyser.

This module allows the user to investigate the central positions of the peak-like structures (i.e., the Gaussian, Lorentzian, SplitLorentzian or PseudoVoigt functions fitted to the ELNES region of the edges of interest). In many instances, changes in the oxidation states or atomic coordination of a certain material can be traced by measuring the shifts in the energy loss positions of the ELNES structures (e.g., the shift to lower energy loss values of the cerium edge white lines when it is reduced from Ce^{4+} to Ce^{3+} in a cerium oxide material[2,46], or the same type of behaviour for the manganese white lines when a manganese oxide material changes between MnO and Mn_3O_4 phases[4,20]).

However, tracing the energy loss position of a single feature can lead to an erroneous conclusion, influenced by random shifts of the zero-loss peak in an

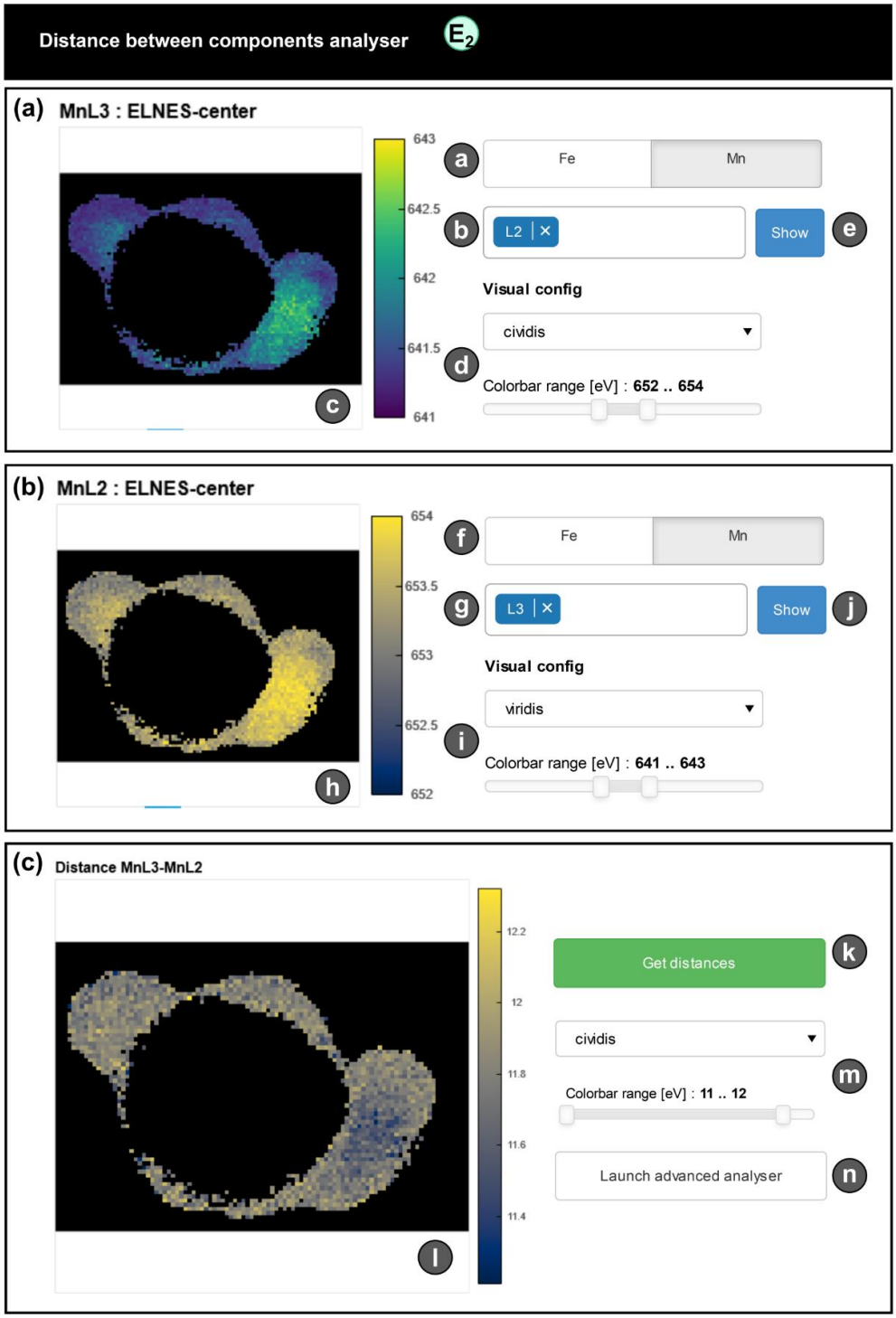


Figure 5.17. Tools for the centre-analysis module of **WhatEELS**. (a) First component selector and visualization controls. (b) Second component selector and visualization controls. (c) Distances map between the components 1 and 2 selected in (a) and (b), visualization controls and advanced analyser launcher.

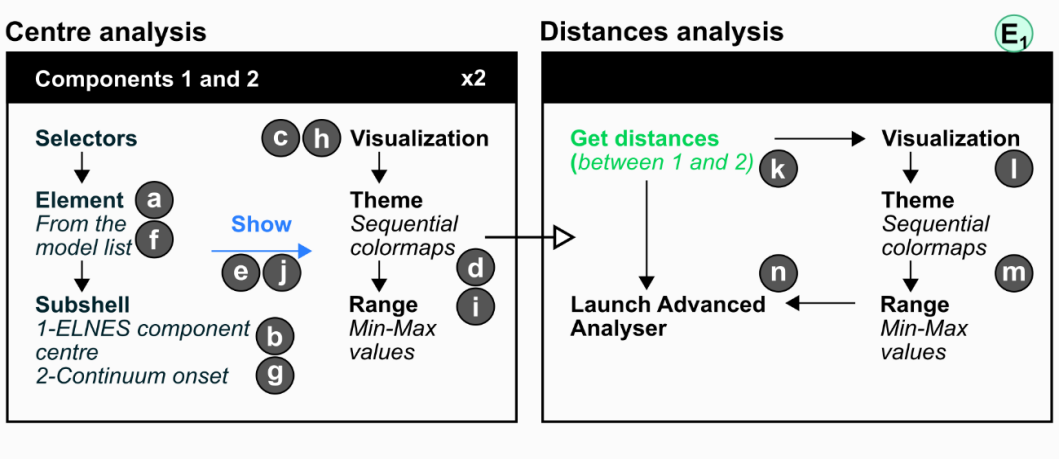


Figure 5.18. Flow chart for utilization of the centres analysis tool **-E2-** of **WhatEELS**.

unstable EELS acquisition. In the literature this problem is sometimes tackled by measuring the energy distances between two different features of the spectrum, assuming that one of them is theoretically known to stay fixed in an energy loss position[20,24]. This feature is also available in this module. Furthermore, the onset energy of edges corresponding to an electron excitation from the K atomic inner-shell (e.g., the O-K edge) are also selectable features this time around.

Figure 5.17 shows the set of tools available for the panel launched by this module. **Figure 5.17** (a) contains the selection and visualization tools for the component 1. The first step for the analysis of centre positions is to select the element containing the desired component. Only those elements with peak-like ELNES structures or K-edges are selectable -a-. Then, a single subshell can be picked -b- (i.e., a single gaussian component, or the K-edge onset) from the available components for the element picked -a-. The centre values mapping -c- shown (pressing -e-) for the element -a- and component -b- selected can be customized with the desired colormap and maximum/minimum values -d- for the range of energy loss values represented.

Figure 5.17 (b) shows the equivalent set of tools (to the ones in **Figure 5.17** (a)) for the selection of the second component under analysis. Once that both components have been selected, the inter-component distance can be

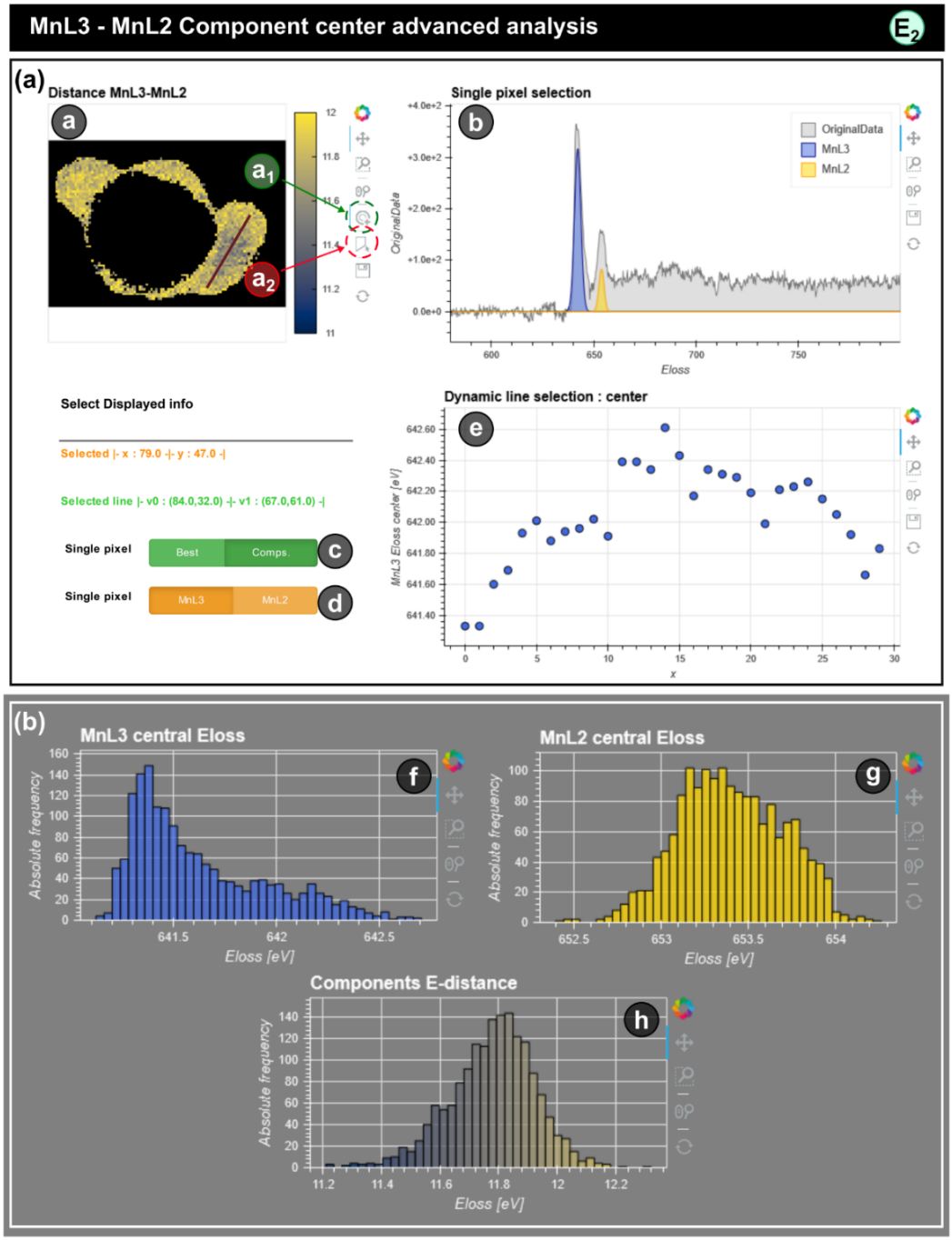


Figure 5.19. Tools from the advanced centres and distances analyser, **WhatEELS**. (a) Visualization tools. (b) Statistical distribution for the analysis of centres and inter-component distances.

computed and displayed (**Figure 5.17** (c) -k-/-h-). This panel also allows for a simple customization of the themes for the distances figure shown -m- .

The centres position and distances analyser tool **-E$_2$-** also gives access to an advanced-analyser panel. A flow chart with the chronological ordering of tasks required to get to the advanced-analyser tool is shown in **Figure 5.18**.

The advance analyser tool is shown in **Figure 5.19**. It contains two distinctive areas: (a) the results visualization area, and (b) the statistical information area. The interactive heatmap -a- for the distances between the centres of the components selected (or the distance from a component centre to the K-edge onset selected) allows the selection of single pixels using the tool -a$_1$-. The original spectrum for the selected pixel is then displayed -b- alongside the areas of the components selected (1 in blue and 2 in yellow). The best-fit curve for the model from **-D-** can be displayed -b- instead, selecting this option in the buttons in -d-.

The user can also draw lines -a$_2$- to inspect the evolution of the centre values of the selected components in specific areas. The line-values are displayed in the area -e-.

Finally, a series of histograms for the statistical distribution of the selected centres for the whole dataset fitted -f-/-g-, as well as the distribution of distances between such components, are included in the panel. Skewed distributions, or bimodal ones, may hint the existence of a real chemical shift in the energy-loss values for the components fitted in the model.

-E$_3$- EELS quantification tool.

This is the last of the results analysis tools included in **WhatEELS**. It allows the user to carry out an elemental quantification analysis based both in the original raw dataset and the fitted models (**-D-**). The quantification is done following the Egerton's method (see section 1.1.2, eq.(9)). The partial (integral) cross sections are extracted from the same database used by the model constructor to get the continuum excitation curves. This module is

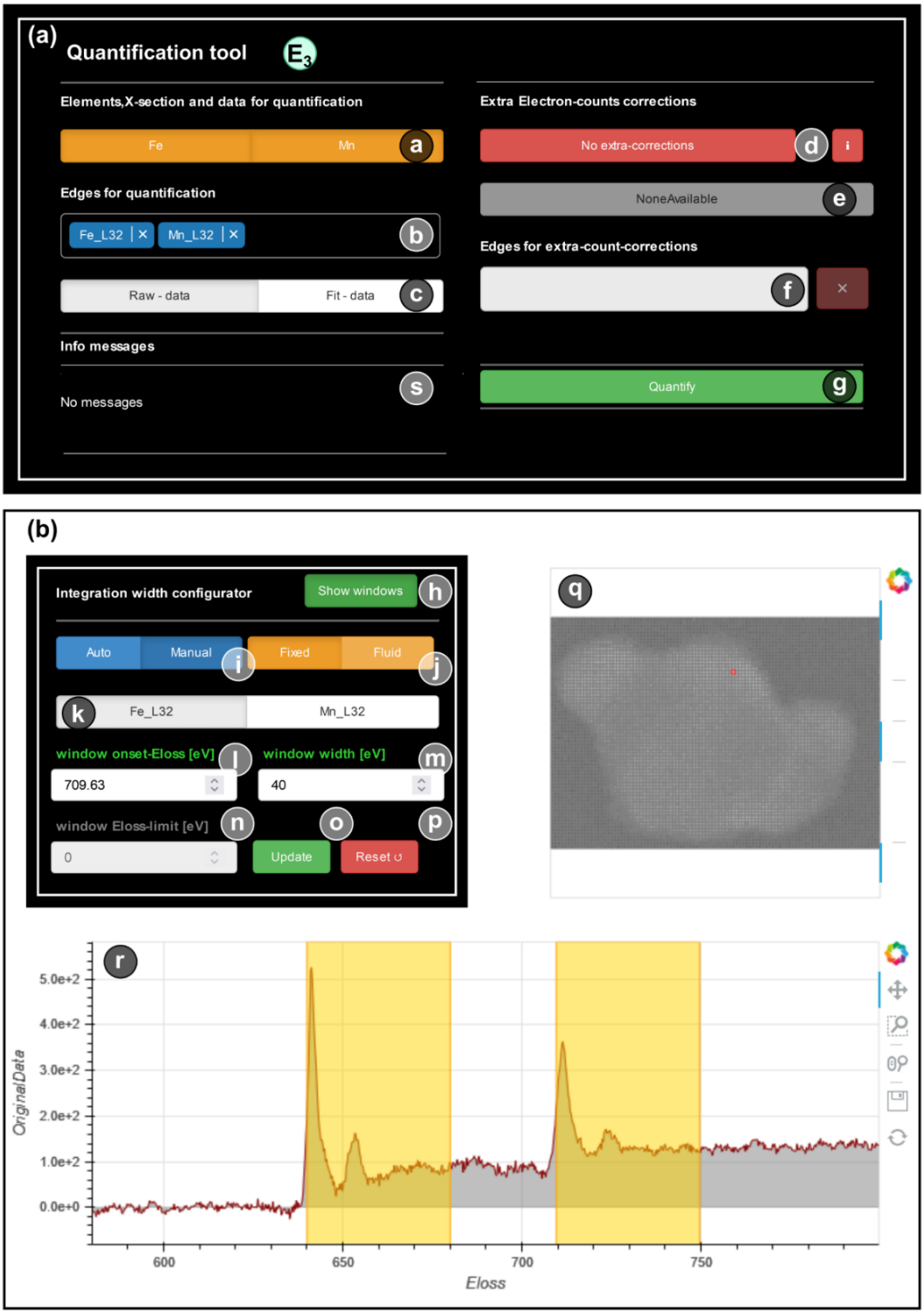


Figure 5.20. Tools from the configuration panel for the elemental quantification in **WhatEELS**. (a) Data selection panel for the quantification. (b) Integration windows configurator.

subdivided in two different panels, one to set up the integration limits and edges to quantify, and a second one to visualize the results.

The tools from the configuration panel for the quantification are shown in **Figure 5.20**. This panel is subdivided in two regions. **Figure 5.20** (a) contains the tools used to select the elements to be quantified -a- and the specific edges from those elements to include in the integration process -b-. Both the elements and edges available are listed from the ones added to the model created in **-D-**.

It also allows to add extra corrections -d-/-e-/-f-. These corrections act over the total integrated intensity of the edges specified -e-, and consist of the subtraction of counts belonging to potentially unaccounted edges that interfere with the ones added to the model in the quantification-process -b-. For example, let's consider the case in which a minor edge unaccounted in the model (i.e., not added to the model) lies bellow an edge (i.e., is located in the same energy loss range) added to the model and relevant for the elemental quantification (i.e., that will be used in the elemental quantification). This could cause an overestimation of the integrated counts for the quantification and, thus, it can introduce bias into the results obtained. The reason behind this is that the signal produced in that energy loss range could be traced back to both excitations from the inner-shell for the major and relevant edge and from the inner-shell for the minor and marginalized edge.

The steps to compute the corrections are the following. **First (1)**, **WhatEELS** carries out an evaluation of possible corrections for the spectra. In order to be able to make any correction at all, the major edge paired with the minor edge must be included within the spectral energy axis and in the model fitted **-D-**. The software will handle this issue automatically by reading all the onset energies at once from the database and tracking the elements from the model. **Second (2)**, using the experimental parameters provided, the ionization cross sections and partial cross sections for both major and minor edges are calculated from the database of GOS surfaces. A ratio is extracted from the partial cross sections, which accounts for the theoretical ratio of expected

electron counts for the minor edge from the electron counts in the major one. **Third (3),** the integrated intensity (modelled data) for the paired major edge is multiplied by the ratio calculated to obtain the expected counts for the minor edge. The energy window for the integrations in the second and third steps corresponds to the energy window placed in the area to be corrected. **Finally (4)**, these electron counts estimated for the unaccounted minor edge are subtracted from the integrated intensity of the affected (and included in the model) edge for the quantification process.

This correction step is not mandatory (and not always allowed, depending on the edges and energy ranges considered). In any case, when activated -d- in the quantification, both the results width and without corrections are later accessible through the visualization panel (i.e., it does not force the user to repeat calculations). Whenever a correction is added to the mix for the quantification, the specifics are visually displayed in the info messages panel in -s-.

Before carrying out the quantification, the integration windows widths and positions have to be checked and configured. The tools for this task are shown in **Figure 5.20** (b). As always with **WhatEELS**, an automatic option is available -i-, which will extract the relevant parameters from the fitted model (windows widths related with the edges positions and onsets from the fitted ELNES centres and K-edges onset positions). The other two manual -i- options are: (1) to set the same integration width for every edge (fixed -j-), in which case the configuration is done by setting said width value as an energy range -m- and the onset energy values for each edge -l-. And (2), to set a unique (fluid -j-) energy width value for each edge, setting the initial energy loss value (onset -l-) and the final value (limit -n-) for the window.

The current energy windows selected for the integration can be displayed -h- over each single pixel spectrum -r- selected interactively on the dataset image -q- area. After any change -j-/-l-/-m-/-n- the update button has to be pressed. The initial values can be recovered at any time -p-.

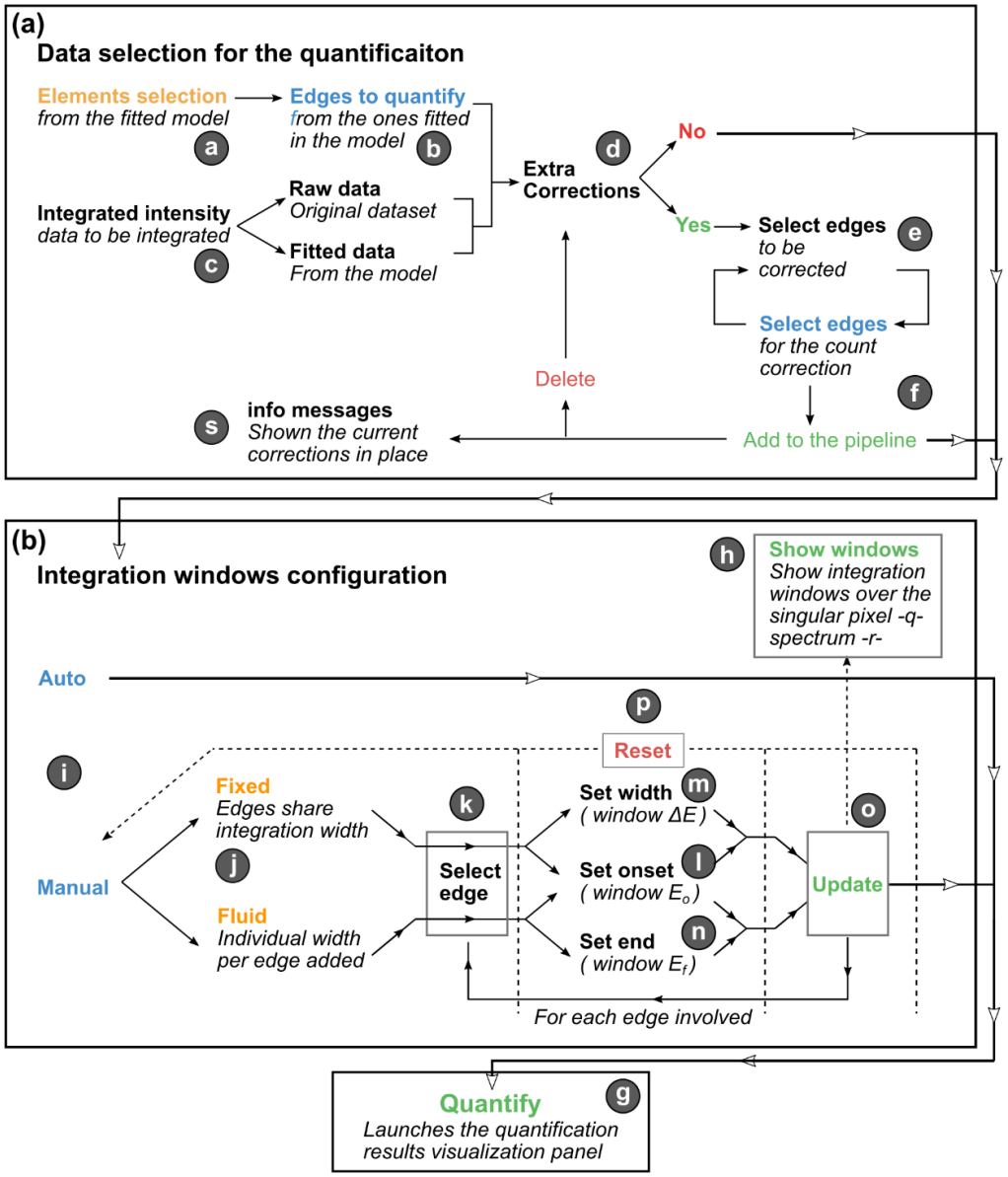


Figure 5.21. Workflow chart for the use of the elemental quantification module in **WhatEELS**. (a) Steps to configure the elements, edges, dataset selection and possible corrections introduced in the quantification. (b) Steps to configure the integration windows width for the quantification. All the lowercase letter tags are consistent with the ones in **Figure 5.20**, as it is also the division of tasks in subfigures (a) and (b).

The scheme for the workflow of tasks described up to this point for the quantification module can be found in **Figure 5.21**.

After completing the whole configuration process for the elemental quantification, and commanding the program to carry out the calculations ('Quantify' -g-), a new panel is launched by **WhatEELS** (automatically) to visualize the results. The tools available in this panel are shown in **Figure 5.22**.

Figure 5.22 (a) shows a summary of the options selected for the quantification in the previous panel -a-, and allows the user to set some visualization options. The base element (N_b in eq.(9), i.e., the denominator) has to be hand-picked (the app does not make any assumption about which one is the major element and which ones are the trace, dopants or secondary ones). If a count-correction step was included in the configuration of the quantification, the user is given the option to visualize the results with such correction active -c-, or to visualize the raw results (no corrections included).

To prompt the appearance of the ratios heatmap -j-, the user has to activate the *show ratios* button -d- (which considers the options -b-/-c- selected). After that, the theme configuration options for the heatmap -f- and the potential scatter plot -j- from a drawn line in -h- are unlocked. Whenever any of these parameters are changed, to translate said changes to the visual representations -j-/-i- the user has to actively press the refresh buttons -f/g-.

The visual representation of the dataset **Figure 5.22** (b) -h- allows the user to draw a line over any area of interest. The objective is to inspect in detail the elemental ratios of the pixels crossed by the drawn segment (tool -h_1-). The path of the line drawn is also shown in the ratios heatmap -j-, by colouring the pixels represented in the scatter plot -j-.

Notice that in the example of **Figure 5.22** a single ratio map is shown, as only two elements (Fe and Mn) where picked for the quantification of this dataset.

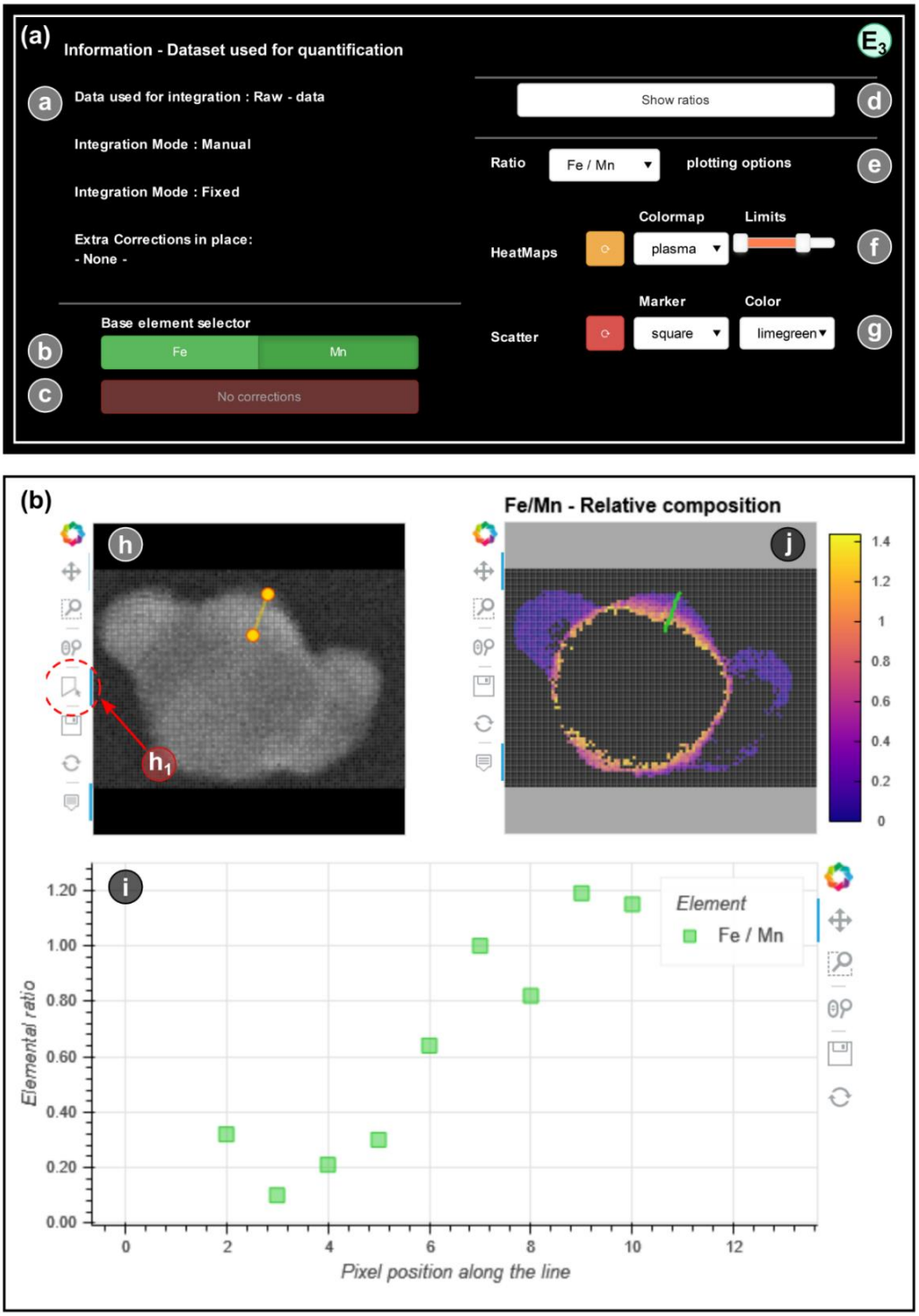


Figure 5.22. Quantification results visualization tools. (a) Configurator of the visualization parameters. (b) Interactive visualization areas in the panel.

Nevertheless, a number of heatmaps equal to the total number of elements minus 1 (the base element -b-) will be displayed for 3 or more elements. Whenever a line is drawn -h_1- in that case, a different marker will be assigned to each of the ratios calculated in the scatter plot -i-. Different theme options can also be assigned manually to each of the ratios computed. To change the heatmap and/or scatterplot marker characteristics for a specific ratio, this has to be indicated -e- before modifying the options.

5.3.4. Others.

-S- The saving panel.

A final extra module is included in **whatEELS**: the saving module **-S-**. Up to this point, the only saving options described in the modules were those in charge of saving the state of the application or the results obtained in order to later be read by **WhatEELS** again (e.g., saving the clustering runs **-B-** to be used in the fitting module **-D-**, or saving the multifit results **-D-** to be launched later again by the results module **-E-** without repeating the fitting process).

However, **WhatEELS** allows the user to launch the saving panel from any of the other modules in order to save both data and images outside the application. For instance, the data can be saved as .csv files to be read externally by some other software such as Origin. The saving panel **-S-** presents always the same set of tools, shown in **Figure 5.23**, and the original place from where the module is launched is indicated at the top (e.g., in this image, the departing point was the clustering module). **WhatEELS** will internally decide the available data (**Figure 5.23** (a)) and

images (**Figure 5.23** (b)) to be saved, depending on the original module from where the panel is launched.

The available data to be saved as datafiles (**Figure 5.23** (a)) is shown in a tabular panel -d-, where the coordinates names of the dataset and the shapes are indicated. These datasets can be saved -c- as NumPy bundles, .cvs, .xlsx

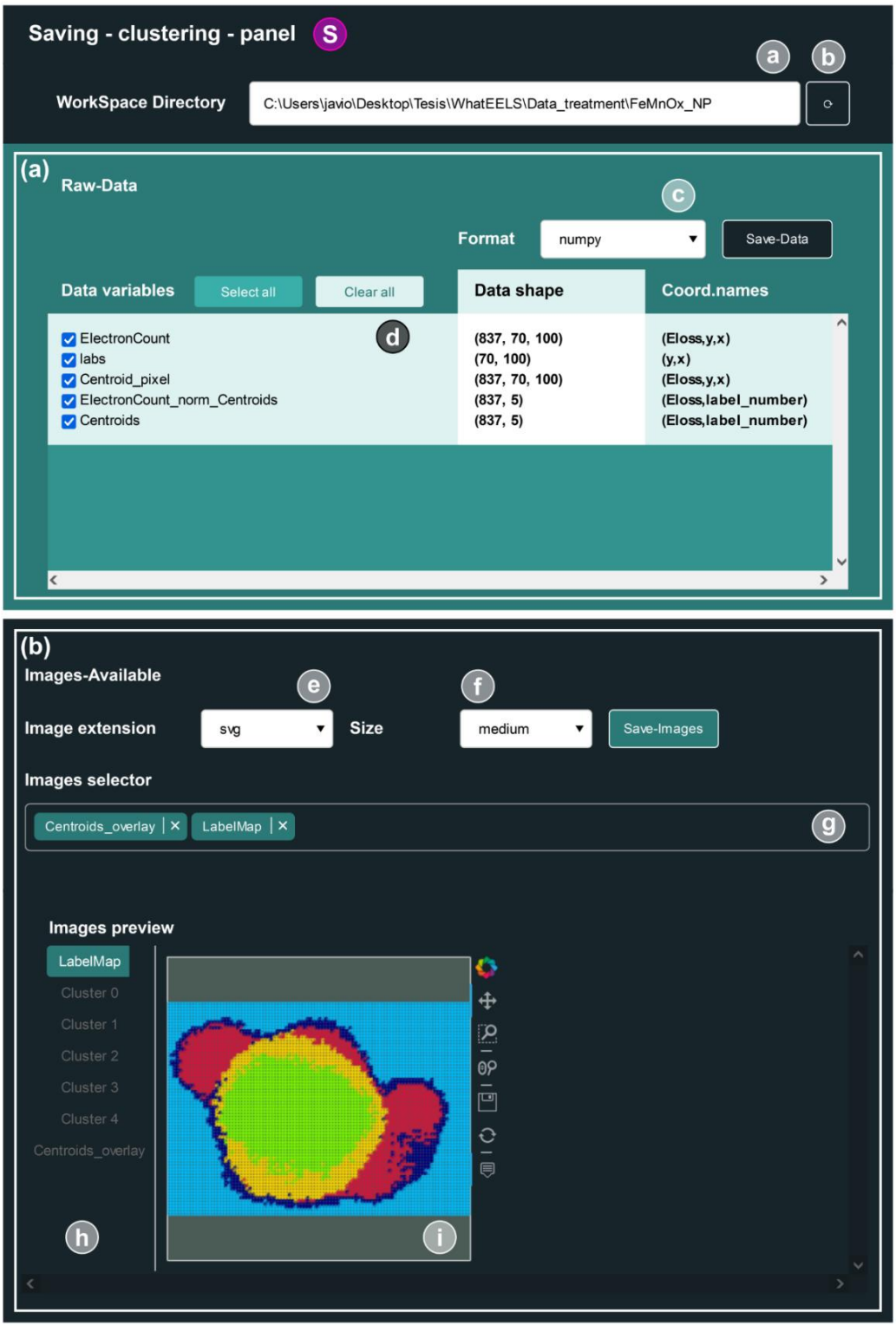


Figure 5.23 Tools from the saving panel of **WhatEELS**. It allows the user to save the images and data from many different modules and at many different states in the characterization process. (a) Tools to save the datasets. (b) Tools to save the images.

(excel) or netCDF4 (a compression format used by **WhatEELS** as well internally).

The images to be saved are shown in the set of tabs -h- (**Figure 5.23** (b)). From the displayed images -i-, the user has to select the desired ones -g-. Two saving -e- formats are offered, .svg and .png. For each one of them, a low, medium or large quality -f- image is offered as well.

All the images will be saved in a folder called '*/Workspace*' created (or read, if already exist) inside the directory shown in the text bar at the top -a-. This directory can be changed manually by the user by copy/pasting the desired one -a-. The original one can always be recovered -b-.

5.4. Iron-Manganese oxide nanoparticles.

This section presents the results of the analysis of the Iron-Manganese oxide core-shell NP[18] using every available tool in the core-fitting-components **-2-** block of **WhatEELS**. The clustering analysis module is also used to extract a suitable segmentation map for the later NLLS fitting.

5.4.1. Clustering Analysis.

The dataset used for the clustering analysis corresponds to the denoised version (via PCA) of Fe-Mn oxide NP EELS SI (with background). The dataset was pre-normalized (L_2) before running the clustering algorithm. As discussed in the previous chapter, using K-Means for non-normalized / non-scaled datasets in EELS data analysis may produce artefactual results, mainly dominated by the L_2 norm (scale) if the euclidean metric is set as the measuring criterium.

A number of 7 clusters was set as the objective for the K-Means algorithm. The fast iterative workflow allowed by **WhatEELS**, combined with the light

[18] This sample has already been presented in the previous chapter (***Machine Learning for EELS data analysis: Clustering and dimensionality reduction methods***). For an even more detailed analysis, the reader is referred to the original publication [20].

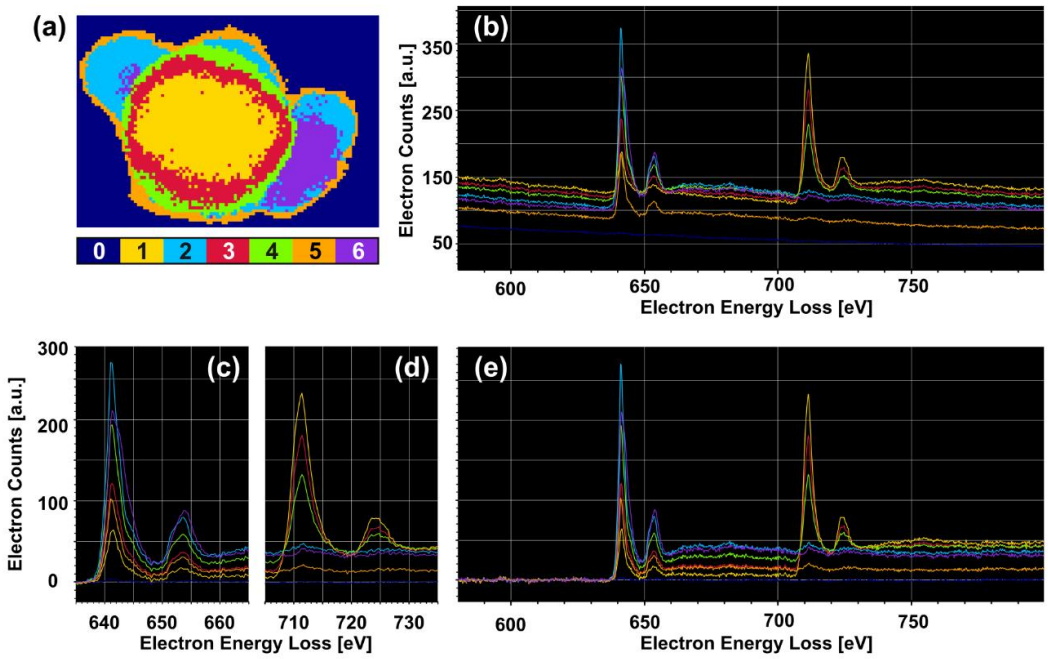


Figure 5.24. (a) Label-map for the clustering segmentation resolved by K-Means setting an objective number of clusters of 7. The dataset was L_2-normalized previously to the clustering analysis. (b) 7 Centroids resolved by the K-Means clustering algorithm. (c-d) Closeups on the Mn and Fe regions (respectively) of the average spectra for the background-removed EELS-SI signals shown in (e). (e) Average signal per cluster from the background-removed EELS-SI and for each of the cluster regions resolved in (a). The colour-scheme is shared by all 5 panels (a-e).

intensity of the calculations carried out by the sklearn implementation of K-Means, facilitated the tests of several combinations of pre-normalizations and objective number of clusters. A number of 7 clusters produced a nice configuration of labelled-regions over the sample area. This is the sole reason to select such a large number of clusters, at least compared to the 3 known materials present in the sample area (i.e., amorphous carbon, iron oxide and manganese oxide).

Figure 5.24 shows the results of the clustering segmentation **-B-**. **Figure 5.24** (a) shows the label-maps. **Figure 5.24** (b) shows the centroid signals. Notice that **WhatEELS** works preferable with background-removed EELS datasets. Thus, although a nice segmentation, it would in principle be unsuitable.

Luckily, **WhatEELS** also allows the later inclusion of segmentation maps from different datasets for the model construction **-D-**, as long as they belong to the same original EELS SI (i.e., the correct dimensionality, the same number of pixels and format, and the same number of energy channels in the spectra). Thus, using this cluster arrangement later was possible. **WhatEELS** basically takes care of obtaining the new 'centroids' (strictly speaking they are not actual centroids) by computing the average signal per cluster for the new (and compatible) dataset. The results in this example are shown in **Figure 5.24** (e). The suitability of using this clustering segmentation for the background-removed dataset is already visible by analysing the closeups for the Mn and Fe areas of the spectra (**Figure 5.24** (c-d), respectively). Notice, for example, the clear chemical shift of the MnL_{32} white lines between cluster 2 (cyan) and cluster 6 (purple). These WLs in cluster 6 (purple) also present a clear broadening of the curves, hinting to the possible convolution of 2 different peak-like structures in a small energy range. This information alone is already hinting a probable dual phase coexistence in the shell region corresponding to cluster 6, but further experiments were required to test this hypothesis.

5.4.2. NLLS model.

The model created for the NLLS fit of the Fe-Mn oxide core-shell NP consisted of: (1) a spline function to model the excitation to the continuum of states for each one of the edges (FeL_{32} and MnL_{32}), extracted from the GOS surfaces modified by the experimental parameters. And (2), two gaussian curves at the onset of each of the included edges, representing the pairs of favoured discrete excitations to other inner shell states (i.e., the WLs).

Before **WhatEELS** can carry out the multifit for the whole EELS SI, it is convenient to adjust the initial parameters for each of the pixels involved. To do so, an initial assessment of the model for the reference spectra of the EELS SI was required. The clustering segmentation shown in **Figure 5.24** was included here, so these reference spectra corresponded to the average EELS signals per cluster.

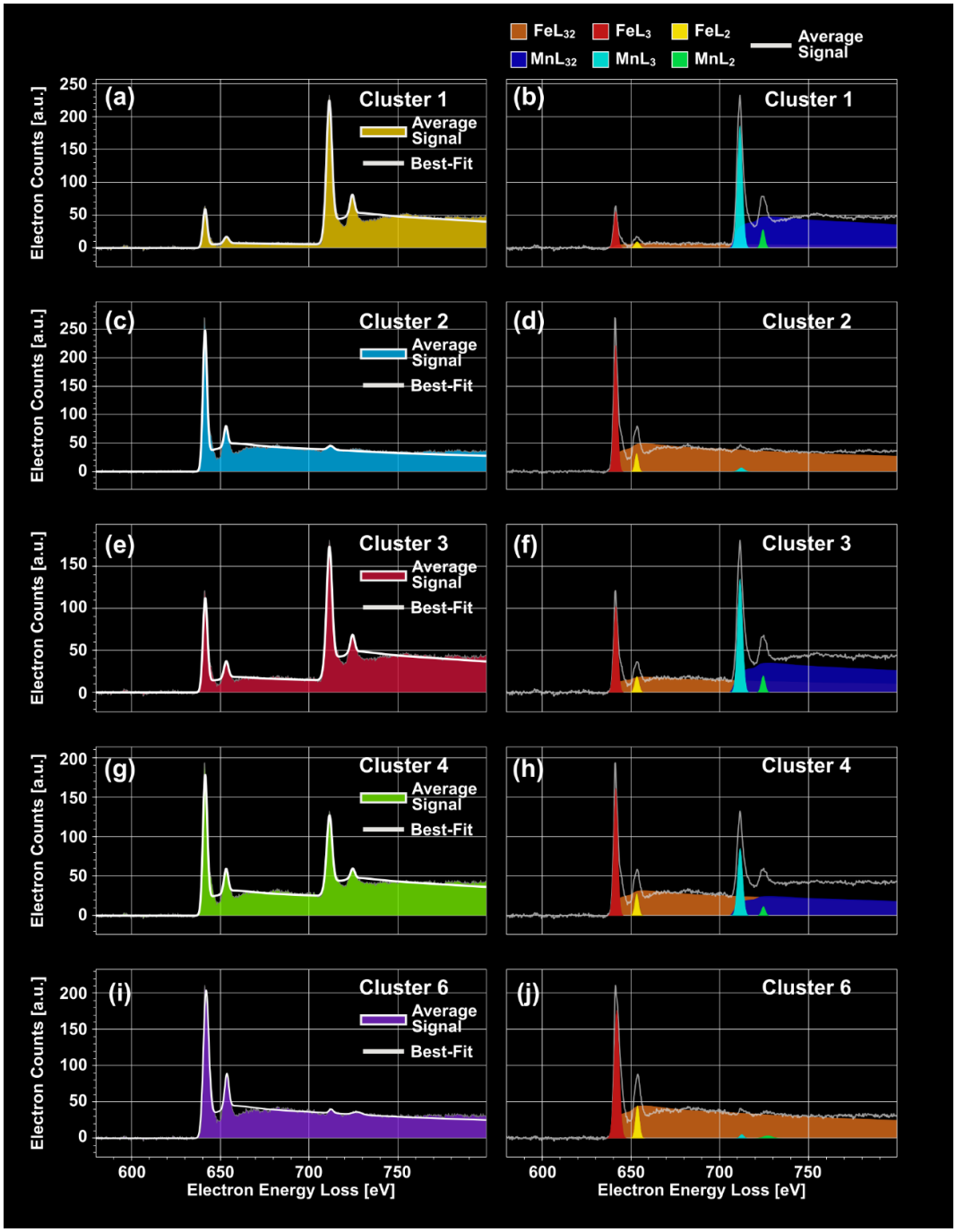


Figure 5.25. (a)(c)(e)(g)(i) Overlays of best-fit curves and raw data for the reference EELS signal of clusters 1,2,3,4 and 6. (b)(d)(f)(h)(j) Overlays of the raw data and the individual components for the best fit for each of the clusters included in the NLLS fitting (1,2,3,4 and 6). The colour scheme is shared by all the individual components representations (b)(d)(f)(h) and (j).

The results for the best-fit curves and the individual components of the NLLS fitting are shown in **Figure 5.25**. Notice that only clusters 1,2,3,5 and 6 are included. Those are the cluster areas selected for the multifit. Cluster 0 corresponds to the background noise, so its spectral interest is minimum. Cluster 4 (**Figure 5.24** (a-b)) was interesting to a certain degree. Nevertheless, the signal to noise ratio in this outermost region of the shell is severely degraded. Although the reference average signal was easily fitted, the single-pixel NLLS fittings oftentimes failed to converge or produced suboptimal results.

Also, clusters 2 **Figure 5.25** (c-d) and 6 **Figure 5.25** (i-j) (NP shell) were fitted by muting the excitation to the continuum function of the FeL_{32} (setting the amplitude to 0). The qualitative analysis of the clustering results showed that these regions of the shell were almost entirely composed by manganese oxide. This helped to speed up the convergence times, at the risk of losing possible contributions of the Fe_{32} edge to the tail of the spectra.

5.4.3. Reduced chi squared.

The first step in the results analysis is to check the reduced chi squared values for each of the datapoints (pixels) in the EELS SI. The results of this analysis are shown in **Figure 5.26**.

Notice how the χ_v^2 mapping show values over 1 for the whole area analysed (**Figure 5.26** (a)). As already discussed, the real objective in a NLLS fitting for an EELS dataset is to obtain a homogeneous distribution of values for the whole dataset, and not to get an actual successful χ_v^2 test (lower than 1).

In the figure, some areas from the cluster 2 (see **Figure 5.26** (b)) present higher values than the average (specially towards the boundary core-shell in the NP). The spectrum best fit model and residual curves of some pixels (marked in **Figure 5.26** (a-b)) are shown in **Figure 5.26** (c-e). Pixel 1 (**Figure 5.26** (c)) shows a very tight fit of the configured model for cluster 3 (red),

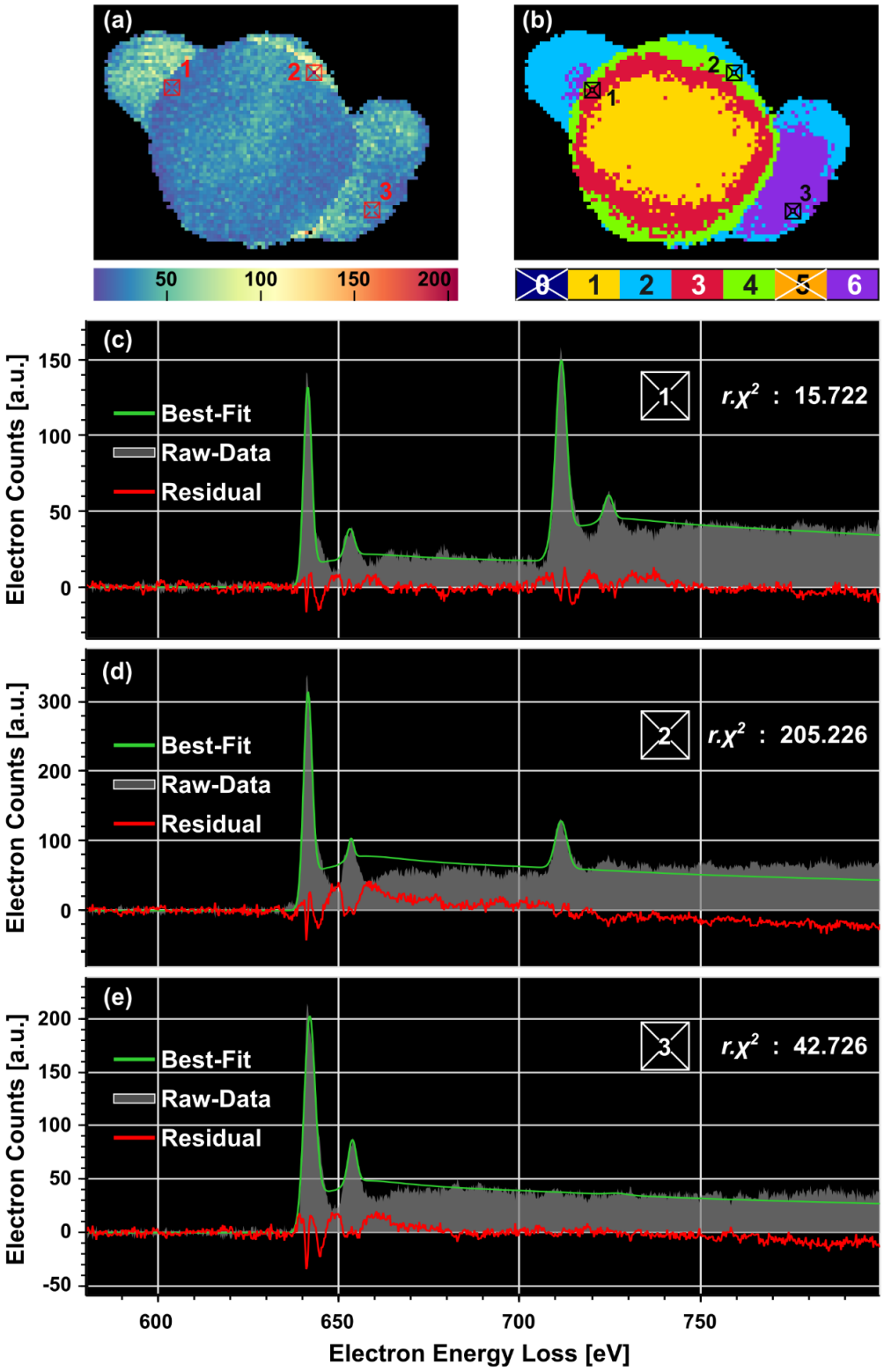


Figure 5.26. (a) Reduced chi squared mapping for the area fitted by **WhatEELS**. (b) Clustering label map for reference (**Figure 5.24**). (c-d) Overlay of the original raw data, best fit and residual curves for the pixels highlighted in (a) and (b).

hence the low χ_v^2 retrieved. Both pixels 2 (**Figure 5.26** (d)) and 3 (**Figure 5.26** (e)) (from the shell areas, clusters 2 and 6 respectively) present worst coefficient numbers. The origin of this degradation of the coefficients lies on the separation of best-fit curve and raw-data towards the end (higher energy loss channels) of the spectra.

This could be driven by an underperforming deconvolution of the low-loss contributions, or even a defective background removal pre-step in these shell areas with an overall lower signal-to-noise ratio. Another possible cause for the problem is the absence of the FeL_{32} excitation to the continuum of states. It was removed (muted) from the shell areas to increase the time performance, but the appearance of the L_3 peak-like structure in pixel 2 (the one with the higher χ_v^2 throughout the whole SI) hints to the probable mixture of iron oxide and manganese oxide in some unaccounted parts of the shell (at the very least, in some pixels of the clusters separated as shells).

In any case, the central positions for the peak-like components are, qualitatively speaking, well resolved. The analysis of centres should not suffer from this uneven χ_v^2 test. The WL ratio results, however, should be taken with a grain of salt, as the models fitted on the shell clusters may tend to overestimate the excitation to the continuum of states for the MnL_{32} edge to compensate for the absence of the FeL_{32} edge (see **Figure 5.26** (d)).

5.4.4. Centres analysis.

Figure 5.27 shows the results for the centres analysis of the clusters belonging to the NP shell (clusters 2, cyan, and 6, purple, **Figure 5.27** (b)). The interest in these areas resides on the fact that a previous analysis of this dataset[20] showed localized changes in the oxidation state of the Mn and a possible coexistence of the MnO and Mn_3O_4 phases.

Figure 5.27 (a) contains the centres mapping for the MnL_3 component. The qualitative visual inspection reveals the existence of distinctive areas within the shell with a displacement of the centres towards higher energy loss values. This behaviour is repeated in the MnL_2 centre mapping **Figure 5.27** (b).

The graphs for the centre positions values along the lines drawn in **Figure 5.27** (a-b) are shown in **Figure 5.27** (d) for the MnL_2 WL and **Figure 5.27** (e) for the MnL_3 WL. A displacement towards higher energy loss values of up to 1.2 eV in the inner part of the shell (closer to the iron oxide core) is visible in both components. This is in good agreement with the known coexistence of MnO and Mn_3O_4 in those concrete regions[20].

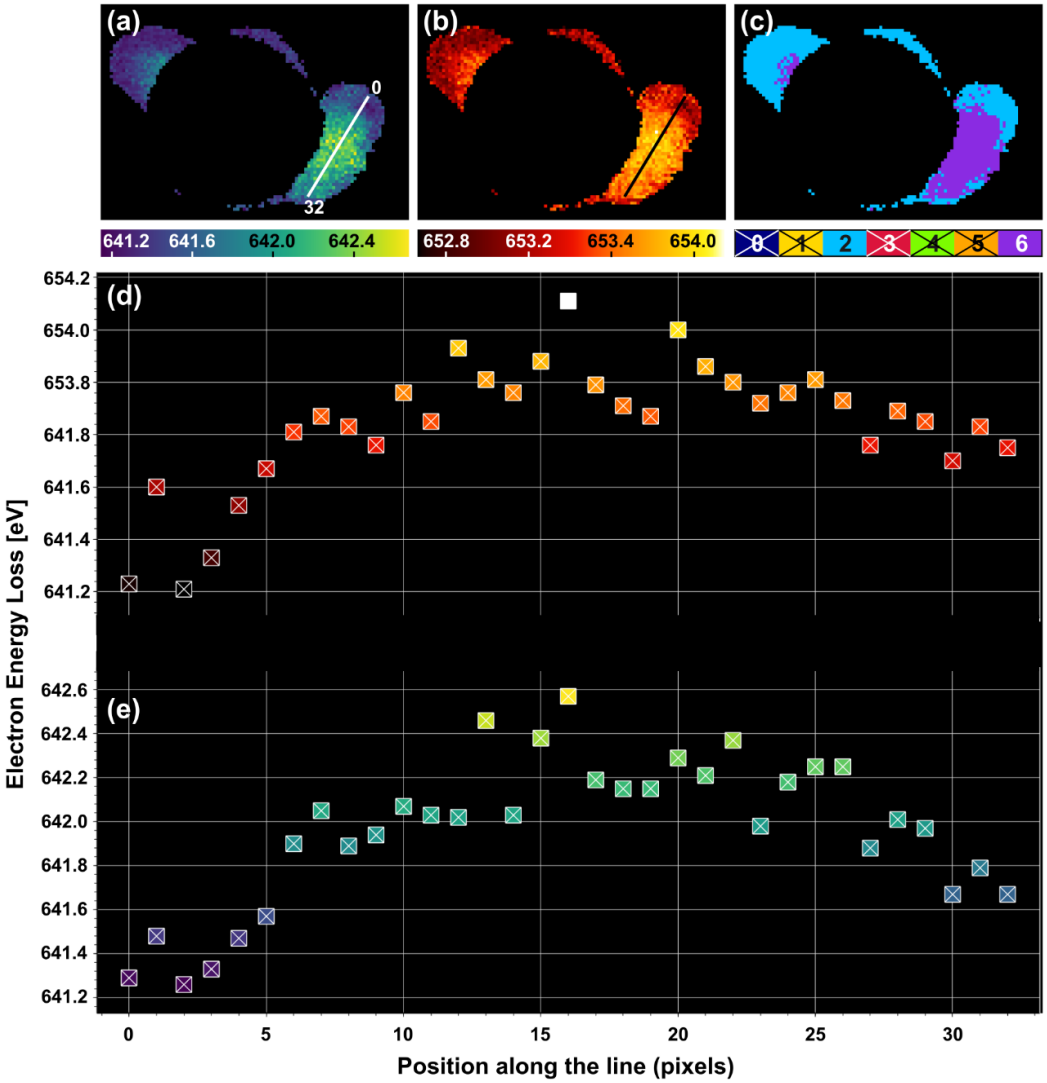


Figure 5.27. Centres analysis for the shell of the Iron-Manganese oxide core-shell NP (shell mainly composed by Mn oxide). (a) MnL_3 component centres for the cluster areas 2 and 6. (b) MnL_2 component centres for the cluster areas 2 and 6. (c) Clusters analysed by the centres analysis tool. (d-e) Scatter plots for the energy loss positions of the centres along the lines highlighted in black (b) and white (a), and for the MnL_2 and MnL_3 components respectively.

5.4.5. White lines analysis.

To cross-check the results of the centres analysis, a WL ratio analysis was carried out for the same areas of the NP shell (clusters 2 and 6, see **Figure 5.28** (b)).

The resulting mapping for the Mn-L_2/Mn-L_3 ratios is shown in **Figure 5.28** (a). Here, the same regions of the shell (closer to the iron core) that those in the centre analysis mappings present a distinctive structure of values (i.e., higher values for the WL ratios). This is in good agreement with the results in the literature[20,23], that state that decreasing the Mn-L_3 / Mn-L_2 (i.e., the inverse of the ratio shown here) is equivalent to a displacement towards higher oxidation numbers in the Mn[19].

The actual value of the ratio is not that relevant, because each of the references in the literature tends to use different values for the integration window (or no integration window at all). Also, the removal of the continuum excitation function is different in each case (which functions are actually used to model this part of the spectra, and where they are exactly fitted). Hence, the important characteristic is the overall behaviour: the displacement towards higher ratios (**Figure 5.28** (a)) in the same areas where it also took place a centre value displacement towards higher energy loss values (**Figure 5.27** (a-b)).

Finally, **Figure 5.28** (d) shows the ratio scatter plot for the pixel highlighted in **Figure 5.28** (a) and for different integration windows used for the calculations. The integration widths are given as a multiplier of the fwhm[20]

[19] MnO presents a Mn^{2+} number, whereas Mn_3O_4 presents a $Mn^{2.66+}$ average number.
[20] The fwhm for a gaussian curve can be recovered from the sigma parameter as

$$fwhm = 2\sqrt{2\ln(2)}\sigma$$

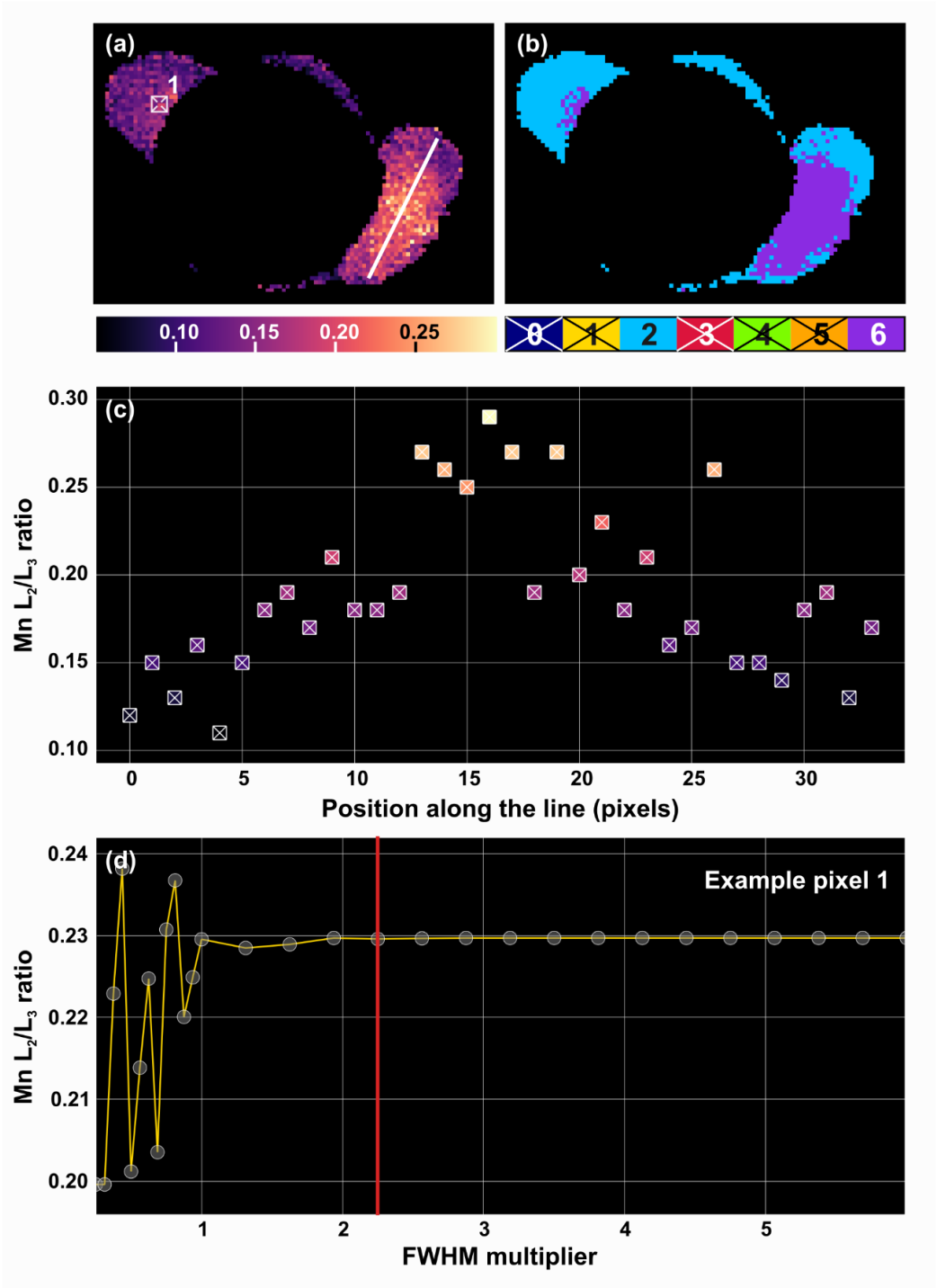


Figure 5.28. (a) Mn-L_2 / Mn-L_3 WL ratio for the shell of the Iron-Manganese oxide NP. (b) Cluster areas analysed by the WL-ratio module. (c) Scatter plot of the WL ratio along the line highlighted in (a) in white. (d) Scatter plot for the ratios measured for pixel 1 (highlighted in (a)) for different integration widths. The x-axis corresponds to the multiplier used by **WhatEELS** to calculate the integration window from the fwhm values of the peak-like components in each pixel.

base value of the fitted component for the MnL_{32} WLs. From 2 onwards, the ratio is stabilized because the signals integrated corresponded to the fitted gaussian models, and that multiplier value would be passed the point at which the whole gaussian curve is considered. The ratios map showcased in **Figure 5.28** (a) corresponds to the multiplier marked with the vertical red line in **Figure 5.28** (d).

5.4.6. Quantification.

The last module used in for this sample was the elemental quantification one. The actual composition of the different regions of the spectra was already known form the clustering analysis results. Nonetheless, the quantification carried out using this module is included to showcase the whole range of tools available in **WhatEELS**.

Recall that in the models proposed for the shell clusters, the excitation to the continuum of the iron oxide edge was muted. As such, and according to the equation for the Egerton's method described in section 2 eq.(9), the ratio Mn/Fe would not be properly evaluated in those regions. Hence, the elemental composition was analysed for the NP core only (that includes both elements and the whole range of components in the NLLS fitting).

The cation Mn/Fe ratio mapping **Figure 5.29** (a) shows an increment of the Mn proportion towards the outer parts of the core. This is also shown in the ratio scatter plot (see **Figure 5.29** (c)) for the pixels along the path highlighted in red in **Figure 5.29** (a).

This result is in good agreement with the expected behaviour of a core-shell structure without mixture of different elements in any of the presented phases, as the Mn/Fe ratio is stable in a base value in the thicker areas of the core (this EELS signal would arise from the superimposed shell of Mn oxide over the Fe oxide core) and quickly grows towards the outer parts of the core (where the shell contents become dominant)

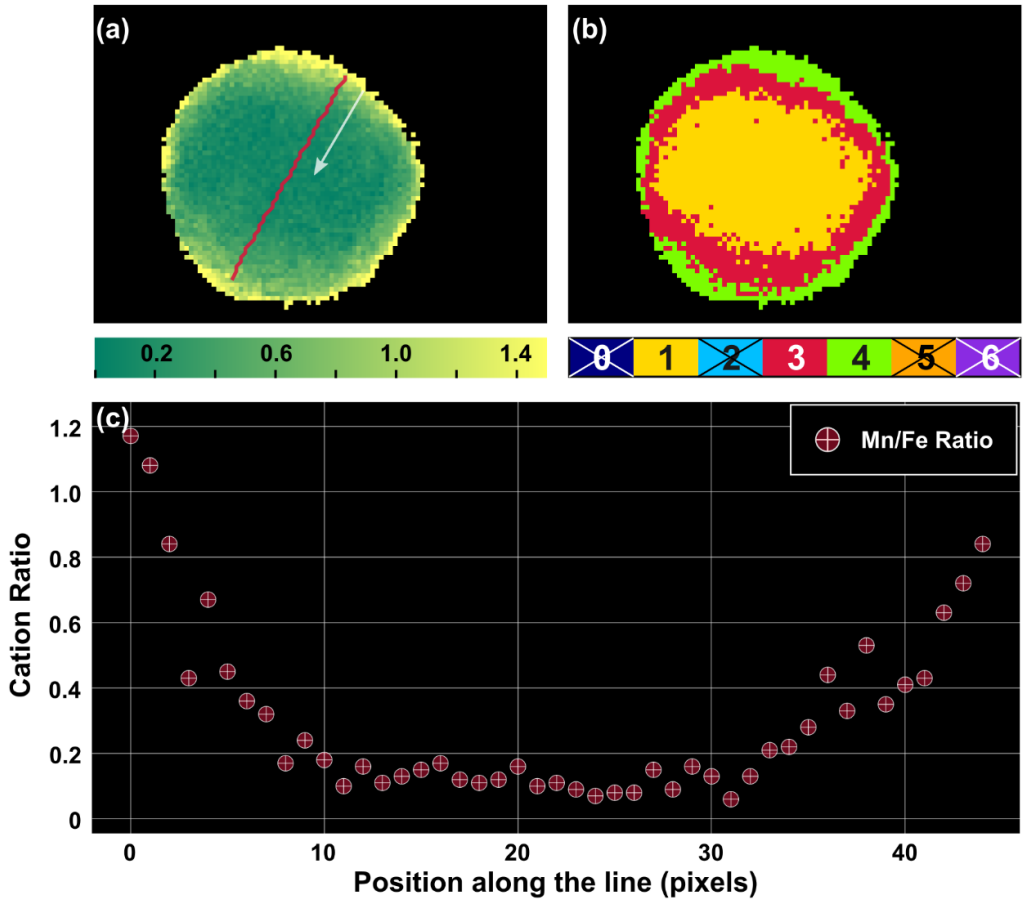


Figure 5.29. (a) Mn / Fe elemental composition mapping for core area of the ion-oxide / manganese oxide NP. (b) Clustering labels mapping for the clusters number 1,3 and 4. (c) Cation ratio resolved by Egerton's method for the pixels in the path highlighted in red in (a).

5.5. The panel structure.

The following pages contain the actual structure of the panels in **WhatEELS**. The tools for each of the modules are labelled identically to those highlighted in section 2 (software architecture). The caption for each figure will indicate which other figures are related with the ones presented here.

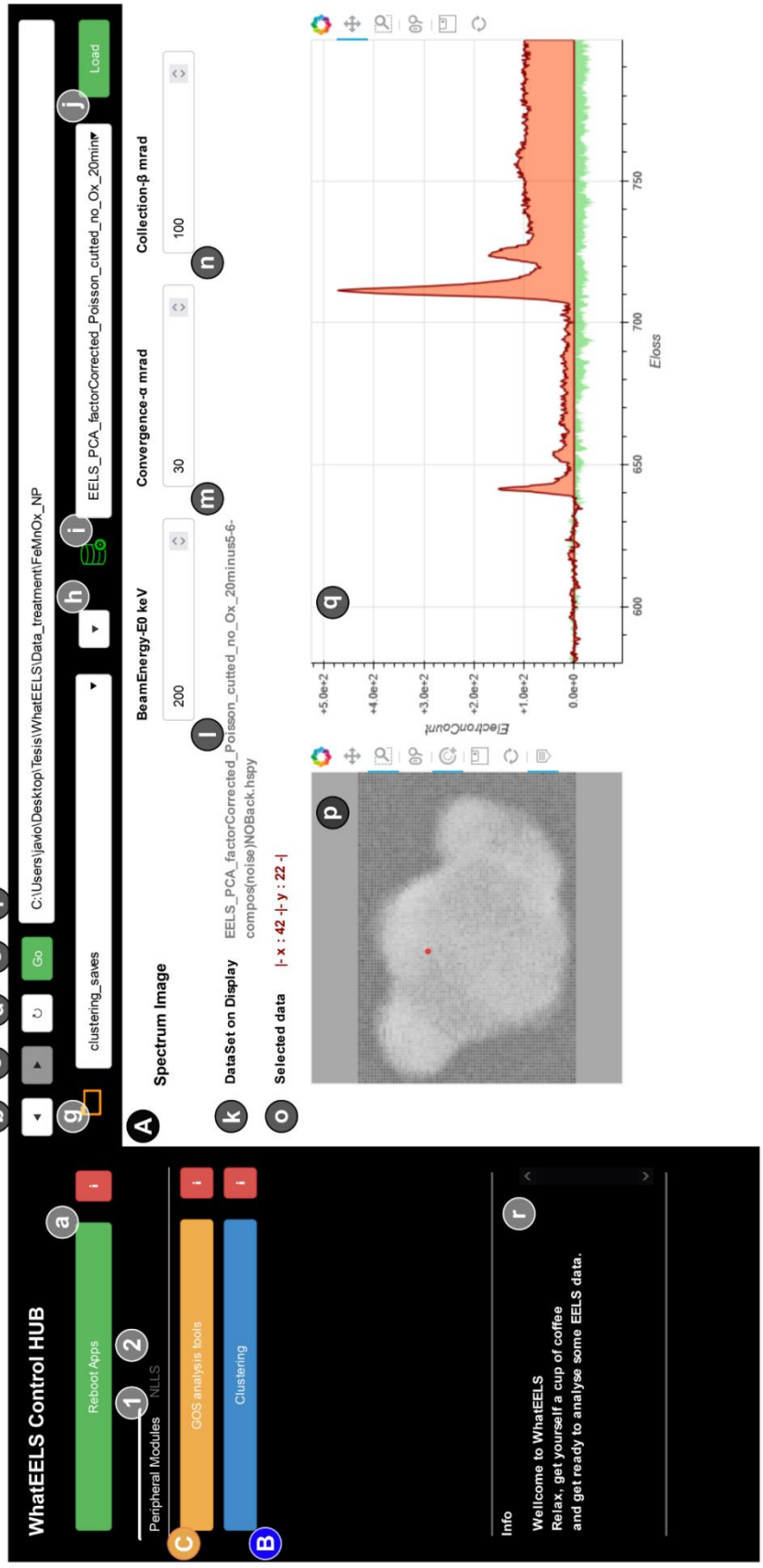


Figure 5.30. WhatEELS Loading panel and main HUB controls. It is equivalent to **Figure 5.6** and **Figure 5.7**.

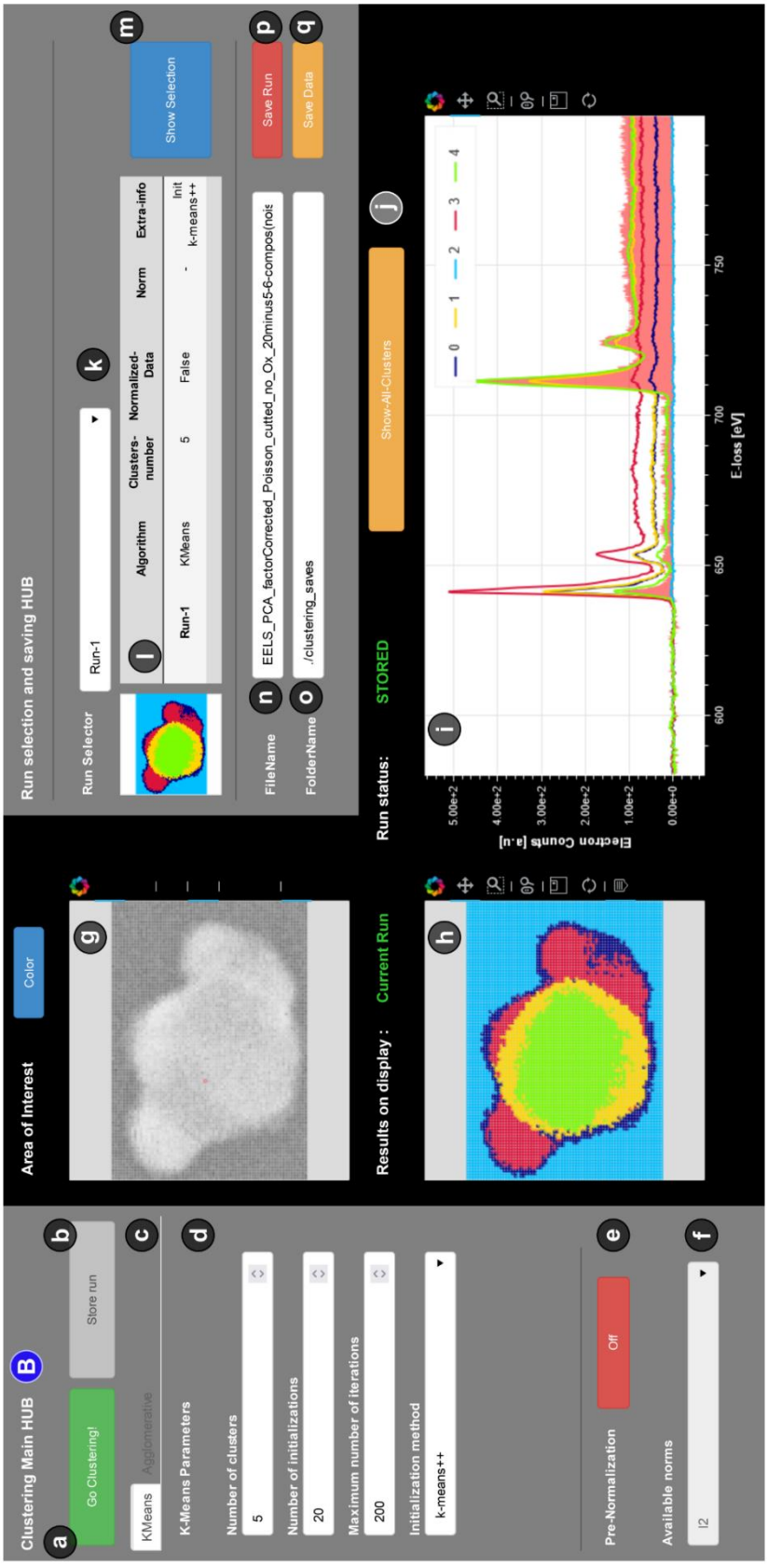


Figure 5.31. WhatEELS clustering panel. The corresponding figure showcasing the individual tools included is **Figure 5.8**.

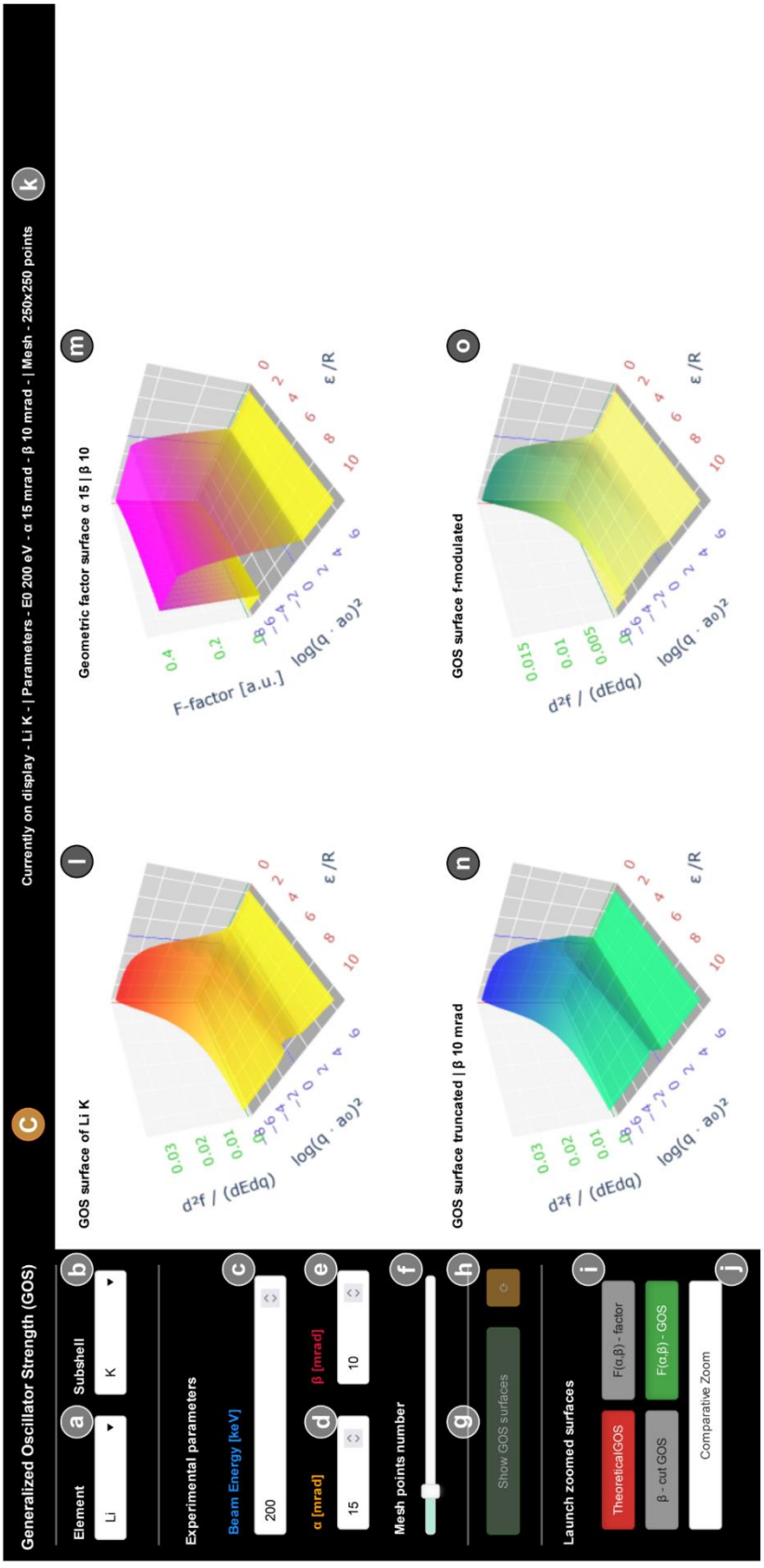


Figure 5.32. GOS surface analyser module from **WhatEELS**. The corresponding figure showcasing the different functionalities included is **Figure 5.9**.

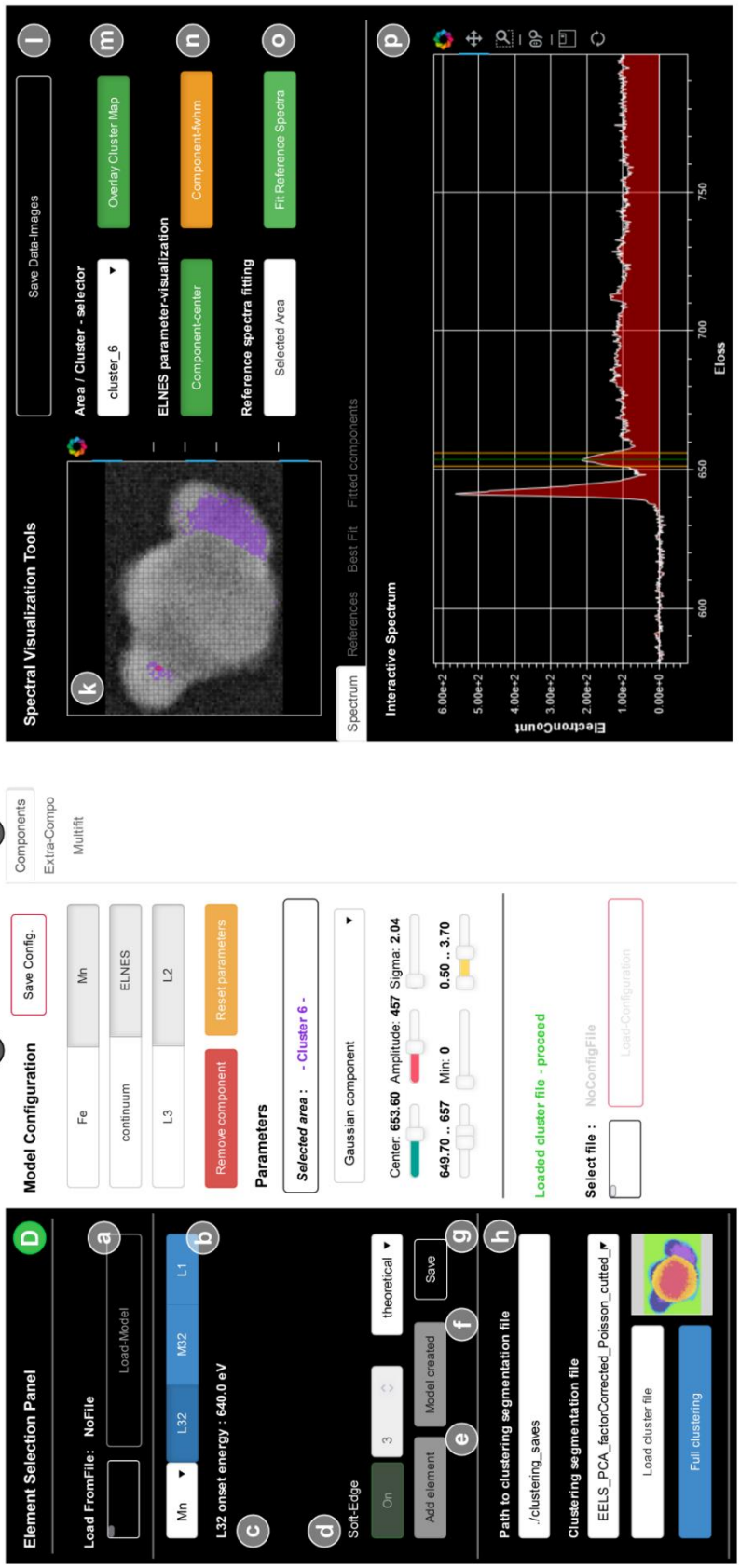


Figure 5.33. Model creator and multifit panel for **WhatEELS**. The corresponding figures showcasing the different functionalities included are **Figure 5.10**, **Figure 5.11** and **Figure 5.12**.

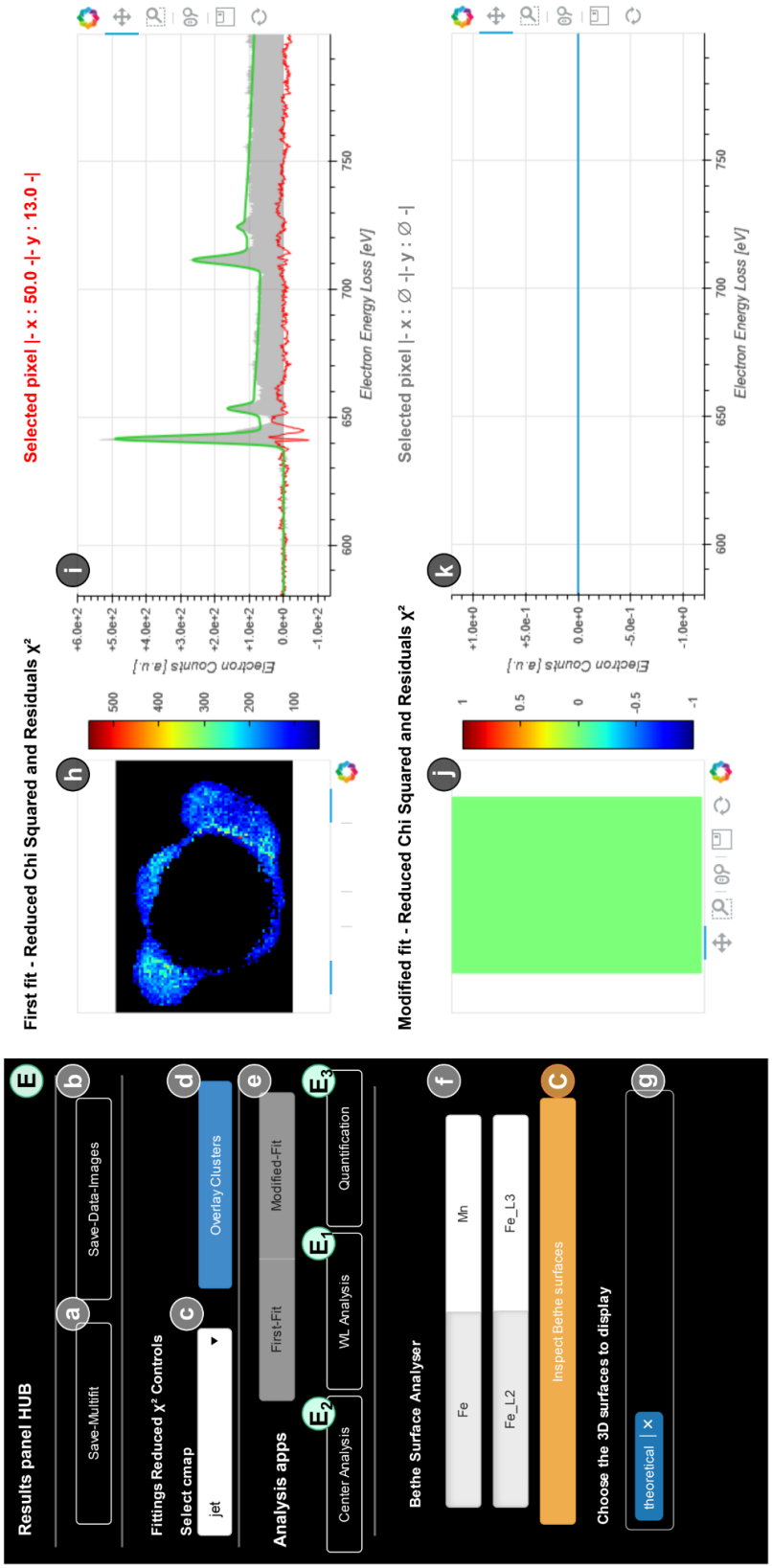


Figure 5.34. Results main HUB. The corresponding figure describing the use of the tools included in this panel is **Figure 5.13**.

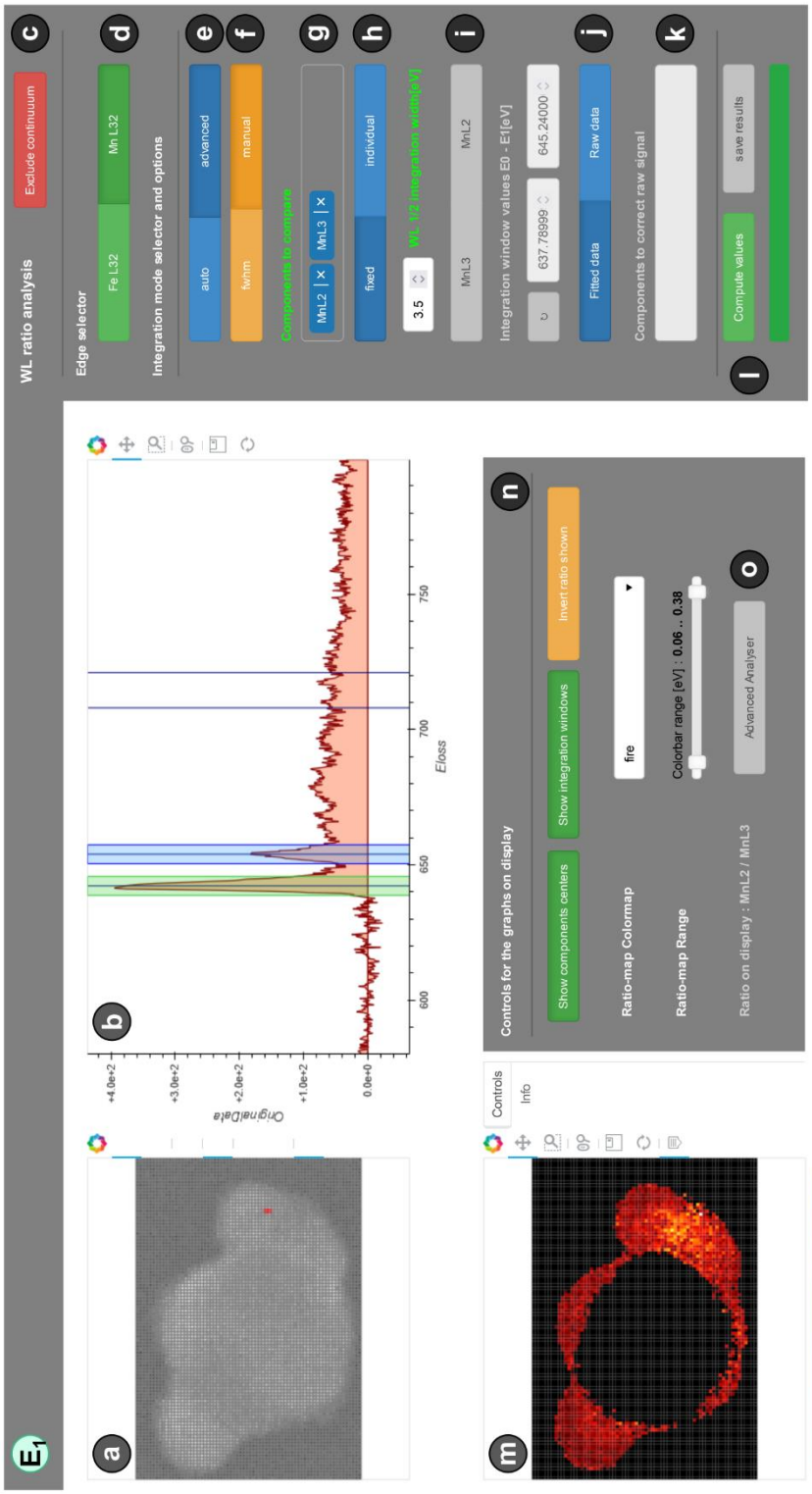


Figure 5.35. White line ratio analysis panel. The module functionalities were described in **Figure 5.14** and **Figure 5.15**.

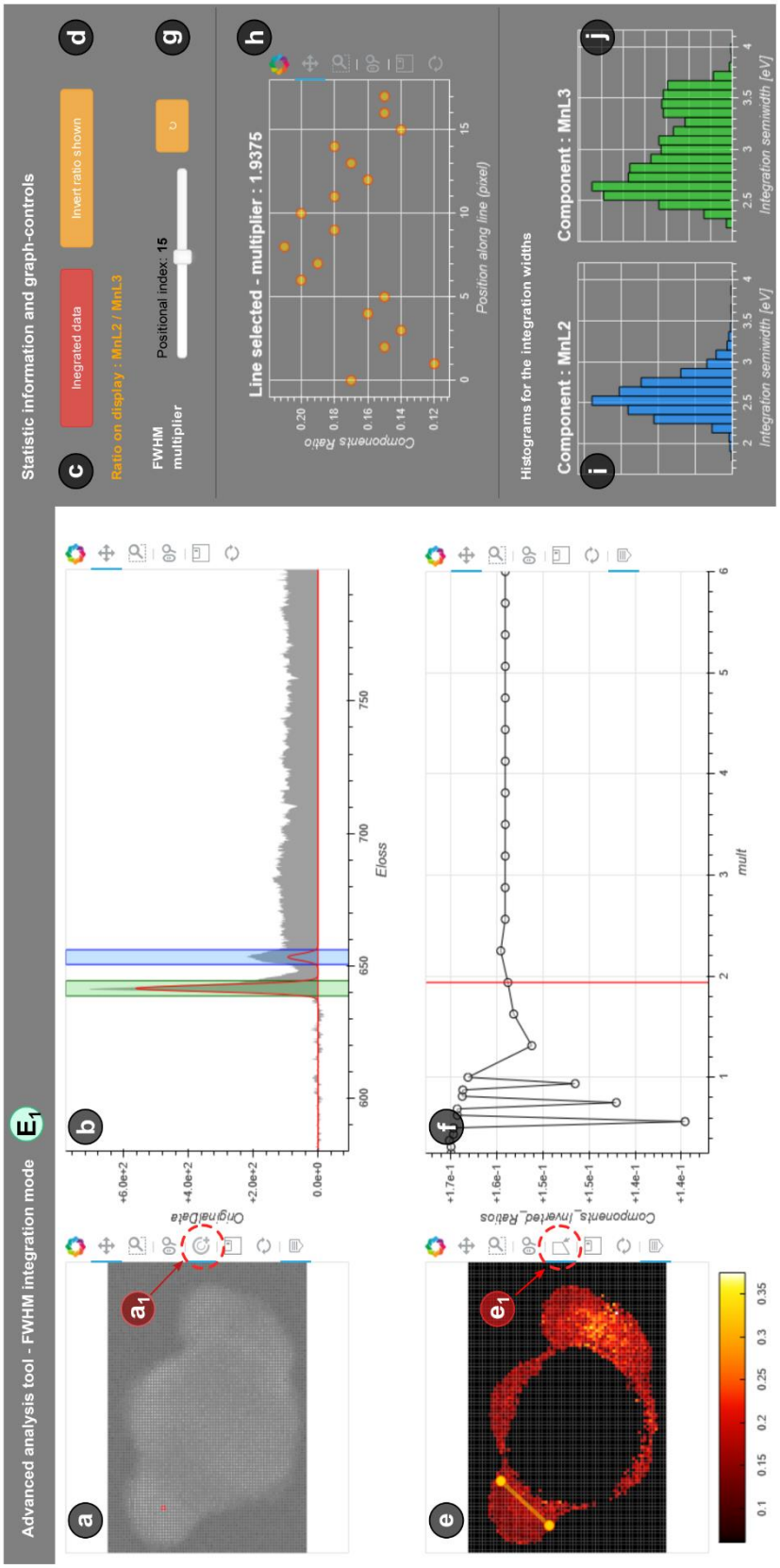


Figure 5.36. Advanced WL ratio analyser. The specifics about the use of this panel are showcased in **Figure 5.16.**

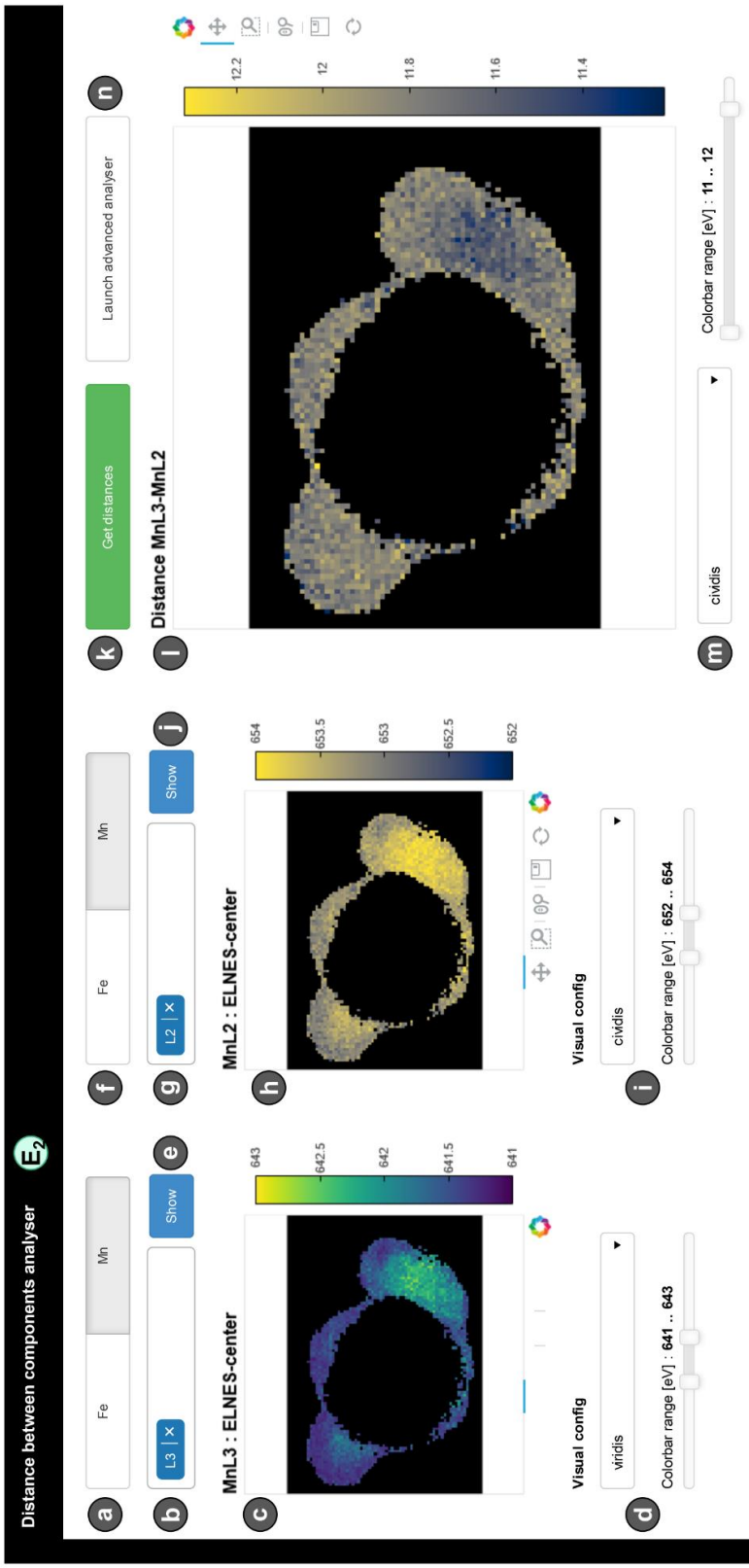


Figure 5.37. WhatEELS centres analysis tool. The specific functionalities form this panel were showcased in **Figure 5.17** and **Figure 5.18**.

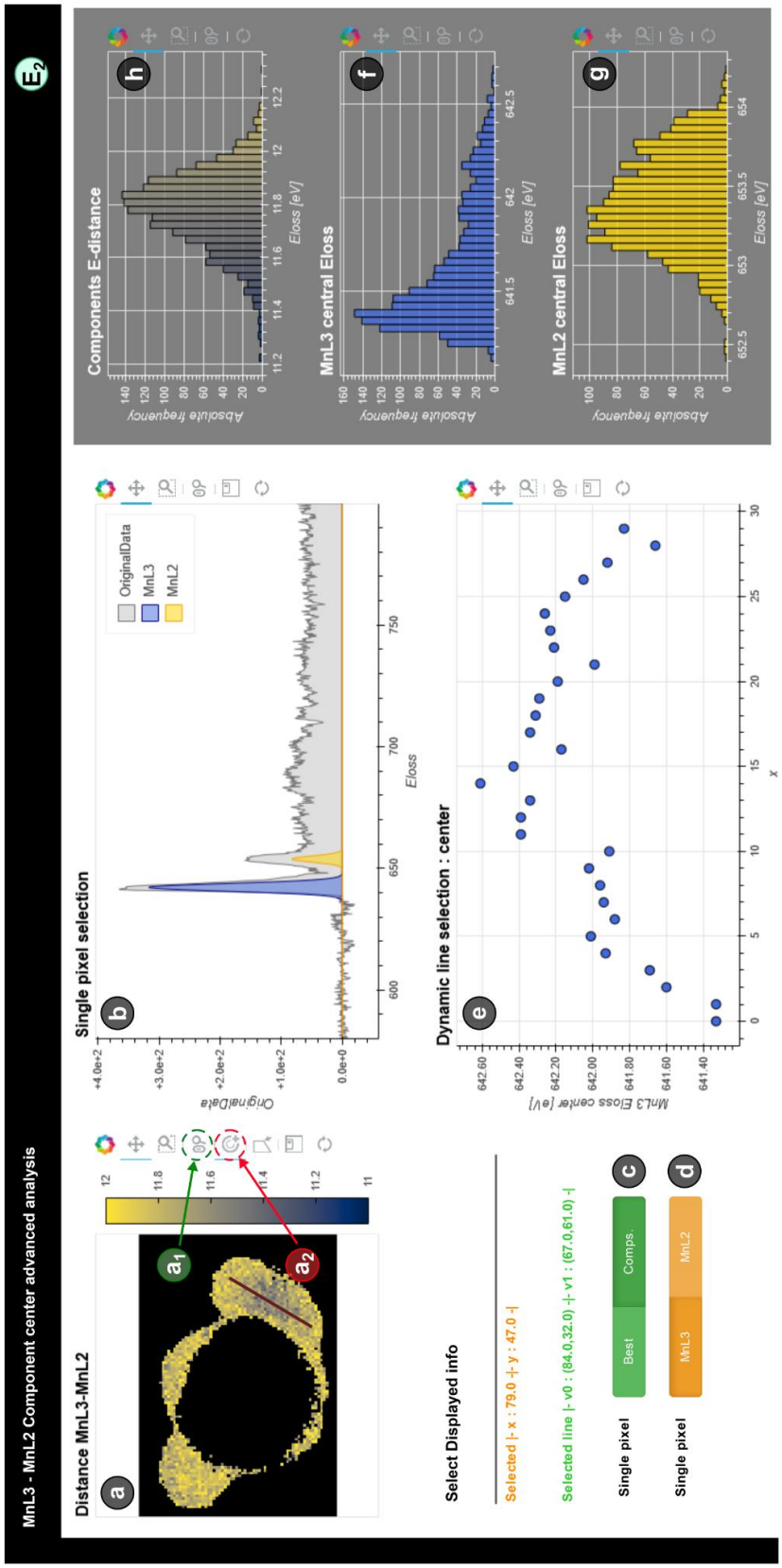


Figure 5.38. Advanced centres analysis tools from **WhatEELS**. The functionalities are showcased as well in **Figure 5.19**.

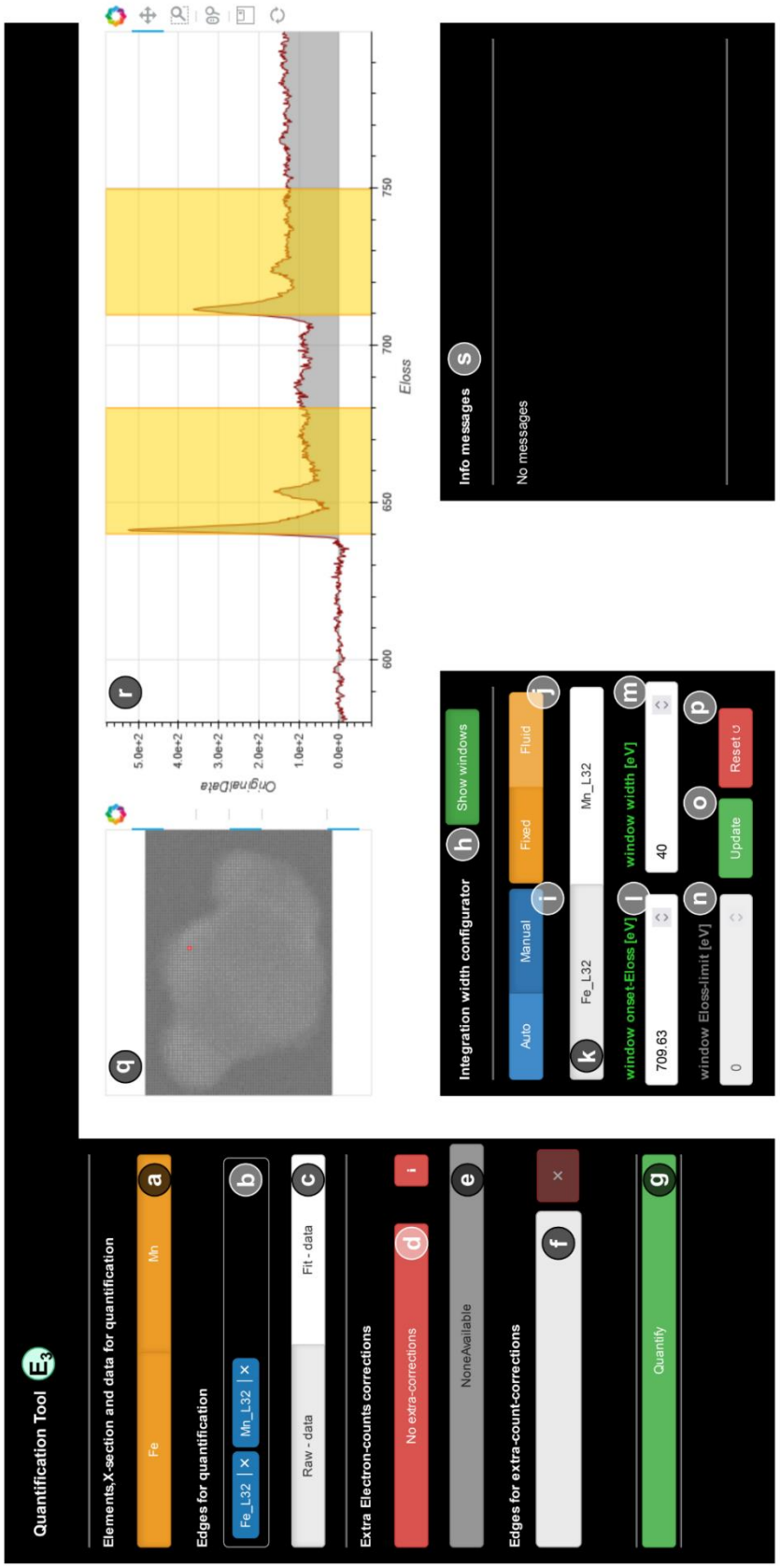


Figure 5.39. Quantification configuration panel. The specific functionalities in this panel were showcased in **Figure 5.20**.

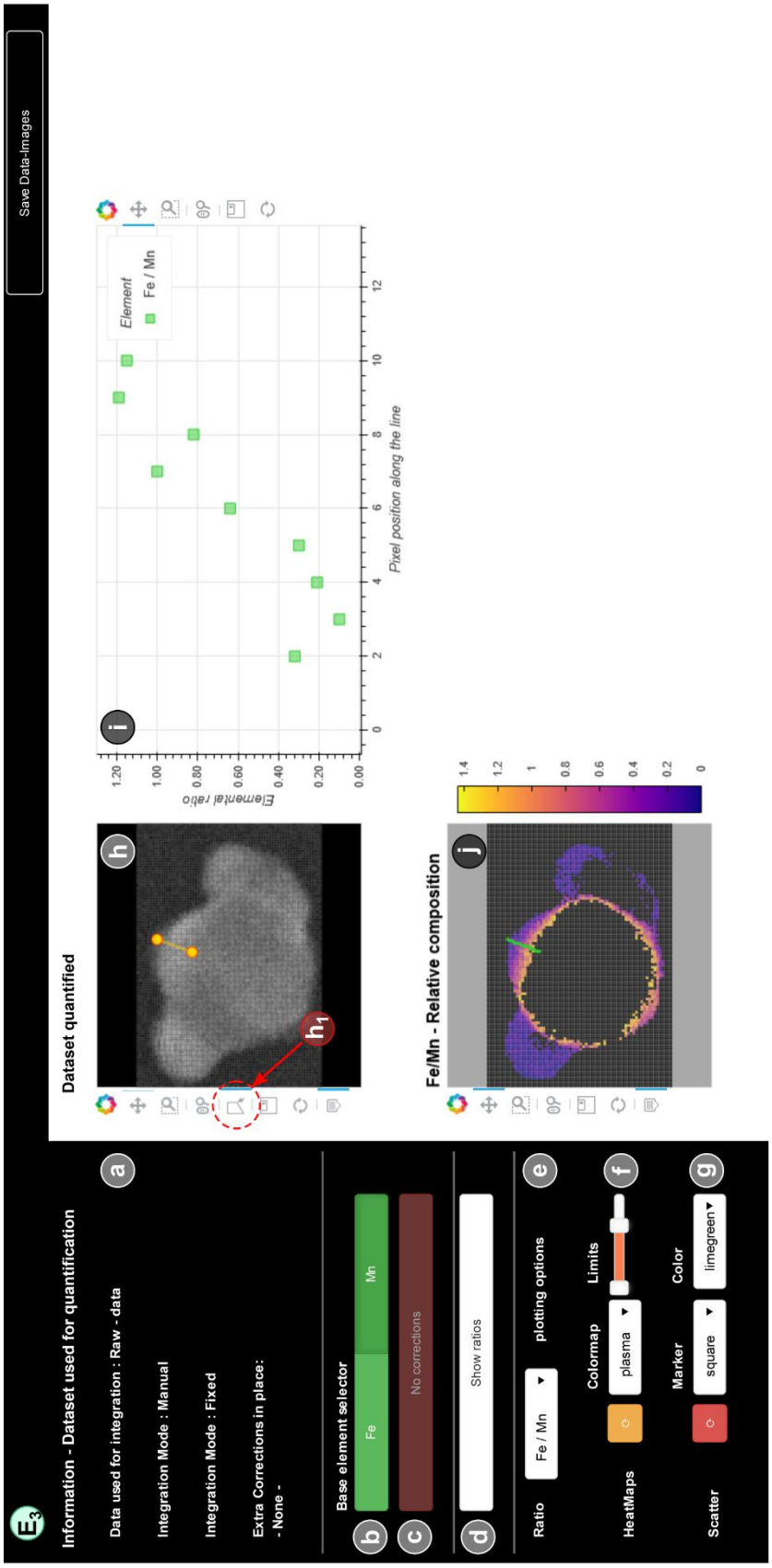


Figure 5.40. Quantification results visualization panel. The specific tools to effectively use this panel were showcased in **Figure 5.22**

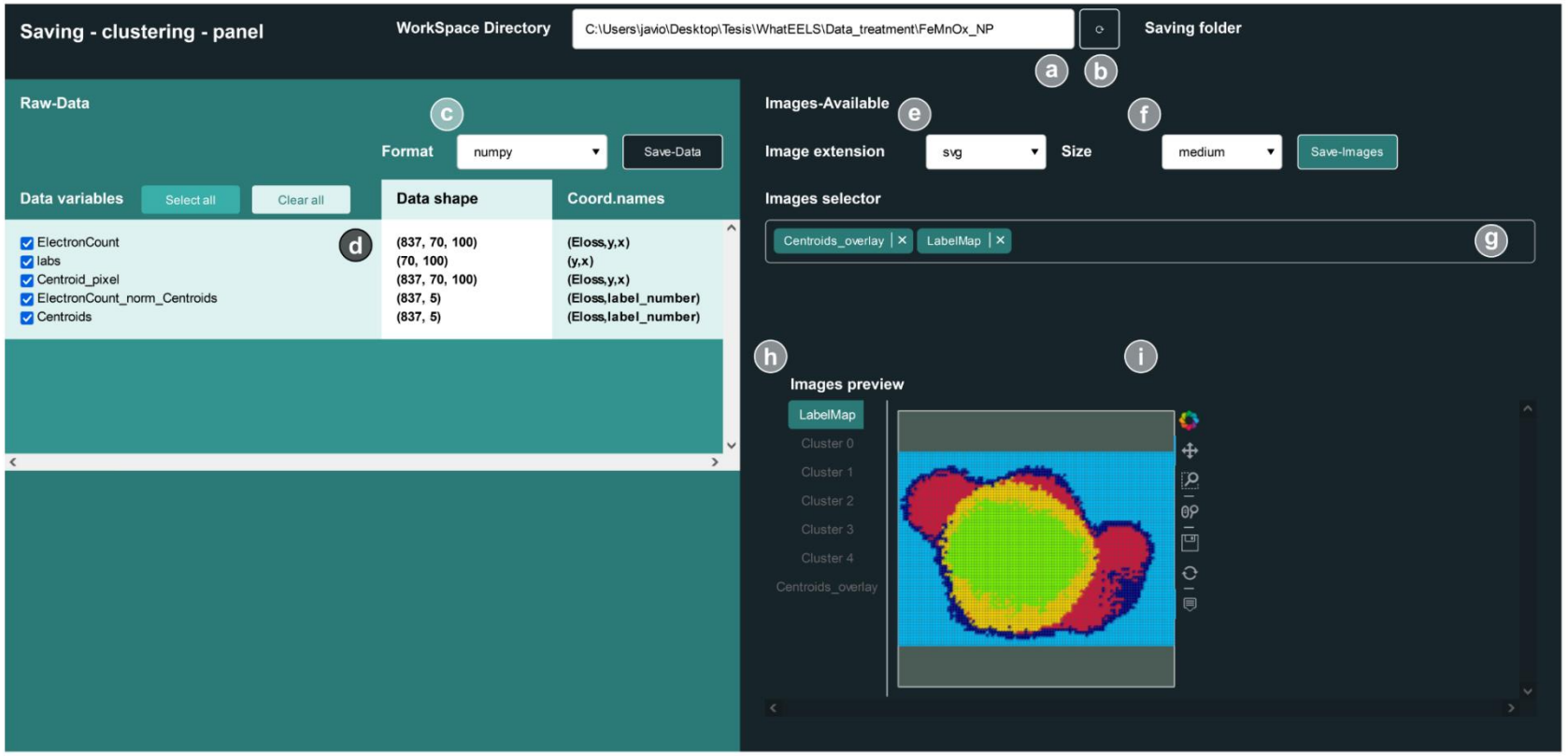


Figure 5.41. Saving panel for the datasets generated by using **WhatEELS**. The specifics about the tools included in this panel were highlighted in **Figure 5.23**

5.6. Conclusions.

This chapter summarizes the development, in the context of the present Thesis, of a new software tool for ELNES analysis: **WhatEELS**. It is designed to combine the power of clustering analysis and the NLLS fitting routines, to make the most of the spectral features near the edge onset of large EELS datasets. **WhatEELS** was laid out with an improved workflow efficiency in mind, one that favours a trial-and-error approach throughout the entirety of the analysis process (e.g., during the model parameters setup, or during the results analysis to set up the integration widths for the WL ratio calculations). **WhatEELS** is fundamentally set on top of three major pillars:

I. It follows a free access philosophy (i.e., it is coded in Python and freely distributed).

II. It presents a modular solution, meaning that any of the functionalities included (e.g., the clustering module **-B-**) can in principle be used independently. This, at the same time, leaves room for further software expansions down the line.

III. As shown throughout this chapter, the **WhatEELS** experience is based on an interactive approach, launching a series of reactive dashboards for each of the included analysis modules that use any of the major web browsers available (e.g., Google Chrome, Mozilla Firefox) as their backend. In principle, this should qualify **WhatEELS** to work in any operative system with a Python interpreter. It also allows any user, with or without any previous programming knowledge, to have access to a solid foundation for ELNES analysis.

A full characterization process using the current **WhatEELS** functionalities for a test sample of iron-manganese oxide core-shell NP was presented here as well. The results revealed localized changes of the manganese oxidation state within the NP shell using two different methods of analysis: WL ratio variations and WL chemical shift measurements. Both routes produced results

in good agreement with the previously measured characteristics of this test sample.

It is worth noticing that the clustering capabilities currently included in **WhatEELS** are far from those showcased in the previous chapter (***Machine Learning for EELS data analysis: Clustering and dimensionality reduction methods).*** At the time of finalizing this manuscript, **WhatEELS** only includes the basic K-means and hierarchical agglomerative clustering solutions from the **Scikit-learn** library. In the future, this advanced dimensionality reduction techniques and clustering algorithms will certainly be included as part of the **WhatEELS** software tool.

Finally, I would like to conclude by putting in perspective once again the importance of the development of **WhatEELS** as a standalone analysis tool (free to use and accessed at http://hdl.handle.net/2445/178745). One of the major problems faced throughout the development of this PhD, especially during the periods dedicated to the pure data analysis, was the clear lack of access to the tools required for an in-depth ELNES characterization. Oftentimes we were faced with insufficient tools provided by the old software solutions, generally unequipped to deal with the large datasets acquired in the current TEM machines. Furthermore, some of the analysis tools described in the literature and given for free by their authors, were coded in a proprietary programming language (i.e., MATLAB) or directly in a proprietary platform (**Digital Micrograph**).

If the TEM community wants to keep up with the path of fast development of analysis techniques driven by the deep learning, clustering and dimensionality reduction fields of machine learning, a fully free-to-use software solution (e.g., coded in Python) capable of an easy assimilation of new modular tools will be required. There is where **WhatEELS** might become an important asset.

5.7. References

[1] R.F.F. Egerton, Electron Energy-Loss Spectroscopy in the Electron Microscope, Springer Science, Boston, MA, 2011. https://doi.org/10.1007/978-1-4419-9583-4.

[2] T. Manoubi, C. Colliex, P. Rez, Quantitative electron energy loss spectroscopy on M45 edges in rare earth oxides, J. Electron Spectros. Relat. Phenomena. 50 (1990) 1–18. https://doi.org/10.1016/0368-2048(90)80001-Q.

[3] F. Hofer, B. Luo, Towards a practical method for EELS quantification, Ultramicroscopy. 38 (1991) 159–167. https://doi.org/https://doi.org/10.1016/0304-3991(91)90117-O.

[4] S. Estradé, L. Yedra, A. López-Ortega, M. Estrader, G. Salazar-Alvarez, M.D. Baró, J. Nogués, F. Peiró, Distinguishing the core from the shell in MnO x/MnO y and FeO x/MnO x core/shell nanoparticles through quantitative electron energy loss spectroscopy (EELS) analysis, Micron. 43 (2012) 30–36. https://doi.org/10.1016/j.micron.2011.04.002.

[5] N. Bonnet, N. Brun, C. Colliex, Extracting information from sequences of spatially resolved EELS spectra using multivariate statistical analysis, Ultramicroscopy. 77 (1999) 1–16. https://doi.org/10.1016/S0304-3991(99)00042-X.

[6] N. Bonnet, D. Nuzillard, Independent component analysis: A new possibility for analysing series of electron energy loss spectra, Ultramicroscopy. 102 (2005) 327–337. https://doi.org/10.1016/j.ultramic.2004.11.003.

[7] R.F. Egerton, Options for Energy-Loss Data Acquisition, in: Electron Energy-Loss Spectrosc. Electron Microsc., Springer US, Boston, MA, 2011: pp. 429–431. https://doi.org/10.1007/978-1-4419-9583-4_11.

[8] N. Dobigeon, N. Brun, Spectral mixture analysis of EELS spectrum-images, Ultramicroscopy. 120 (2012) 25–34. https://doi.org/10.1016/J.ULTRAMIC.2012.05.006.

[9] S. Lichtert, J. Verbeeck, Statistical consequences of applying a PCA noise filter on EELS spectrum images, Ultramicroscopy. 125 (2013) 35–42. https://doi.org/10.1016/j.ultramic.2012.10.001.

[10] P. Potapov, On the loss of information in PCA of spectrum-images, Ultramicroscopy. 182 (2017) 191–194. https://doi.org/10.1016/j.ultramic.2017.06.023.

[11] P. Torruella, M. Estrader, A. López-Ortega, M.D. Baró, M. Varela, F. Peiró, S. Estradé, Clustering analysis strategies for electron energy loss spectroscopy (EELS), Ultramicroscopy. 185 (2018) 42–48. https://doi.org/10.1016/J.ULTRAMIC.2017.11.010.

[12] R.D. Leapman, C.R. Swyt, Separation of overlapping core edges in electron energy loss spectra by multiple-least-squares fitting, Ultramicroscopy. 26 (1988) 393–403. https://doi.org/https://doi.org/10.1016/0304-3991(88)90239-2.

[13] H. Shuman, A.P. Somlyo, Electron energy loss analysis of near-trace-element concentrations of calcium, Ultramicroscopy. 21 (1987) 23–32. https://doi.org/https://doi.org/10.1016/0304-3991(87)90004-0.

[14] R. Door, D. Gängler, Multiple least-squares fitting for quantitative electron energy-loss spectroscopy - an experimental investigation using standard specimens, Ultramicroscopy. 58 (1995) 197–210. https://doi.org/10.1016/0304-3991(94)00198-V.

[15] K. Riegler, G. Kothleitner, EELS detection limits revisited: Ruby - a case study, Ultramicroscopy. 110 (2010) 1004–1013. https://doi.org/10.1016/j.ultramic.2010.02.010.

[16] M.-T. Chang, R.-F. Cai, C.-C. Chen, S.-C. Lo, Development of Clustering Algorithm Applied for the EELS Analysis of Advanced Devices, Microsc. Microanal. 26 (2020) 2112–2114. https://doi.org/10.1017/S1431927620020486.

[17] Y. Wang, M.R.S. Huang, U. Salzberger, K. Hahn, W. Sigle, P.A. van Aken, Towards atomically resolved EELS elemental and fine structure mapping via multi-frame and energy-offset correction spectroscopy, Ultramicroscopy. 184 (2018) 98–105. https://doi.org/10.1016/j.ultramic.2017.10.014.

[18] T. Akita, N. Taguchi, Practical analysis of Li distribution by EELS, Surf. Interface Anal. 48 (2016) 1226–1230. https://doi.org/10.1002/sia.6082.

[19] X. Mu, A. Kobler, D. Wang, V.S.K. Chakravadhanula, S. Schlabach, D. V. Szabó, P. Norby, C. Kübel, Comprehensive analysis of TEM methods for LiFePO4/FePO4 phase mapping: spectroscopic techniques (EFTEM, STEM-EELS) and STEM diffraction techniques (ACOM-TEM), Ultramicroscopy. 170 (2016) 10–18. https://doi.org/10.1016/j.ultramic.2016.07.009.

[20] L. Yedra, E. Xuriguera, M. Estrader, A. López-Ortega, M.D. Baró, J. Nogués, M. Roldan, M. Varela, S. Estradé, F. Peiró, Oxide Wizard: An EELS Application to Characterize the White Lines of Transition Metal Edges, Microsc. Microanal. 20 (2014) 698–705. https://doi.org/10.1017/S1431927614000440.

[21] R.D. Leapman, L.A. Grunes, P.L. Fejes, Study of the L23 edges in the 3d transition metals and their oxides by electron-energy-loss spectroscopy with comparisons to theory, Phys. Rev. B. 26 (1982) 614–635. https://doi.org/10.1103/PhysRevB.26.614.

[22] P.A. Van Aken, B. Liebscher, V.J. Styrsa, Quantitative determination of iron oxidation states in minerals using Fe L2,3-edge electron energy-loss near-edge structure spectroscopy, Phys. Chem. Miner. 25 (1998) 323–327. https://doi.org/10.1007/s002690050122.

[23] T. Riedl, T. Gemming, K. Wetzig, Extraction of EELS white-line intensities of manganese compounds: Methods, accuracy, and valence sensitivity, Ultramicroscopy. 106 (2006) 284–291. https://doi.org/10.1016/j.ultramic.2005.09.002.

[24] Á.M. Arévalo-López, M.Á. Alario-Franco, Reliable method for determining the oxidation state in chromium oxides, Inorg. Chem. 48 (2009) 11843–11846. https://doi.org/10.1021/ic901887y.

[25] F. Baiutti, J. Blanco-Portals, S. Anelli, P. Torruella, M. López-Haro, J.J. Calvino, S. Estradé, M. Torrell, F. Peiró, A. Tarancón, J.B. Portals, S. Anelli, P. Torruella, J.J. Calvino, Tailoring the Transport Properties of Mesoporous Doped Cerium Oxide for Energy Applications, J. Phys. Chem. C. 125 (2021) 16451–16463. https://doi.org/10.1021/acs.jpcc.1c04861.

[26] E.P. Francisco de la Peña Vidar Tonaas Fauske, Pierre Burdet, Petras Jokubauskas, Magnus Nord, … Andreas Garmannslund, hyperspy/hyperspy: HyperSpy v1.5.2, (2019). https://doi.org/10.5281/zenodo.3396791.

[27] J. Blanco-Portals, P. Torruella, F. Baiutti, S. Anelli, M. Torrel, A. Tarancón, F. Peiró, S. Estradé, WhatEELS . A new python-based interactive software solution for ELNES analysis combining clustering and NLLS, Ultramicroscopy. 232 (2021) 113403. https://doi.org/10.1016/j.ultramic.2021.113403.

[28] M. Scheinfein, M. Isaacson, Design and Performance of Second Order Corrected Spectrometers for Use With the Scanning Transmission Electron Microscope, Scanning Electron Microsc. (1984) 1681–1696.

[29] H. Kohl, A simple procedure for evaluating effective scattering cross-sections in STEM, Ultramicroscopy. 16 (1985) 265–268. https://doi.org/10.1016/0304-3991(85)90081-6.

[30] P. Rudiger, X. Artusi, J.A. Bednar, M.S. Madsen, C. B, J. Signell, J.-L. Stevens, M. Liquet, Hoxbro, J. Mease, Andrew, Arne, M. Paprocki, kbowen, E. Jung, H.-Y. Amanieu, J. Winkelmann, L. Talirz, P. A, A. Randelhoff, B. Sullivan, J. Barhak, N. Ghenzi, hoseppan, kleavor, miliante, R. Mulpuri, G. Bischof, A. Thorve, holoviz/panel: Version 0.11.0, (2021). https://doi.org/10.5281/ZENODO.4573728.

[31] P. Rudiger, J.-L. Stevens, J.A. Bednar, B. Nijholt, J. Mease, Andrew, C. B, A. Randelhoff, V. Tenner, maxalbert, M. Kaiser, ea42gh, J. Samuels, stonebig, K. Pevey, F. LB, A. Tolmie, D. Stephan, Hoxbro, J. Bois, S. Lowe, J. Bampton, henriqueribeiro, ruoyu0088, I. Lustig, A. Klein, B. Van de Ven, J. Signell, L. Talirz, L. Barth, holoviz/holoviews: Version 1.14.2, (2021). https://doi.org/10.5281/ZENODO.4581995.

[32] B.V. de V.P.E.M.Z.D.A. Brendan Collins, Bokeh: Essential Open Source Tools for Science, (2020). https://doi.org/10.5281/ZENODO.4317718.

[33] P.T. Inc., Collaborative data science, (2015). https://plot.ly.

[34] T.A. Caswell, M. Droettboom, A. Lee, E.S. de Andrade, J. Hunter, T. Hoffmann, E. Firing, J. Klymak, D. Stansby, N. Varoquaux, J.H. Nielsen, B. Root, R. May, P. Elson, J.K. Seppänen, D. Dale, J.-J. Lee, D. McDougall, A. Straw, P. Hobson, C. Gohlke, T.S. Yu, E. Ma, hannah, A.F. Vincent, S. Silvester, C. Moad, N. Kniazev, E. Ernest, P. Ivanov, matplotlib/matplotlib: REL: v3.4.1, (2021). https://doi.org/10.5281/ZENODO.4649959.

[35] M. Newville, R. Otten, A. Nelson, A. Ingargiola, T. Stensitzki, D. Allan, A. Fox, F. Carter, Michał, D. Pustakhod, lneuhaus, S. Weigand, R. Osborn, Glenn, C. Deil, Mark, A.L.R. Hansen, G. Pasquevich, L. Foks, N. Zobrist, O. Frost, A. Beelen, Stuermer, kwertyops, A. Polloreno, S. Caldwell, A. Almarza, A. Persaud, B. Gamari, B.F. Maier, lmfit/lmfit-py 1.0.2, (2021). https://doi.org/10.5281/ZENODO.4516651.

[36] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, S.J. van der Walt, M. Brett, J. Wilson, K.J. Millman, N. Mayorov, A.R.J. Nelson, E. Jones, R. Kern, E. Larson, C.J. Carey, İ. Polat, Y. Feng, E.W. Moore, J. VanderPlas, D. Laxalde, J. Perktold, R. Cimrman, I. Henriksen, E.A. Quintero, C.R. Harris, A.M. Archibald, A.H. Ribeiro, F. Pedregosa, P. van Mulbregt, A. Vijaykumar, A. Pietro Bardelli, A. Rothberg, A. Hilboll, A. Kloeckner, A. Scopatz, A. Lee, A. Rokem, C.N. Woods, C. Fulton, C. Masson, C. Häggström, C. Fitzgerald, D.A. Nicholson, D.R. Hagen, D. V. Pasechnik, E. Olivetti, E. Martin, E. Wieser, F. Silva, F. Lenders, F. Wilhelm, G. Young, G.A. Price, G.L. Ingold, G.E. Allen, G.R. Lee, H. Audren, I. Probst, J.P. Dietrich, J. Silterra, J.T. Webber, J. Slavič, J. Nothman, J. Buchner, J. Kulick, J.L. Schönberger, J.V. de Miranda Cardoso, J. Reimer, J. Harrington, J.L.C. Rodríguez, J. Nunez-Iglesias, J. Kuczynski, K. Tritz, M. Thoma, M. Newville, M. Kümmerer, M. Bolingbroke, M. Tartre, M. Pak, N.J. Smith, N. Nowaczyk, N. Shebanov, O. Pavlyk, P.A. Brodtkorb, P. Lee, R.T. McGibbon, R. Feldbauer, S. Lewis, S. Tygier, S. Sievert, S. Vigna, S. Peterson, S. More, T. Pudlik, T. Oshima, T.J. Pingel, T.P. Robitaille, T. Spura, T.R. Jones, T. Cera, T. Leslie, T. Zito, T. Krauss, U. Upadhyay, Y.O. Halchenko, Y. Vázquez-Baeza, SciPy 1.0: fundamental algorithms for scientific computing in Python, Nat. Methods. 17 (2020) 261–272. https://doi.org/10.1038/s41592-019-0686-2.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, others, Scikit-learn: Machine learning in Python, J. Mach. Learn. Res. 12 (2011) 2825–2830.

[38] O. Grisel, A. Mueller, Lars, A. Gramfort, G. Louppe, P. Prettenhofer, M. Blondel, V. Niculae, J. Nothman, A. Joly, T.J. Fan, J. Vanderplas, manoj kumar, H. Qin, N. Hug, N. Varoquaux, L. Estève, R. Layton, J.H. Metzen, G. Lemaitre, A. Jalali, R. (Venkat) Raghav, J. Schönberger, R. Yurchak, W. Li, C. Woolam, T.D. la Tour, K. Eren, J. du Boisberranger, Eustache, scikit-learn/scikit-learn: scikit-learn 0.24.1, (2021). https://doi.org/10.5281/ZENODO.4450597.

[39] D. del-Pozo-Bueno, F. Peiró, S. Estradé, Support vector machine for EELS oxidation state determination, Ultramicroscopy. 221 (2021). https://doi.org/10.1016/j.ultramic.2020.113190.

[40] J. Blanco-portals, F. Peiró, S. Estradé, Strategies for EELS Data Analysis . Introducing UMAP and HDBSCAN for Dimensionality Reduction and Clustering, Microsc. Microanal. (2021) 1–14. https://doi.org/10.1017/S1431927621013696.

[41] D.B. Williams, C.B. Carter, The Transmission Electron Microscope, in: Transm. Electron Microsc., Springer US, Boston, MA, 1996: pp. 3–17. https://doi.org/10.1007/978-1-4757-2519-3_1.

[42] T. Manoubi, M. Tencé, C. Colliex, Quantification of white lines in electron energy loss spectroscopy (EELS), Ultramicroscopy. 28 (1989) 49–55. https://doi.org/10.1016/0304-3991(89)90269-6.

[43] A.B. Shah, Q.M. Ramasse, J.G. Wen, A. Bhattacharya, J.M. Zuo, Practical spatial resolution of electron energy loss spectroscopy in aberration corrected scanning transmission electron microscopy, Micron. 42 (2011) 539–546. https://doi.org/10.1016/j.micron.2010.12.008.

[44] M. Varela, M.P. Oxley, W. Luo, J. Tao, M. Watanabe, A.R. Lupini, S.T. Pantelides, S.J. Pennycook, Atomic-resolution imaging of oxidation states in manganites, Phys. Rev. B - Condens. Matter Mater. Phys. 79 (2009) 1–14. https://doi.org/10.1103/PhysRevB.79.085117.

[45] L.A.J. Garvie, P.R. Buseck, Determination of Ce4+/Ce3+ in electron-beam-damaged CeO2 by electron energy-loss spectroscopy, J. Phys. Chem. Solids. 60 (1999) 1943–1947. https://doi.org/10.1016/S0022-3697(99)00218-8.

[46] J.P. Winterstein, C.B. Carter, Electron-beam damage and point defects near grain boundaries in cerium oxide, J. Eur. Ceram. Soc. 34 (2014) 3007–3018. https://doi.org/10.1016/j.jeurceramsoc.2014.02.017.

Chapter 6

Analysis of dopant segregation and Ce oxidation state in doped cerium oxide mesoporous materials.

6.1. Introduction.

This chapter presents a summary of the major results of the transmission electron microscopy (TEM) characterization of a set of cerium oxide (CeO_2) mesoporous materials. Each one of them was doped with a different quantity of gadolinium (Gd) and praseodymium (Pr), targeting the following three compositions: $CeGd_{0.2}O_{0.8}$ (or CGO), $CePr_{0.2}O_{0.8}$ (or CPO) and $CeGd_{0.1}Pr_{0.1}O_{0.9}$ (or CGPO). These stoichiometric values are approximations based on the balance of the precursor solutions used during the synthesis procedure. A fourth mesoporous material was also considered, consisting of a CGO mesoporous structure post-processed with an etching bath in hydrofluoric acid (HF, for 5 minutes) and the addition of an atomic cobalt (1% mol) surface decorator. This set of samples were synthesised by Dr. Albert Tarancón and co-workers at the Catalonia Energy Research Institute (IREC).

All these materials were nanocasted by hard-templating methodology[1], impregnating the porous structure of a KIT-6 SiO_2 (*Ia3d* symmetry) with the solutions of rare earth (RE) precursors. This silica template material presents a chiral double gyroidal (D-gyroid) pattern of channels. Hence, the resulting doped CeO_2 mesoporous materials obtained after the template removal retained the morphology of the complementary structure to the D-gyroid surface, defined by channels of the silica template. An example is provided in

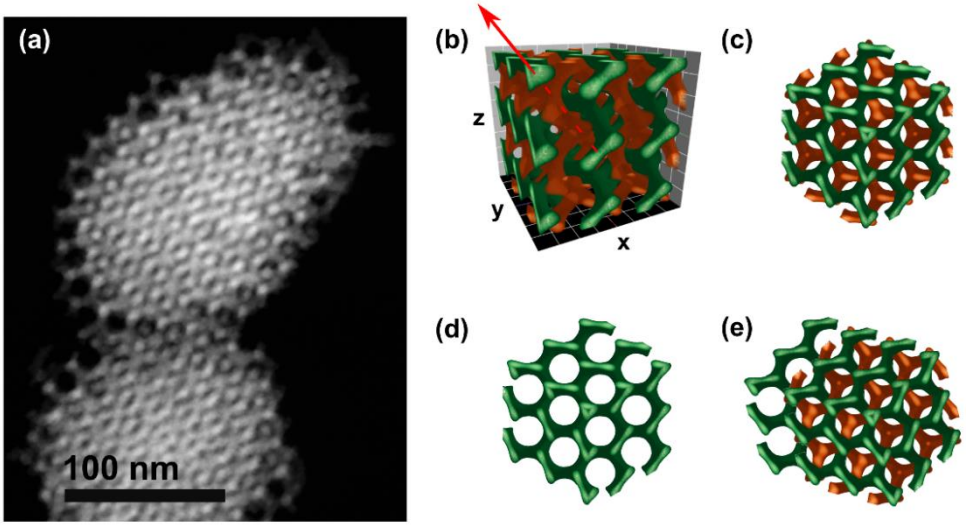


Figure 6.1. (a) HAADF image of two CGPO mesoporous structures. (b) Perspective image of a model of the D-gyroid surface of the KIT-6 silica. The two chiral gyroidal interconnected surfaces are shown in different colours. The [111] symmetry axis is identified with the red arrow. (c) Isometric perspective of the D-gyroid from the [111] symmetry axis. (d) Isometric perspective of the single gyroid from the [111] symmetry axis. (e) Isometric perspective of the composed model of a D-gyroid and open pore single gyroid in the outer corona of the structure.

Figure 6.1. The HAADF image shown in **Figure 6.**1 (a) corresponds to the CGPO sample, although it serves as a representative case for the rest of materials under study as they all showed the same mesoporous morphology. A model of the complementary D-gyroid surface is shown in a perspective view in **Figure 6.**1 (b), and again in **Figure 6.**1 (c) viewed from the [111] axis of the structure (with respect to the *Ia3d* symmetry group of the complementary D-gyroid, marked with a red arrow in **Figure 6.**1 (b)). Each one of the chiral complementary gyroids are coloured differently to serve as visual guides to separate the two set of channels. The entirety of some of the mesoporous structures, and the corona (outer) part in the majority of them presented a partial filling of the KIT-6 channels. This would be the consequence of a partial filling of the template by the precursor solutions, reaching only one of the two existing gyroidal channels in the silica structure. A model of the single complementary gyroid morphology is shown in **Figure 6.**1 (d), viewed from

the [111] axis. Finally, a model of a partial filling event confined to the corona part of the mesoporous material is shown in **Figure 6.**1 (e). Precisely this partial filling in the outer part of the materials is perfectly observable in the example included in **Figure 6.**1 (a).

From now on, these mesoporous structures will be referred to as a D-gyroid, although formally speaking they would correspond to the complementary surface of a D-gyroid.

The formation of dopant clusters was expected to occur in all three, CGO, CPO and CGPO cases. However, the actual cation ratio inside these formations, their location within the granular structure of the mesoporous materials and the homogeneity in their distribution were unknown a priori. Some previous studies showed that the dopants in granular films had a tendency towards the segregation in grain boundaries (GB)[2], so it was a major focal point for this characterization. In fact, the HF etching process and sintering with the Co decorators in the CGO-HF sample was intended to work as a removal procedure of the excess of Gd dopants segregated towards the grain surfaces (GS) and boundaries (GB). In the literature, the formation of these clusters of Gd and Pr has been extensively studied, as they are responsible for the physical mechanisms that affect (positively or negatively) the base electrochemical responses of these types of materials. Moreover, cerium is known to show the capability to hop between the Ce^{3+} and Ce^{4+} oxidation states, which in a reductive ambient or with a suitable metallic dopant concentration would allow the CeO_2 materials (otherwise electron insulators) to act as mixed ion-electron conductors (MIEC)[3–7]. Hence, part of the study was focused on the characterization of the local Ce oxidation state and its relation to the dopant distribution.

This set of mesoporous samples played a central role in the evolution of the work done during the period of this PhD thesis. As a matter of fact, the difficulties encountered to complete the intended experiments for each one of the materials listed above pushed the exploration of the new analysis and ET reconstruction techniques described in each one of the previous chapters. For

instance, the proximity of the Ce-M_{54} and Pr-M_{54} edges, and the overlap between the Ce-M_{32} and Gd-M_{54} edges made it impossible to obtain an accurate elemental quantification based on the Egerton method for the larger datasets required in these experiments using the implementations of the already available tools (GATAN - Digital Micrograph and HyperSpy[8]). These problems were tackled by resorting to the non-linear least-squares (NLLS) fitting analysis of the spectra, although, once again, the available solutions at the time did not provide solutions. This was one of the major driving forces behind the development of the WhatEELS software toolkit, described in the third chapter of this work, which implemented a combined approach of clustering analysis and NLLS to increase the speed, control, and results stability of these types of fitting routines.

The field of clustering analysis was introduced as an EELS characterization technique at the time of the proposal of these experiments with CeO_2 mesoporous materials[9]. The experience gained using these types of algorithms during the characterization of this set of samples revealed a wide variety of potential pitfalls in the methodology, driving the exploration of new routes. As a result, the combined use of dimensionality reduction methods (DRM) based on manifold learning (UMAP[10]) and hierarchical density based clustering algorithms (HDBSCAN[11]) was proposed as a promising solution towards a fully data-driven analysis, extensively detailed in the second chapter of this work[12].

Finally, the intricate morphology of the D-gyroidal channel structure in the mesoporous materials, combined with a poor signal-to-noise ratio (SNR) caused by the register time limitations of SI for the X-EDS analytical reconstructions, acted as the main driving forces behind the implementation in Python of the advanced TVM methodology[13–15] for ET and analytical-ET described in the first chapter of this work.

The timeline of the experiments carried out for these CeO_2 mesoporous materials is shown in **Figure 6.2**, alongside a schematic representation of the

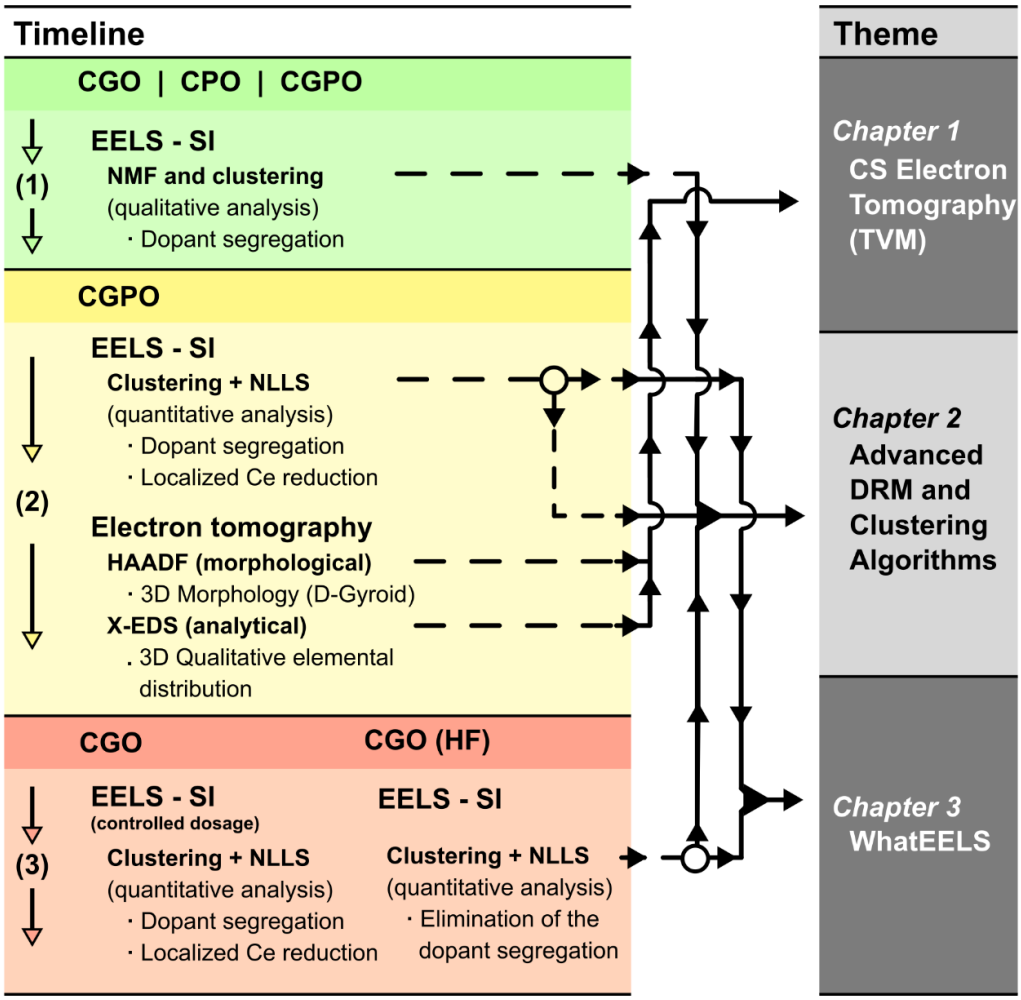


Figure 6.2. Timeline of the experiments carried out for the doped CeO2 set of samples. The main analysis techniques to extract the results are listed below each one of the samples. The flow diagram links the experiments with the analysis techniques described in each one of the previous chapters of this work.

specific contribution of each one of them to the main themes treated in this work.

The structure of this chapter will follow this timeline for the experiments.

- ***CGO, CPO and CGPO. Qualitative analysis.***

  The first set of experiments, carried out in a JEOL-ARM at 200 keV, provided a qualitative view of the dopant distribution in the mesoporous

structures. They also raised awareness about the possible cation reduction induced by an excessive illumination current from the electron beam.

- ***CGPO. Quantitative analysis and electron tomography.***

  The second set of experiments were centred around the CGPO mesoporous material. A series of EELS SI were acquired in the same JEOL ARM at 200 keV. The quantitative characterization of these datasets was carried out combining clustering analysis and NLLS fitting (**WhatEELS**).

  Furthermore, a series of analytical electron tomography (ET) experiments were conducted on this CGPO sample to confirm the hypothesis of dopant segregation towards GB in the whole volume of the mesoporous structures. The acquisition of the datasets used in the reconstructions was carried out in a TITAN Themis at 80 keV (low acceleration voltage in an attempt to reduce the sample damage in the long exposure times of a tomographic acquisition)

- ***CGO and CGO(HF). Quantitative analysis.***

  The last set of experiments were focused on the Gd doped CeO_2 mesoporous materials, both the pristine sample and the postprocessed by the HF etching bath and decorated with cobalt one. An EELS SI analysis was repeated for the CGO, with a special emphasis on the adequate illumination current to avoid artificial Ce cation reduction induced by the electron beam, and being able to resolve the hypothesised local changes in the Ce oxidation state towards GB and GS.

6.2. CGO, CPO and CGPO. Qualitative analysis.

A series of conventional EELS experiments were carried out to test the hypothesis of dopant segregation towards GB and GS within the polycrystalline structure of the CPO, CGO and CGPO mesoporous materials. These initial experiments were done without carefully considering the electron dosage over the samples. As a consequence, any localized change of

the Ce oxidation state observed in the results from these experiments could not be confidently discarded as a beam damage artifact.

All the results presented here (for the CGO, CPO and CGPO) follow the same structure. (i) A first panel including the location of the regions of interest (ROIs) for the EELS-SI acquired from the mesoporous materials is presented. (ii) A second panel shows the results of a non-negative matrix factorization (NMF)[16] DRM applied directly on the raw EELS-SI. The analysis of the archetype signals resolved by NMF algorithm constituted the first attempt to qualitatively identify possible dopant segregation regions within the samples. Finally, (iii) a non-supervised segmentation of the raw dataset is presented using the combination of UMAP DRM and HDBSCAN clustering analysis, following the methodology described in chapter 2 of this work (***Machine Learning for EELS data analysis: Clustering and dimensionality reduction methods***).

6.2.1. Pr-doped cerium oxide mesoporous material (CPO).

Beginning with the CPO, a first example of the mesoporous materials under analysis is shown in **Figure 6.3** (a). The ROI (**Figure 6.3** (b), (c)) was placed in the corona part of a mesoporous material, encompassing a GB region between CeO_2 grains. The raw spectrum presented in **Figure 6.3** (d) provides an idea of the noise levels of the dataset under analysis, indicating a underwhelmingly poor sNR.

The NMF results for this first area of the CPO sample are shown in **Figure 6.4**. Only the 3 (of 10) major components, the ones with the higher variance ratio values, are included in the figure. The rest of them corresponded to noisy components in the dataset. The component labelled as 0 (i.e., the loading image in **Figure 6.4** (b) and the archetype signal coloured in grey in **Figure 6.4** (e)) can be ascribed to the background signal, and accounts for the thickness variation on the sample. Component 1 (i.e., the loading image in **Figure 6.4** (c) and archetype signal coloured in green in **Figure 6.4** (e)) corresponds to the major contribution of Ce signal to the sample. The energy loss position of the

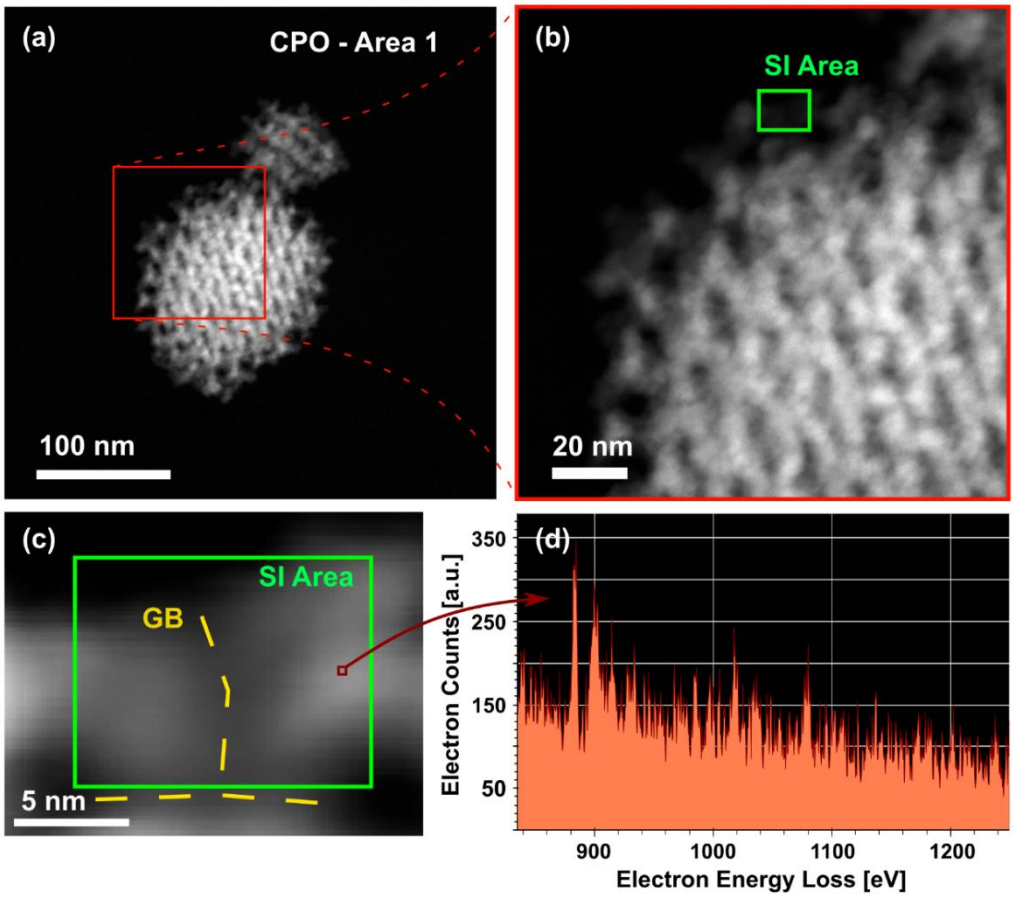


Figure 6.3 (a) HAADF image of a CPO mesoporous material. (b) Zoomed image of the area of interest, including the ROI area where the EELS SI was acquired. (c) Co-acquired HAADF image of the ROI. (d) Example of a single pixel spectrum extracted from the EELS-SI in (c). The GB is highlighted with yellow markings in (c).

CeM_{54} WLs and their apparent intensity ratio ($I(CeM_5)/I(CeM_4)$) correspond to the Ce^{3+} oxidation state. Comparing **Figure 6.4** (a), (b) and (c) it becomes apparent that the Ce signal clearly follows the grain structure of the ROI area. Finally, component 2 (i.e., the loading image in **Figure 6.4** (d) and the archetype signal coloured in light blue in **Figure 6.4** (e)) would correspond to the Pr dopant spectral component. The morphology of the image in **Figure 6.4** (d) is indicative of the formation of a series of clusters of higher Pr concentration in localized areas of the ROI.

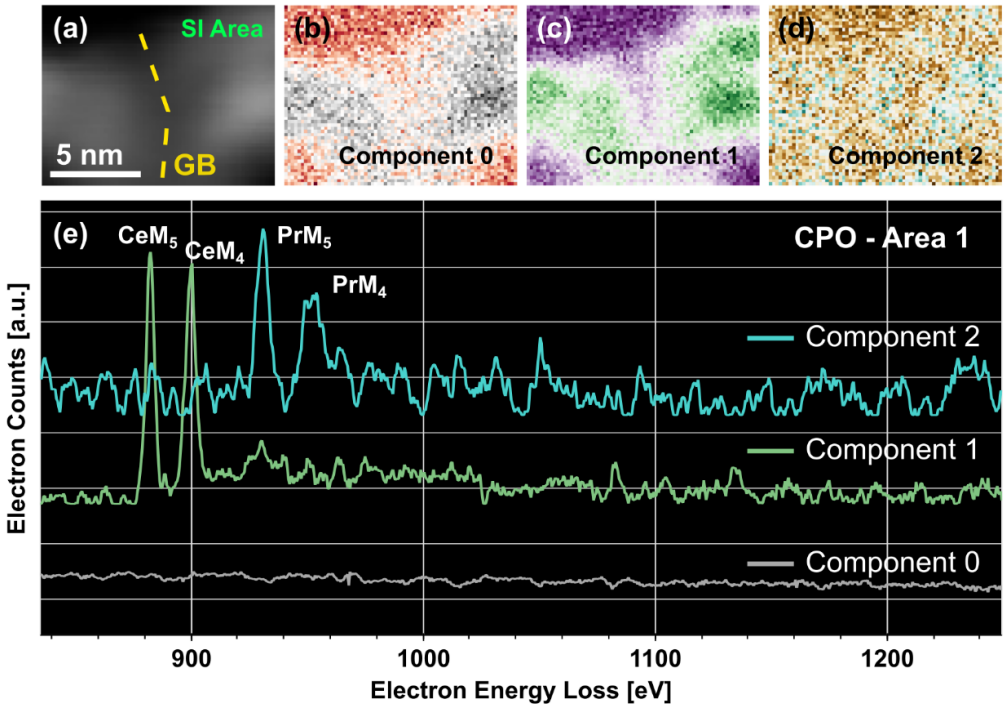


Figure 6.4 (a) Co-acquired HAADF image of the ROI area for the EELS SI shown in **Figure 6.3**. (b-d) Representation images (loadings) of the 3 major components resolved by the NMF algorithm. (e) Archetype signals for the components shown in (b-d). The spectra are normalized individually and staggered to be able to visualize all the different details.

The results of the clustering analysis for this first area of interest in the CPO sample are shown in **Figure 6.5**. The 2D projection of the UMAP embedding resolved for the noisy raw dataset, and coloured according to the HDSCAN clustering results, is shown in **Figure 6.5** (b). A 'cosine' distance measurement was selected for the computation of the high dimensional embedding of the dataset, and the energy range was limited to the interval between 850 and 1050 eV (i.e., focused on the spectral variations in the energy loss regions containing the onsets of the Ce and Pr edges). As explained in chapter 2, this configuration for the UMAP computations induces a damping effect on the weight of the total L_2-norm for the SI. This distance measurement (cosine) will be kept for the UMAP calculations in the ongoing section.

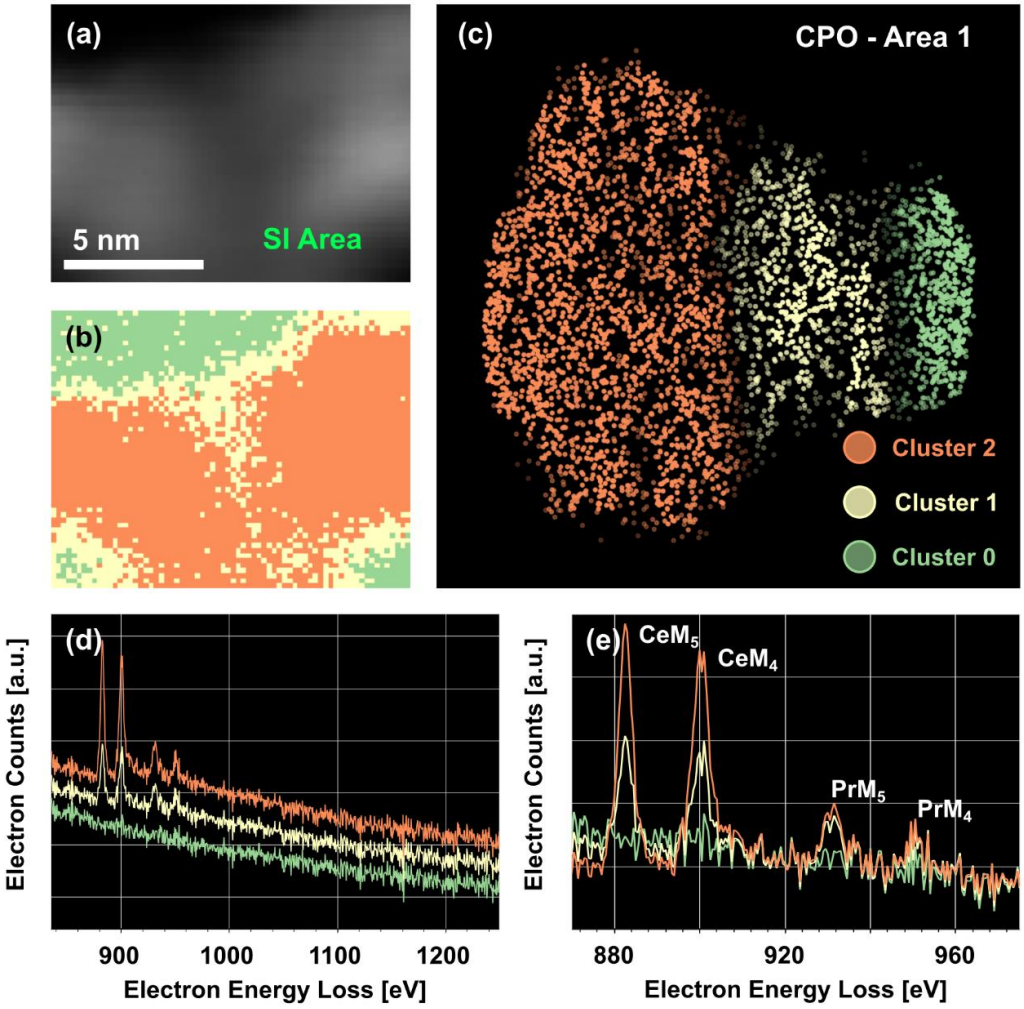


Figure 6.5. (a) Co-acquired HAADF image of the ROI area for the EELS SI shown in **Figure 6.3**. (b) Label map resolved by HDBSCAN over the manifold projection computed by UMAP and shown in (c). (d) Average EELS signal for each one of the clusters resolved by HDBSAN, artificially staggered in the y-axis to help with the visualization of details. (b) Detailed view of the normalized spectra focused on the energy loss region for the Ce and Pr edges.

The embedding projection shows a somehow fuzzy distribution of datapoints, mainly driven by the low sNR of the spectra. Nevertheless, the HDBSCAN clustering was capable of separating 3 spectral regions, which are in good agreement with the morphology of the grains highlighted in **Figure 6.3** (c). The average signals for the clusters resolved (**Figure 6.5** (d)) show an increment of the Pr/Ce ratio towards GB and GS (cluster 1 in yellow). This effect is clearly visible in the detailed view of the L_2-normalized average

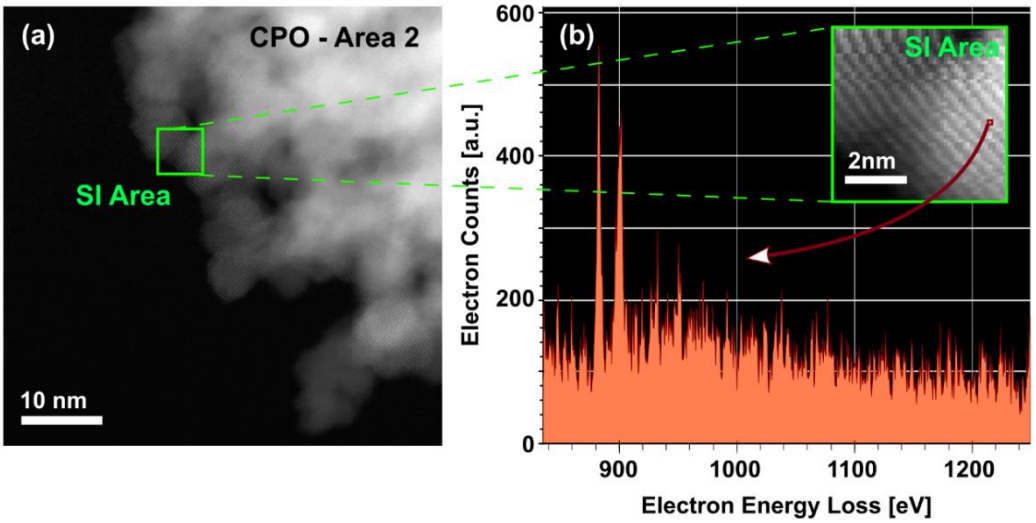


Figure 6.6. (a) HAADF image of a CPO mesoporous material, that includes highlighted in green the ROI area for the EELS SI acquisition. (b) Example of a single pixel spectrum extracted from the EELS-SI. The green inset corresponds to the co-acquired HAADF image of the ROI during the EELS SI acquisition experiment.

spectra shown in **Figure 6.5** (e). The normalization was carried out for the whole energy loss range of the spectrum image and serves the purpose of scaling all the components to facilitate the qualitative comparison. In both the reference spectra for cluster 1 (GB-GS) and cluster 2 (bulk), the CeM_{54} WL configuration is indicative of a dominant Ce^{3+} oxidation state throughout the whole ROI.

A second region was analysed to cross-check the results obtained for this CPO sample. The second ROI for the EELS SI acquisition, along an example of a single spectrum within the SI area, are shown in **Figure 6.6**. Once again, the ROI is placed over a boundary interface between CeO_2 grains in the outermost part (corona) of the mesoporous structure. Comparing the spectrum of a single pixel in this case (**Figure 6.6** (b)) with the single pixel spectrum of the previous sample (**Figure 6.3** (d)), it becomes apparent the improvement of the sNR in this particular example.

The NMF decomposition results are shown in **Figure 6.7**. The components 0-2 (i.e., the loadings images in **Figure 6.7** (b-d) and grey, green, and light blue

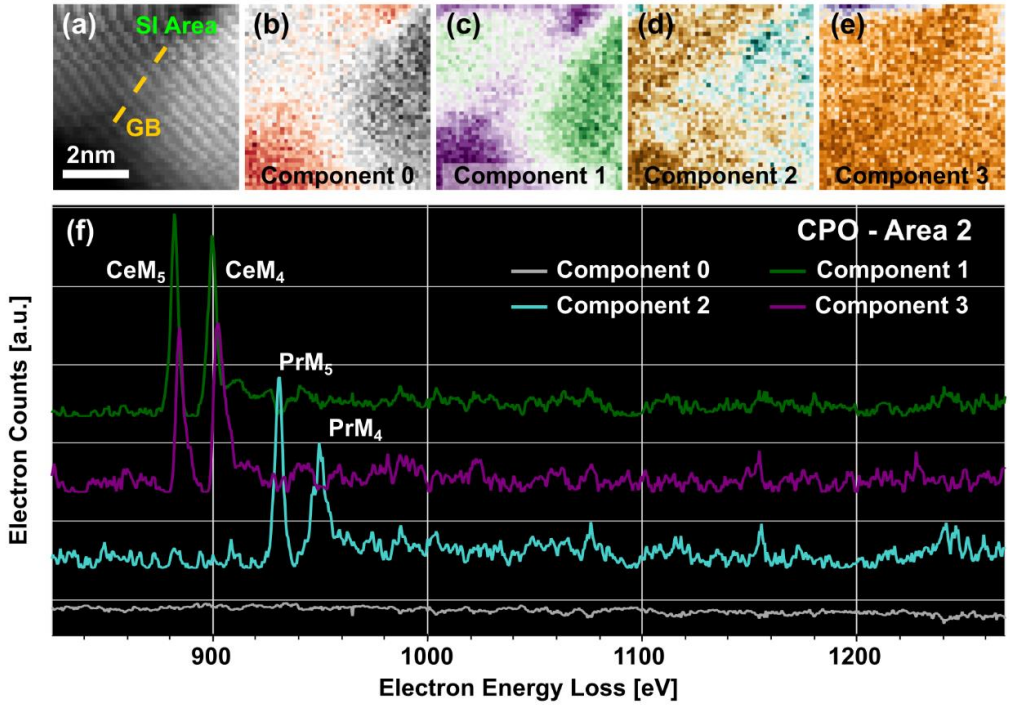


Figure 6.7 (a) Co-acquired HAADF image of the ROI area for the EELS SI shown in **Figure 6.6**. (b-e) Representation images (loadings) of the 4 major components resolved by the NMF algorithm. (f) Archetype signals for the components shown in (b-e). The spectra are individually normalized and staggered to be able to visualize all the different details in them.

coloured signals **Figure 6.7** (f)) would correspond once again to spectral contributions of the thickness, CeO_2 (Ce^{3+}) and Pr dopant respectively. However, owed to the higher sNR the spectral separation of these components and their interpretability is much easier for this second ROI. A fourth component (**Figure 6.7** (e) and purple signal in **Figure 6.7** (f)) was separated from the noise background this time. Component 3 (purple) would also correspond to a region of the CeO_2 within the sample (much like component 1, in green). However, the change in the Ce M_{54} ratio and the displacement towards higher energy loss values would be indicative of a dominant Ce^{4+} oxidation state in this area.

The UMAP-HDBSCAN clustering results for this second ROI are shown in **Figure 6.8**. The higher sNR provided a better ground for the manifold approximation and projection done by UMAP, which in turn resulted in a

better separation of the spectrally relevant areas by clustering with HBDSCAN (**Figure 6.8** (b) and (c) respectively). This methodology was capable of separating the Ce^{4+} rich area at the top of the ROI (cluster 0, in blue) already hinted by the NMF results, as well as the GB and surface areas (cluster 4, in

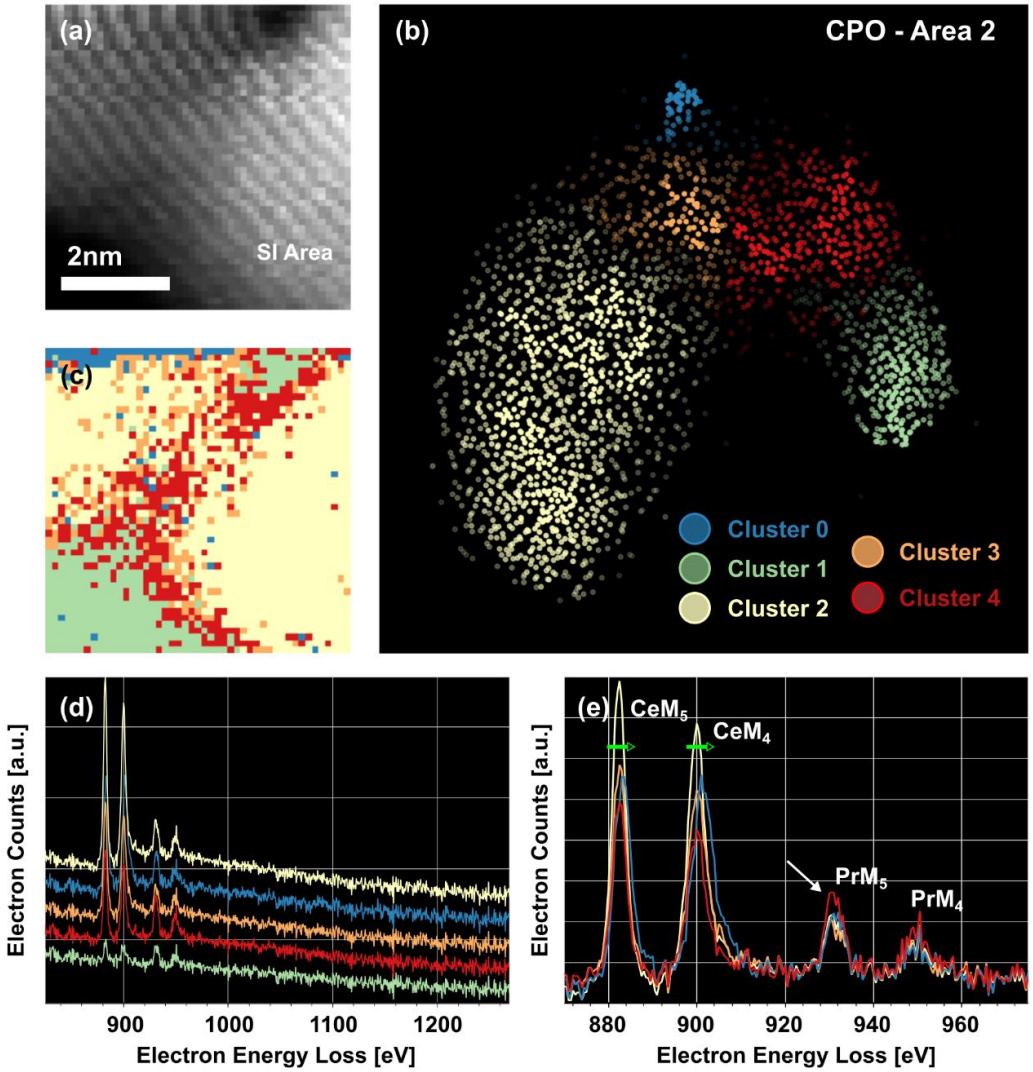


Figure 6.8. (a) Co-acquired HAADF image of the ROI area for the EELS SI shown in **Figure 6.6**. (b) Projection of the embedding resolved by UMAP to the 2D plane for the ROI shown in (a), coloured according to the label map resolved by HDBSCAN shown in (c). (d) Average EELS signal for each one of the clusters resolved by HDBSAN, artificially staggered in the y-axis to help with the visualization of the details. (b) Detailed view of the normalized spectra focused on the energy loss region for the Ce and Pr edges. The cluster 1 signal, background areas, is not presented here as it opaques partially the other spectra.

red). The L_2-normalized average signals detailed for the Ce and Pr edge energy loss region in **Figure 6.8** (e) clearly show the displacement of the CeM_{54} WLs towards higher energy loss values and the change in the M_5/M_4 ratio in blue for cluster 0 (i.e., the Ce^{4+} dominance in the region) At the same time, the relative increment of the PrM_{54} WL intensity with respect to the CeM_{54} WLs in clusters 3 (orange) and 4 (red) indicate a dopant segregation effects in these areas, especially in the case of cluster 3. At the same time, the CeM_{54} WLs ratio and position are indicative of a dominant Ce^{3+} oxidation state in clusters 2 (yellow), 3 (orange) and 4 (red).

These results for the CPO samples indicate a tendency of the Pr dopant to segregate towards the GB and GS, and an overall dominance of the Ce^{3+} oxidation state. However, the little region identified with a Ce^{4+} oxidation state in the second ROI raises a reasonable doubt about the possible cation reduction effects induced by the uncontrolled electron dose during the acquisition. Its location at the top-left part of the SI area coincides with the initial position in the scanning process during the acquisition (i.e., the spectra that was acquired with the lower cumulative dosage transmitted to the sample, as these pixels are the first areas illuminated). The Ce signal for

the rest of the spectra in this ROI area shows a quick transition towards the Ce^{3+} oxidation state (i.e., a larger M_5 WL and a displacement towards lower energy loss values of the CeM_{54} onset). According to the component 0 resolved by the NMF (i.e., the thickness related one in **Figure 6.7** (b) with

the grey spectrum signal in **Figure 6.7** (f)) these parts of the SI over the CeO_2 grain structure that do not belong to cluster 0 (blue in **Figure 6.8** (c)) would correspond to thicker areas on the sample. The fact that they present a predominant Ce^{3+} oxidation state, in conjunction with the direction of the beam scanning motion, is in good agreement with the hypothesis of cation reduction induced by the cumulative charge transmitted to such small ROI area of the sample (notice that the SI covers a practically small patch with an area 4 nm^2) during the acquisition process. The thinner area of Ce^{4+} only

remains in its hypothetical initial oxidation state because is measured at the very beginning of the acquisition.

6.2.2. Gd-doped cerium oxide mesoporous material (CGO).

The ROI area where the EELS SI for the mesoporous CGO material was acquired is showcased in **Figure 6.9** (a). It corresponds once again to a region of the mesoporous corona with an interface between several grains dangling over the vacuum in a grid hole. An example of a single pixel spectrum from this SI is shown in **Figure 6.9** (b). The sNR is similar to that of the second area investigated for the CPO sample (**Figure 6.6** (b)).

The NMF results are shown in **Figure 6.10**. In this case a grand total of 7 components contained relevant information about the sample. For example, component 1 (the loadings image in **Figure 6.10** (c) and the green archetype signal in **Figure 6.10** (i)) show spectral characteristics corresponding to a dominant composition of CeO_2 with a Ce^{4+} oxidation state. Components 4, 6 and 9 (loading mappings in **Figure 6.10** (e), (f) and (h), and archetypes in dark blue, pink and light blue in **Figure 6.10** (i), respectively) show a clear

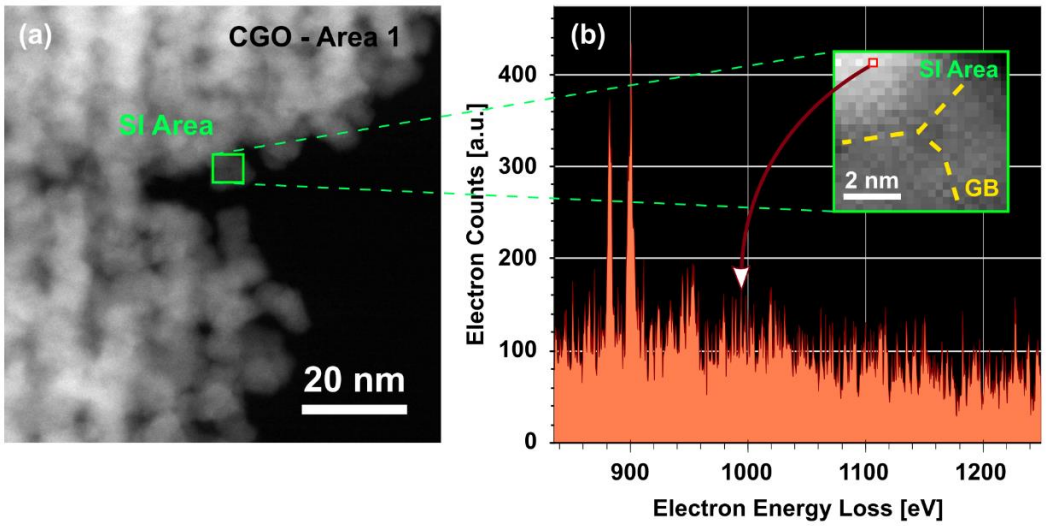


Figure 6.9 (a) HAADF image of a CGO mesoporous material. The square highlighted in green corresponds with the ROI area for the EELS SI acquisition. (b) Example of a single pixel spectrum extracted from the EELS-SI. The green inset corresponds to the co-acquired HAADF image of the ROI during the EELS SI acquisition experiment.

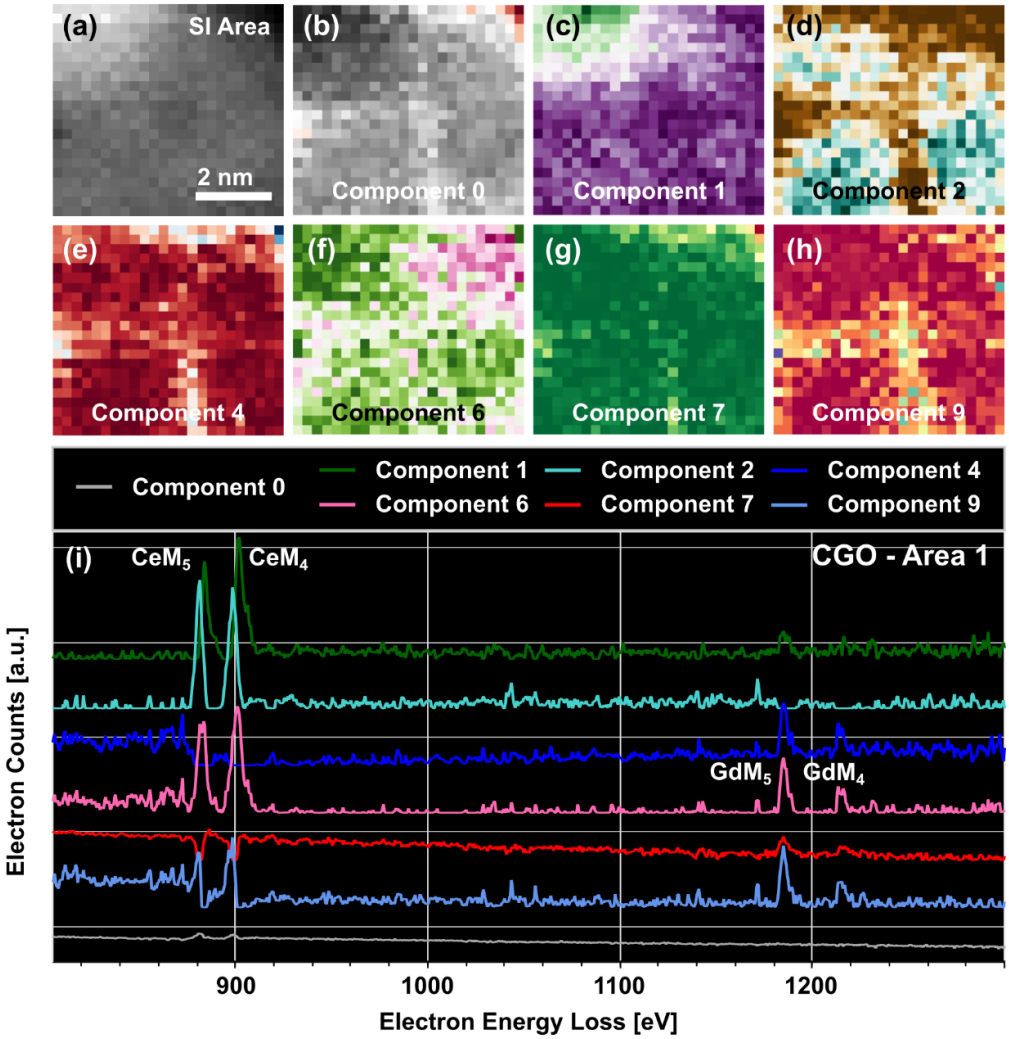


Figure 6.10. (a) Co-acquired HAADF image for the EELS SI of the CGO mesoporous sample presented in **Figure 6.9**. (b-h) Representation images (loadings) of 7 of 10 components resolved by the NMF algorithm. (i) Archetype signals for the components shown in (b-h). The spectra are individually normalized and staggered to be able to visualize all the different details in them.

increment in the Gd WLs signal, spatially localized in the GBs and GS. Meanwhile, component 2 (loadings image in **Figure 6.10** (d) and archetype signal in cyan in **Figure 6.10** (i)) show the typical spectral characteristic of a transition of the Ce to a 3+ oxidation state. The problem with such a large number of spectrally relevant components is that some of them present colliding characteristics in the same spatial positions (i.e., the same pixel in the

loading images). This clearly complicates the analysis and segmentation of the SI in spectrally relevant regions.

The clustering results for this ROI area are presented in **Figure 6.11**. Despite the lower number of pixels in the SI (i.e., the limited statistical knowledge about the sample for the computations carried out by UMAP, done once again selecting a cosine distance measurement), the projected manifold (**Figure 6.11** (c)) allowed HDBSCAN to produce a sensible clustering classification (**Figure 6.11** (b)). The detailed view of the normalized spectra included in the inset regions of (**Figure 6.11** (d)), shows a tendency of the Gd dopant to segregate towards the GB and GS regions (clusters 3 in red and 4 in orange). They also show the existence of a region in the bulk of one of the grains (top-left corner of the ROI area) with a dominant contribution of CeO_2 to the spectra and a Ce^{4+} oxidation state (cluster 0, blue). Conversely, the bulk of two grains in the bottom of the ROI present a dominant contribution of CeO_2 to the spectra with a predominant Ce^{3+} oxidation state (cluster 2, in yellow). Finally, the area of cluster 2 (green) shows an increment of the Gd dopant and a transition towards a Ce^{3+} oxidation state in the GS from the bulk area of cluster 0 (blue). This mixture of Ce^{3+} and Ce^{4+} oxidation stated is represented by the widening Ce M_{54} WLs and their ratio variation in the spectrum coloured in green (cluster 1) inside the green inset in **Figure 6.11** (d).

Once again, the dopants present a clear tendency to segregate towards the GB and GS. Also, similarly to the case of the CPO material, the spatial location of the only region resolved with a Ce^{4+} oxidation state raises suspicions about the validity of the conclusions about the oxidation state variation extracted from this qualitative analysis. Given that cluster 0 (in blue in **Figure 6.11** (b)) is placed at the beginning of the scanning ROI area, the apparent spatially resolved cation reduction for the rest of the clusters could be a direct consequence of electron beam damage produced during the acquisition process.

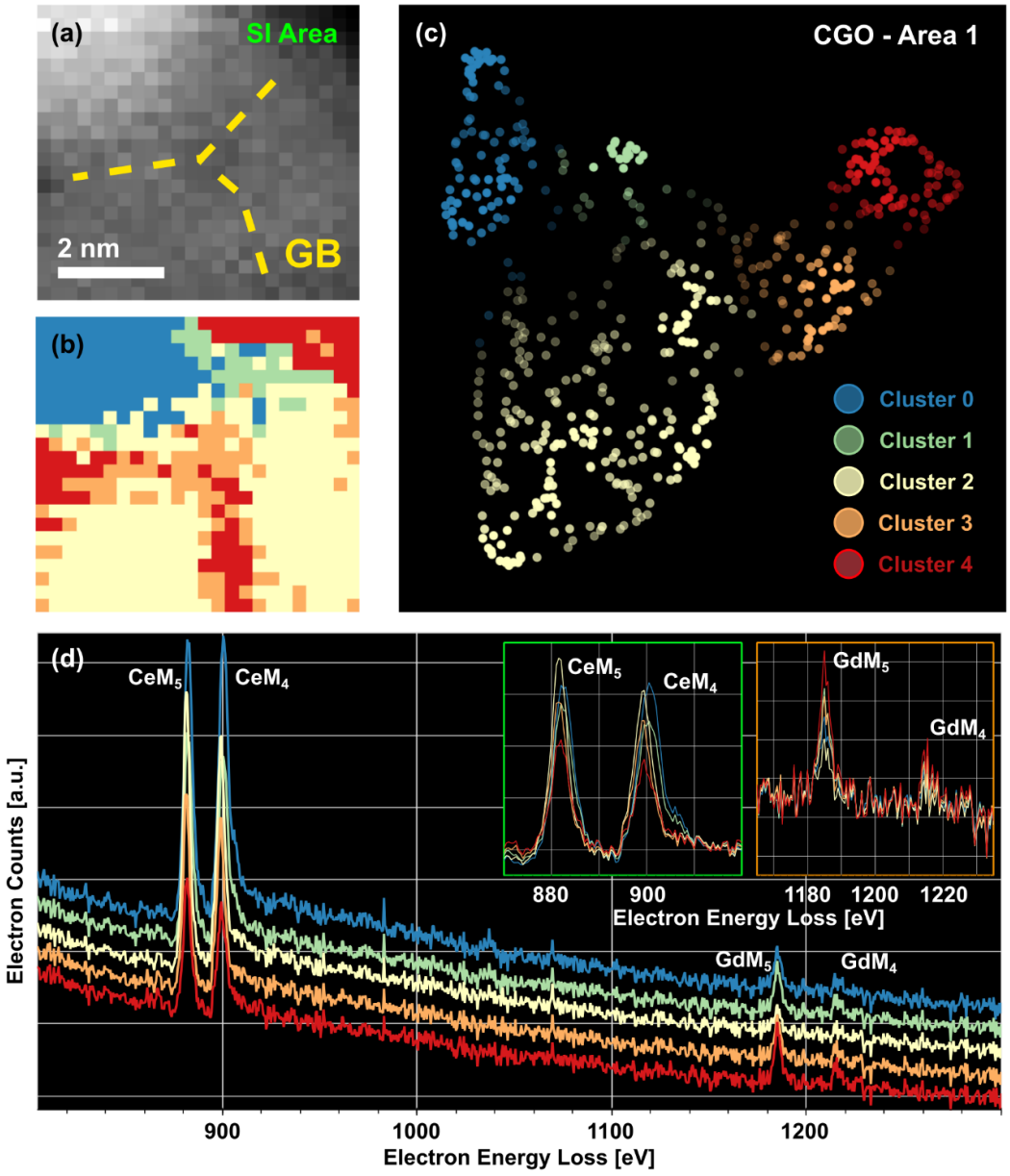


Figure 6.11. (a) Co-acquired HAADF image of the ROI area for the CGO EELS SI shown in **Figure 6.9**. (b) Label map resolved by HDBSCAN over the manifold projection computed by UMAP and shown in (c). (d) Average EELS signal for each one of the clusters resolved by HDBSAN, artificially staggered in the y-axis to help with the visualization of the details. The green inset area shows a detailed view of the normalized average spectra in the energy loss range of the Ce edge onset. The orange inset area shows a detailed view of the normalized average spectra in the energy loss range of the Gd edge onset.

6.2.3. Gd-Pr-doped cerium oxide mesoporous material (CGPO).

Two different regions belonging to a single structure of the CGPO mesoporous sample are included here for an initial qualitative analysis. The ROI areas are shown in **Figure 6.12** (a), along with a single pixel spectrum example for each one of them in **Figure 6.12** (b).

The results for the NMF decomposition for the EELS SI from the first area in **Figure 6.12** are shown in **Figure 6.13**. The background component is not included this time. The archetype for component 1 (green signal from **Figure 6.13** (d)) shows ELNES characteristics compatible with a predominant Ce^{4+} oxidation state in the CeO_2, whereas component 2 (cyan signal in **Figure 6.13** (d)) shows both the appearance of the GdM_{54} WLs and an ELNES structure for the CeM_{54} WLs consistent with a Ce^{3+} oxidation state. Component 3 appears to describe a possible Pr dopant segregation towards the GB region.

The clustering classification from **Figure 6.14** confirms the interpretations given to the NMF decomposition results. As with the case of the previous CGO sample, the small number of datapoints provided by the EELS SI produces a

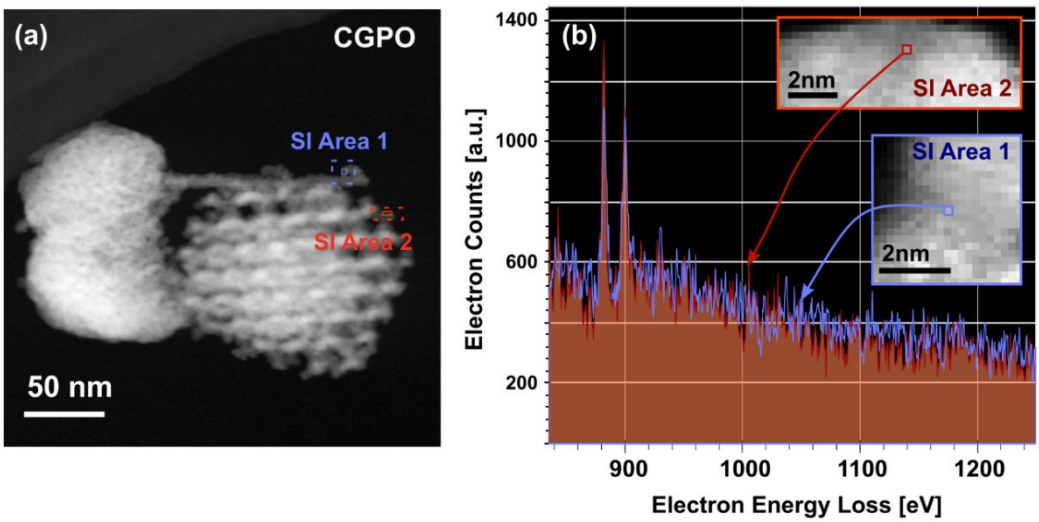


Figure 6.12. (a) HAADF image of a CGPO mesoporous structure nearby a residue of the KIT-6 silica template. Two ROI areas are highlighted over the mesoporous structure. (b) Overlay of single pixel spectra examples for the ROI areas 1 (blue) and 2 (red).

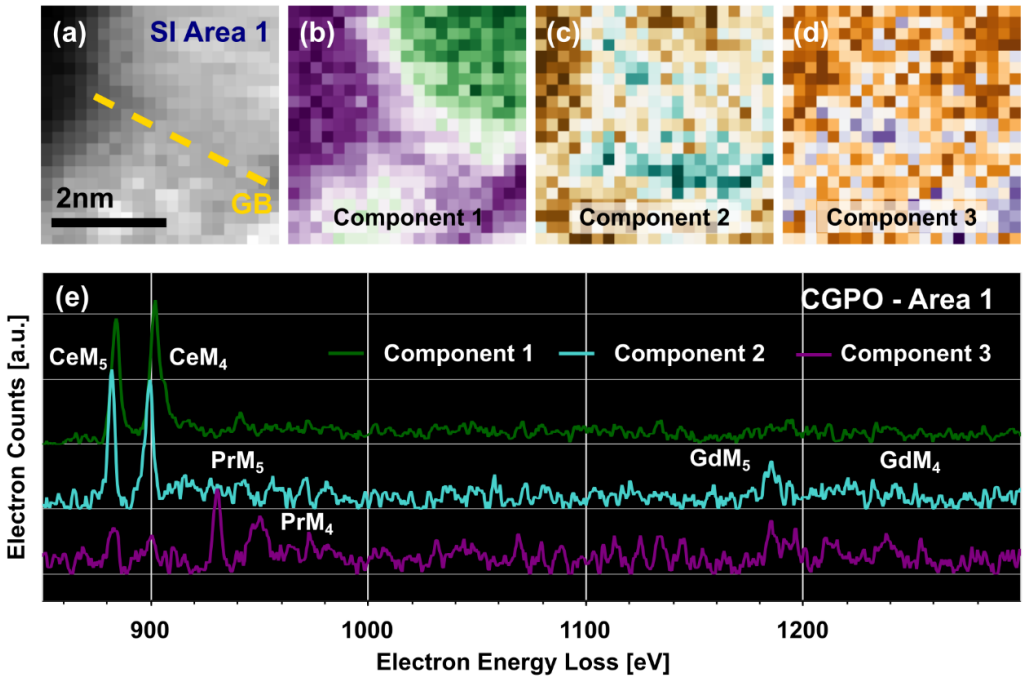


Figure 6.13. (a) HAADF image co-acquired with the EELS SI in the ROI area number 1 shown in **Figure 6.12** (a). (b-d) Representation images (loadings) for 3 of the 10 components resolved by the NMF algorithm. (e) Archetype signals for the components shown in (b-d). The spectra are individually normalized and staggered for visualization purposes.

rather sparse UMAP projection (computed using a cosine distance measurement for the manifold learning process). Nonetheless, once again the HDBSCAN clustering algorithm was able to segment the different spectrally relevant areas of the dataset. The detailed view of the normalized spectra in the energy range containing both the CeM_{54} and PrM_{54} edges (i.e., the inset highlighted green in **Figure 6.14** (d)) indicates that cluster 0 (blue) and 3 (orange) would contain a mixture of Ce^{3+} and Ce^{4+} oxidation states, with a higher concentration of the latter in the case of cluster 0. Also, cluster 3 (orange) shows an increment in the relative intensity values for both the Pr (green inset) and Gd (orange inset in **Figure 6.14**(d)) with respect to the Ce WLs. Clusters 1 (green) and 4 (red) correspond to two different areas of the GB contained within the ROI. Both show a predominant content in Ce^{3+}. However, cluster 1 shows a more predominant segregation of the Pr whereas

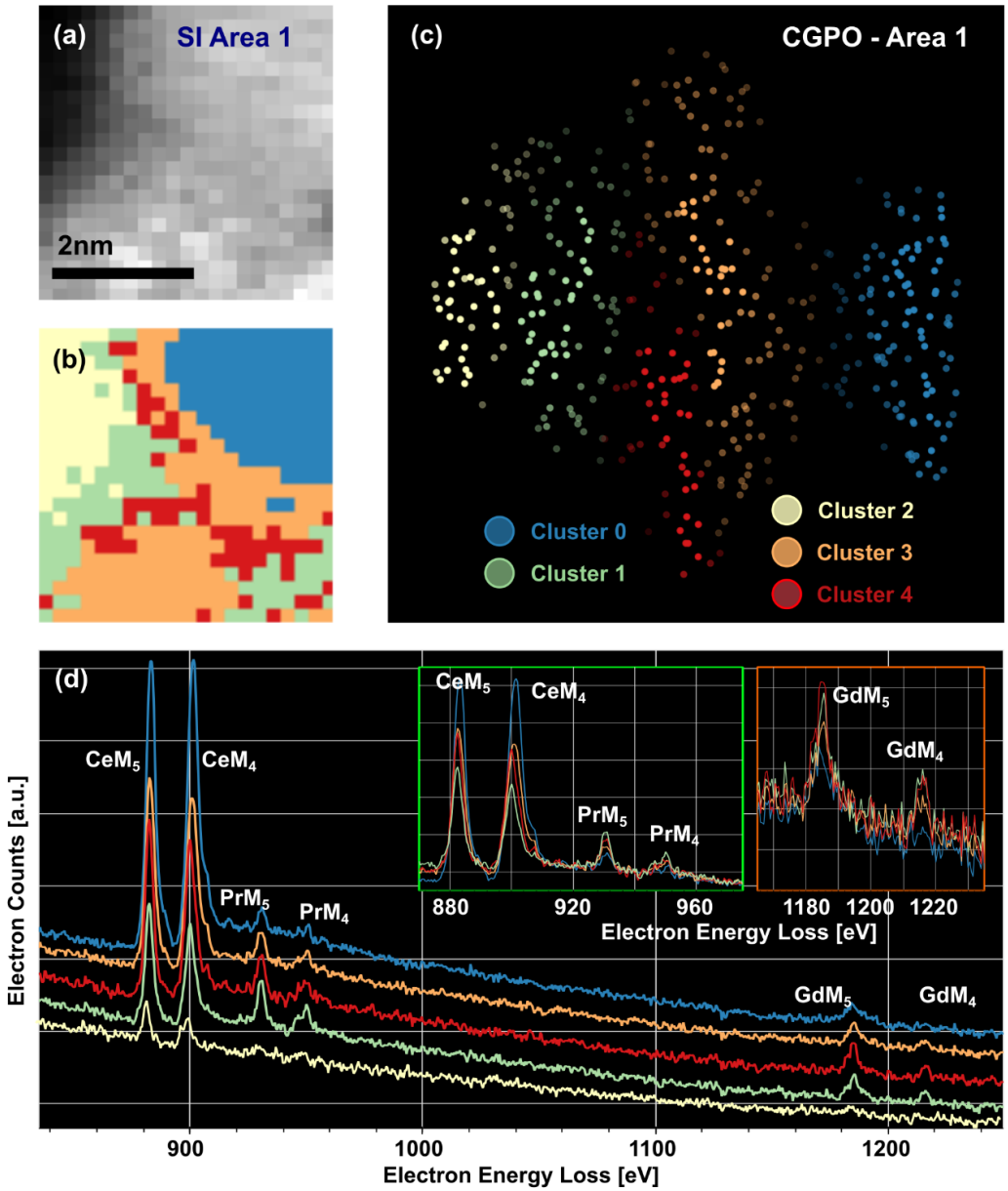


Figure 6.14 Co-acquired HAADF image of the ROI area number 1 for the CGPO EELS SI shown in **Figure 6.12**. (b) Label map resolved by HDBSCAN over the manifold projection computed by UMAP and included in (c). (d) Average EELS signal for each one of the clusters resolved by HDBSAN, staggered in the y-axis for visualization reasons. The green inset area shows a detailed view of the normalized average spectra in the energy loss range of the Ce and Pr edges onsets. The orange inset area shows a detailed view of the normalized average spectra in the energy loss range for the Gd edge onset.

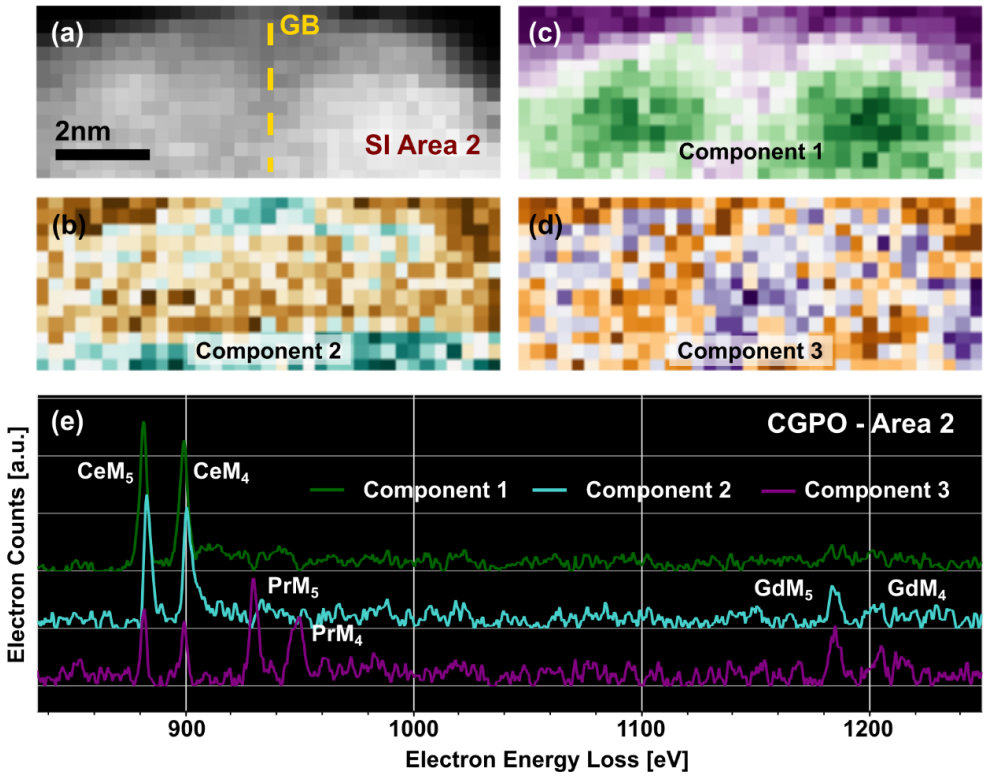


Figure 6.15. (a) HAADF image co-acquired with the EELS SI in the ROI area number 2 shown in **Figure 6.12** (a). (b-d) Image loadings for 3 of the 10 components resolved by the NMF algorithm. (e) Archetype signals for the components shown in (b-d). The spectra are individually normalized and staggered for visualization purposes.

the average signal for cluster 4 indicates a more predominant segregation of the Gd dopant.

The NMF decomposition results for area 2, highlighted in red in **Figure 6.12**, are shown in **Figure 6.15**. Overall, they mimic the behaviour described for the area number 1 (**Figure 6.13**). Nonetheless, this time the Ce^{4+} oxidation state appeared more predominantly in the archetype corresponding to the component number 2 (cyan curve in **Figure 6.15** (d)) that includes also the GdM_{54} WLs, whereas component 1 shows a predominant content of Ce^{3+}.

The clustering results for this area are shown in **Figure 6.16**. This time the Ce oxidation state appears to be Ce^{3+} in all the SI area. However, it is possible that part of the spectral variation discovered by the NMF decomposition (the Ce^{4+} rich region) was lost in the UMAP projection (**Figure 6.16** (c), where the

manifold is learnt using a cosine distance measurement) given the limited statistical information available in such a small SI. In any case, the different segregation behaviour of the Pr and Gd dopants is once again observable, as the Pr segregation appears similar for clusters 1 (green) and 4 (red) and the Gd segregation seems more predominant in in cluster 4 (red) (see the inset detail of normalized average signals in **Figure 6.16** (d)).

Recapitulating, all the samples (CPO, CGO and CGPO) presented some sort of dopant (Pr and Gd) segregation towards GB and GS. This behaviour was observed both in the NMF decompositions and the clustering analysis results. With regards to the Ce oxidation state variations, only the CGPO showed hints of a behaviour compatible with an actual Ce cation reduction towards the GB. However, in one of the areas the NMF and clustering analysis showed conflicting results with regards to this property. The samples of CGO and CPO also showed the coexistence of two different Ce oxidation states within the areas analysed. Unfortunately, given the actual spatial distribution resolved by both the NMF and clustering methodologies, this selective reduction effect could be attributed to the electron beam damage over the samples.

Overall, the NMF decomposition results provided a good initial intuition of the spectral distribution for all samples involved. The conflicts between some of the spectral features packed in each one of the NMF components were later avoided in the clustering analysis combining UMAP and HDBSCAN. This methodology was once again proven capable to perform well under severe noise conditions.

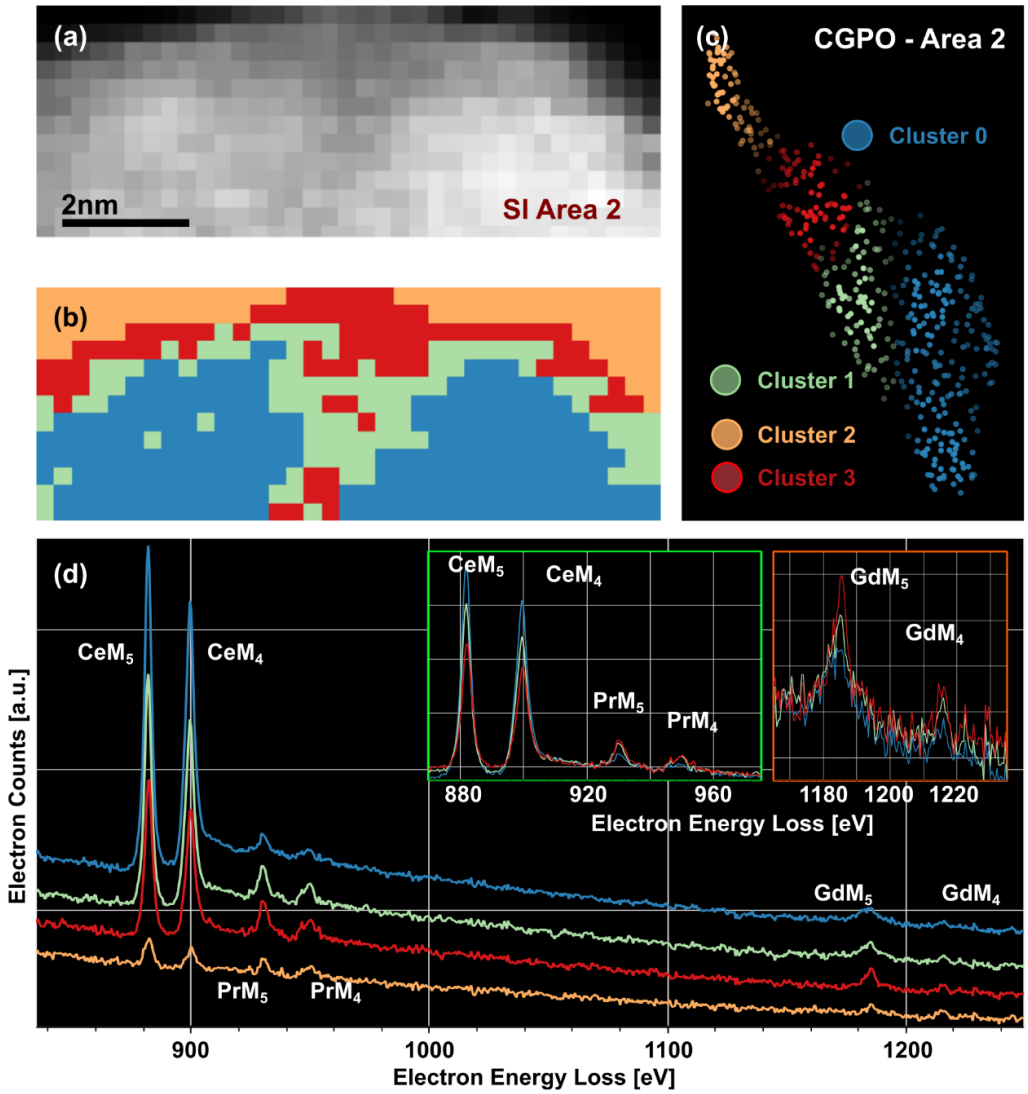


Figure 6.16. Co-acquired HAADF image of the ROI area number 2 for the CGPO EELS SI shown in **Figure 6.12**. (b) Label map resolved by HDBSCAN over the manifold projection computed by UMAP and included in (c). (d) Average EELS signal for each one of the clusters resolved by HDBSAN, staggered in the y-axis for visualization reasons. The green inset area shows a detailed view of the normalized average spectra in the energy loss range of the Ce and Pr edges onsets. The orange inset area shows a detailed view of the normalized average spectra in the energy loss range for the Gd edge onset.

6.3. CGPO. Quantitative analysis and electron tomography.

As the CGPO mesoporous structures were proven to be the samples from the set with the higher degree of stability under the electron beam, they were selected as the best candidates to further explore the possible localized changes in the Ce oxidation state and dopant concentrations. Hence, a second round of experiments was carried out focusing on the samples with this composition.

This time, the conventional EELS SI acquired in the outer corona parts of the CGPO mesoporous material were quantitatively analysed. The cerium oxidation state was tracked by two distinctive methodologies: one involving the CeM_5 energy loss position and a second one studying the M_5/M_4 WL ratio variations. The dopant segregation was measured quantitatively using the Egerton method for the elemental quantification through EELS data analysis[17]. However, the existing software solutions for this quantification process relied on the background removal on the onset of each one of the elements under analysis. Given the superposition of the Ce excitation to the continuum of states signal, and the onset and fine structure of the Pr edge, the background removal procedure was extremely unstable for the latter and produced a severe and systematic underestimation of this element in the quantification. Also, the CeM_{32} minor edge was suspected to be contributing to an overestimation of the quantity of Gd across the whole range of the SIs under analysis, as the M_3 line position falls precisely under the GdM_{54} edge onset. Providing a solution to these complicated quantification problems was the main driving force behind the development of ***WhatEELS***[18], as the previously available NLLS fitting solutions were too limited in their capabilities (as described in chapter 3 of this work).

In addition to the EELS analysis, a 3D X-EDS analytical ET reconstruction was carried out to resolve the 3D chemical structure for the whole volume of the material. The reconstruction was done following the TVM methodology

described in the first chapter of this work, as it allowed a precise reconstruction using a very limited number of projections (i.e., limiting the electron dose over the sample).

6.3.1. EELS-SI quantitative analysis for the CGPO.

The ROI for the EELS SI acquired in this second round of experiments for the CGPO material is shown in **Figure 6.17**. Notice how the acquisition ROI covers an extensive area of the sample corona, including a large number of overlapping grains and partially oriented GB with respect to the illumination direction. This larger dataset provided a better statistical ground for the posterior UMAP projection and clustering analysis. **Figure 6.17** (b) shows a high resolution HAADF image of the ROI from **Figure 6.17** (a). The FFT indexation for two adjacent grains separated by a GB (zones 1 **Figure 6.17** (c) and 2 in **Figure 6.17** (d)) confirmed the polycrystalline nature of these materials, where the CeO_2 grains filling the silica template are randomly

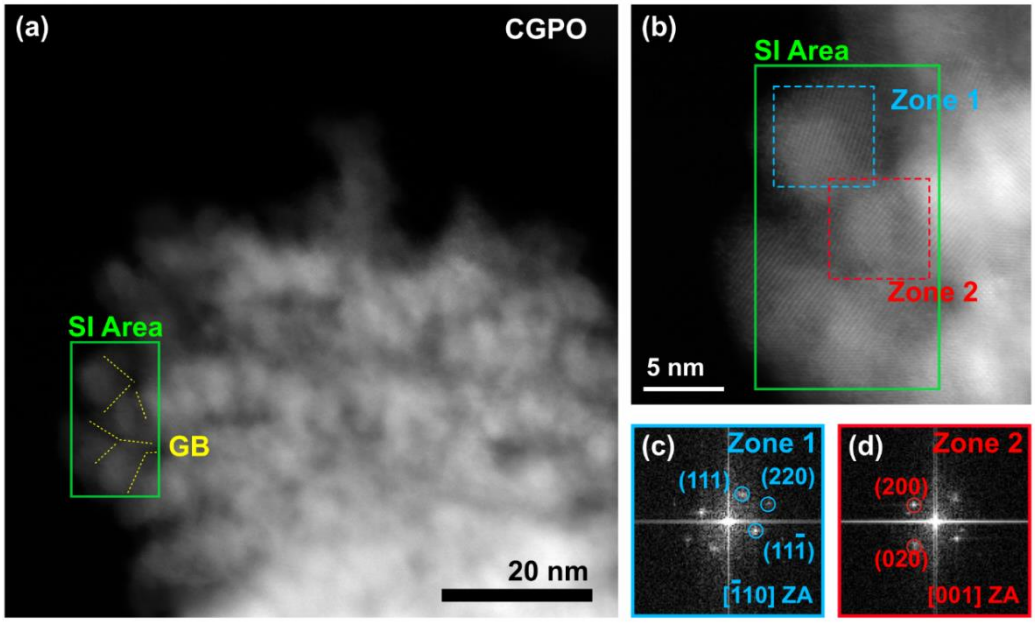


Figure 6.17. (a) HAADF image for the CGPO ROI. The yellow lines highlight some GBs included within the SI area. (b) High resolution HAADF image of the ROI for the EELS Si acquired. Two adjacent grains are indicated in a blue and red square ROI, and their indexed FFT results are shown in (c) and (d) respectively.

oriented within the larger mesoporous structure during and after the synthesis process (instead of forming a single large crystal).

The quantitative results were obtained using the **WhatEELS** software solution, that combined a clustering segmentation analysis (to separate the spectrally similar regions), and a NLLS fitting routine to resolve the elemental distribution in the SI. As the **WhatEELS** NLLS fitting required a sample with the background signal removed preceding the element with the lower energy onset value, a PCA denoising procedure that included a large number of components (way below the knee of the variance ratio curve) was carried out on before the background removal. The reason to maintain a large number of the components initially separated by PCA, otherwise typically ascribed to noise in the SI, was to avoid the potential bias induced by handpicking a selection of them without considering the possible noise dilution of relevant information from the original EELS dataset (i.e., to avoid some of the known problems for EELS SI denoising using a PCA decomposition[19]). On the other hand, a reasonable balance between the noisy component kept and a successful background removal had to be considered.

The objective of the combination of clustering and NLLS in **WhatEELS** is to provide an identical departure point for the model parameters in each one of the pixels belonging to a common cluster, computed from the average cluster signal (i.e., the natural centroid). Therefore, the clustering analysis was carried out over the same background-removed SI. Also, the manifold projection computation by UMAP was carried out selecting a Euclidean metric measurement for the distances between the datapoints (i.e., single pixel spectra) in their native hyperdimensional space, in order to keep the information about the thickness effects for the fitting procedure (majorly codified by the L_2-norm value for the spectra, see chapter 3 of this work). This metric will be maintained for the rest of the UMAP computations in this chapter, as every single clustering segmentation is intended to be used in conjunction with the NLLS fitting in WhatEELS and, therefore, the inclusion of the thickness effects in the segmentation is important.

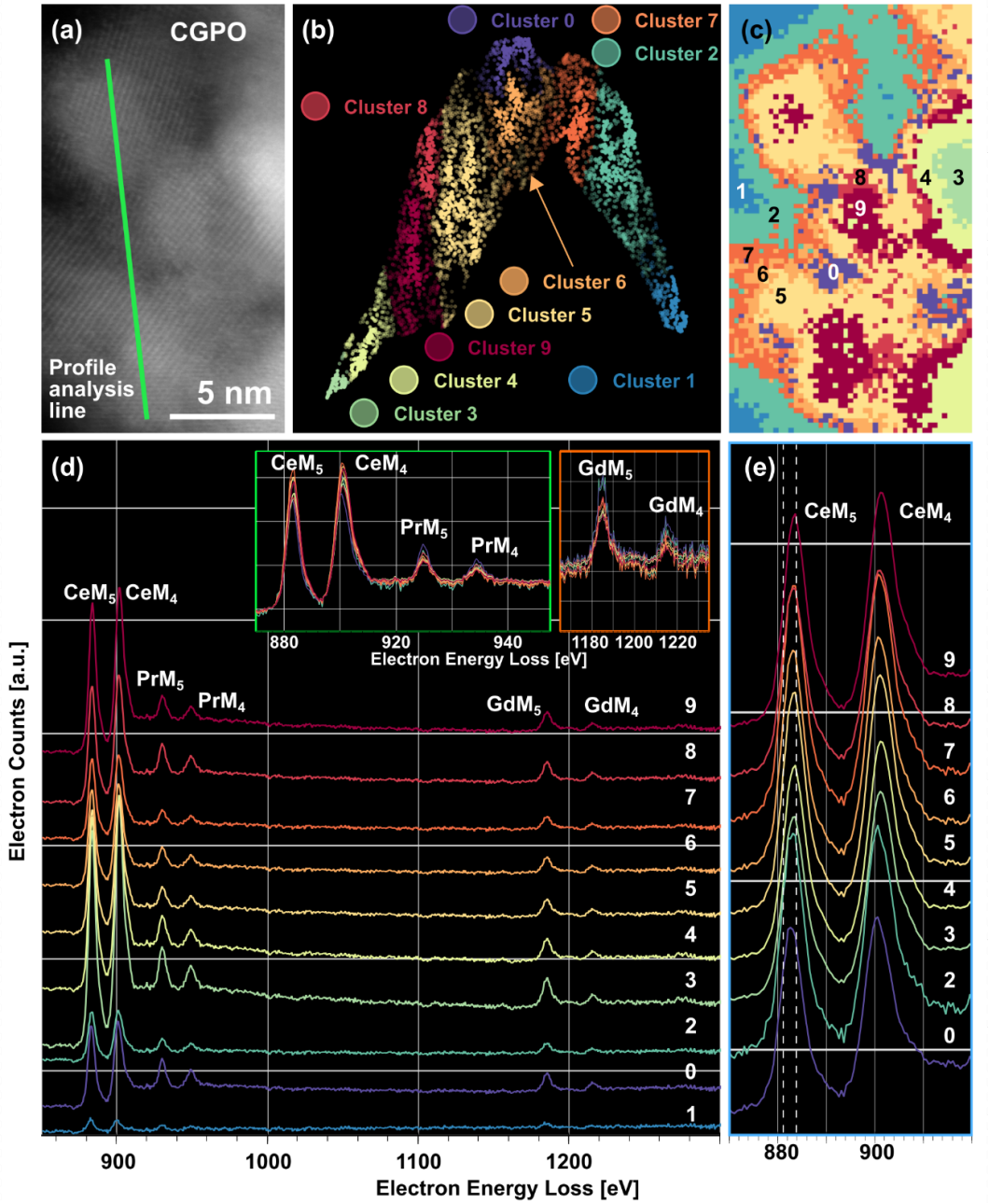


Figure 6.18. High resolution HAADF image of the ROI area where the CGPO EELS SI was acquired (**Figure 6.17**). (b) UMAP projected manifold for the EELS dataset, coloured according to the clustering results obtained by HDBSCAN and shown in (c). (d) Average EELS signal for each one of the clusters resolved by HDBSCAN, staggered in the y-axis for visualization reasons. The green inset area shows a detailed view of the normalized average spectra in the energy loss range of the Ce and Pr edges onsets. The orange inset area shows a detailed view of the normalized average spectra in the energy loss range for the Gd edge onset. (e) Staggered version for the detailed view of the normalized average signals per cluster in the energy loss region of the Ce edge.

The clustering results for this CGPO dataset are presented in **Figure 6.18**. The line coloured in green over the HAADF high resolution image in **Figure 6.18** (a) highlights the path used later for the profiling analysis of the quantification results for the Ce oxidation state and Pr-Gd dopant composition. The number of natural clusters discovered by HDBSCAN (**Figure 6.18** (c)) over the UMAP projection map (**Figure 6.18** (b)) is larger than in the previous examples. This number could be reduced by fine tuning the hyperparameters in both algorithms, but as it stands it was deemed as a favourable case scenario for the **WhatEELS** NLLS fitting procedure. Already, the qualitative analysis of the average (and normalized average) signals per cluster (**Figure 6.18** (d), (e) and detailed insets) hints a plausible dopant segregation towards localized regions in the GBs and surfaces, through an increment in the relative intensity values of the Pr and Gd WLs with respect to the Ce WLs for clusters 0, 2 and 8. Also, a change in the Ce oxidation state towards GB and GS (clusters 0 and 2) becomes apparent from the analysis of the staggered normalized signals in the CeM_{54} edge energy loss region in **Figure 6.18** (e), as precisely the Ce M_5 lines in those clusters show a displacement of their central positions towards lower values of electron energy losses.

The line ratios and the elemental quantification are shown in **Figure 6.19** (a quantitative results mappings extracted from **WhatEELS** for the CeM_5 central position, the Ce M_5/M_4 white), (c), (e) and (f). The profiles for these measurements along the line highlighted in green in **Figure 6.18** (a) are shown in **Figure 6.19** (b), (d) and (g).

The fitted results for the Ce M_5 WL centre show a displacement to lower energy loss values consistent with a reduction of Ce over the GB regions and towards the GS (**Figure 6.18** (a)). This result is consistent with the observed Ce WL ratio variations (**Figure 6.18** (c)). The increment of the M_5/M_4 ratio has been related in the literature with a change towards a Ce^{3+} oxidation state from the nominal Ce^{4+} (mainly observed in this sample in the bulk areas of the grains). The ratio values calculated from the fitted components integrated intensities

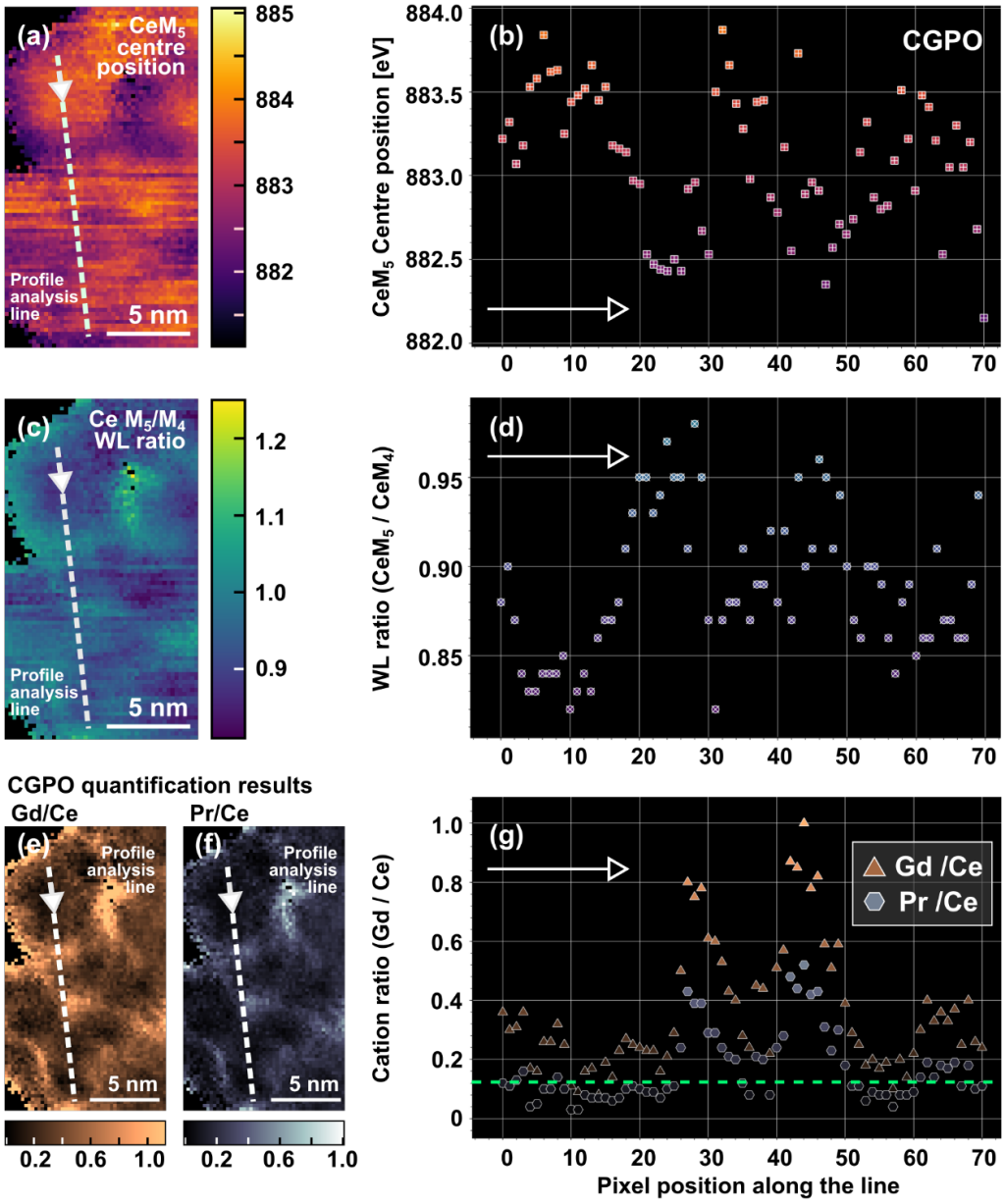


Figure 6.19. Quantitative results for the CGPO mesoporous material (a) Mapping of the centre position energy loss value for the CeM_5 WL component resolved by **WhatEELS**. (b) Scatter plot for the centre positions of the CeM_5 WL along the pixels covered by the dashed line in (a).(c) Mapping of the Ce M_5/M_4 WL ratio variations from the component resolved by **WhatEELS**. (d) Scatter plot for the Ce M_5/M_4 WL ratio variations along the dashed line highlighted in (c). (e)(f) Gd/Ce and Pr/Ce ratio quantification mappings extracted from **WhatEELS**. (g) Gd/Ce and Pr/Ce ratio quantification for the pixels along the lines highlighted in (e-f).

resolved by WhatEELS (and showcased in **Figure 6.19** (c)) are consistent with those reported in the literature [20–22].

Regarding the quantification measurements, a clear segregation of both dopants (Pr and Gd) towards GB and GS is observable in the ratio mappings presented in **Figure 6.19** (e-f).

The line profiles provided an extra layer of analysis to the results. The line included in **Figure 6.18** and **Figure 6.19** clearly traverses 2 different GBs, departing from the surface of a CeO_2 grain and finishing in the surface of a different grain at the end of the ROI area. The local minima for the CeM_5 centre positions, the maxima for the M_{54} WL ratio and the maximum concentration of both the Pr and Gd dopants fall at the same values of pixel position, corresponding to the 2 GBs traversed by the line. This would indicate a coordinated Ce reduction and dopant segregation towards GBs. These quantities also show values consistent with this same dopant segregation and Ce reduction behaviours at beginning and end of the line (i.e., towards the grain surfaces traversed). The green line in **Figure 6.19** (g) marks the stoichiometric ratios expected for the CGPO material. Hence, a good agreement is observed for the bulk regions, and an excess of dopants in the GB.

Overall, these quantitative results confirm what was already established during the qualitative analysis of the CGPO. That is, dopant segregation and Ce reduction towards GB and GS.

6.3.2. ADF and X-EDS electron tomography.

A series of electron tomography experiments for the CGPO sample were carried out in an attempt to generalize the study of the dopant segregation to GB and GS to the whole 3D volume of the mesoporous structure. Given the 3D morphology of these mesoporous materials, the conventional TEM EELS analysis was impossible beyond the outer-corona regions of the structures. In a nutshell, the intricated channel structure of the mesoporous materials made impossible the acquisition of spectral datasets with identifiable GB structures by conventional means (i.e., planar EELS) in the inner-most parts of the

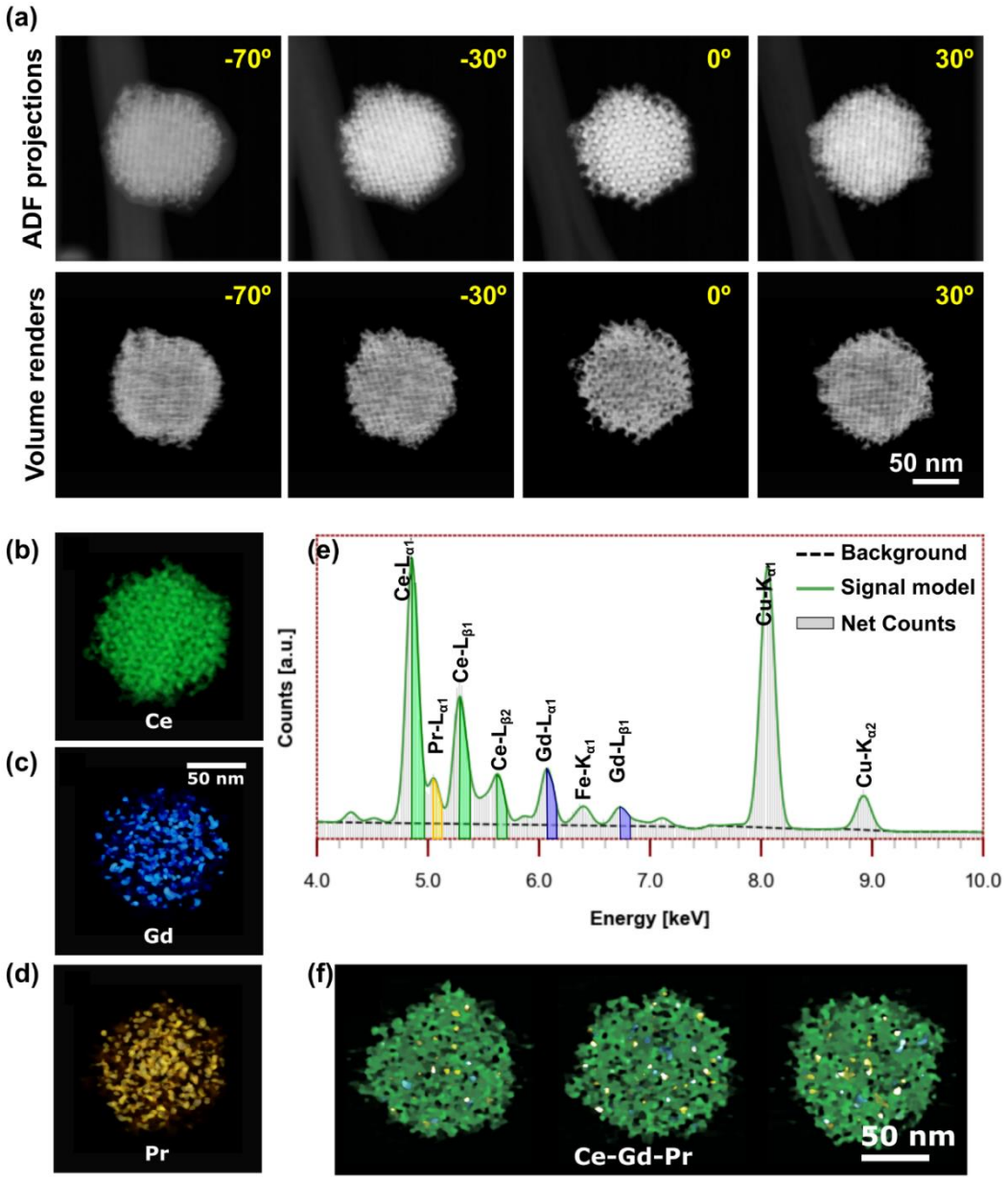


Figure 6.20. (a) ADF projections acquired and projected views of the volume renders for the ADF tomography reconstructions at 4 different angles. (b-d) Volume renderings for the analytical reconstructions of the Ce, Gd and Pr signals, recovered from the X-EDS models showcased in (e). (f) Overlays of 3 different XY-slices through the Ce, Gd and Pr elemental volumes from (b-d). The colours are consistent between panels.

mesoporous structures. Furthermore, as the average size of the mesoporous structures ranged between 150 and 200 nm the X-EDS analytical acquisition mode was the optimum choice for an analytical-ET experiment. This technique avoids the absorption effects and multiple scattering events related with EELS spectroscopy on thick materials, that would lead to capping artefacts in the ET reconstructions. As a side effect, the lateral resolution was decreased, but a qualitative analysis of the chemical composition in 3D was still possible.

The results shown in **Figure 6.20** correspond to a dual morphological and analytical ET reconstruction carried out for a single mesoporous structure of CGPO. The beam current was set to 105 pA to increase the SNR of the X-EDS mappings extracted, whereas the accelerating voltage was set as 80 kV to increase the scattering cross section and ensure a higher number of collisions. As consequence, to minimize the final total electron dose transferred to the sample during the acquisition of the whole set of X-EDS SI, the number of projections was restricted to 9. The angular range for this experiment was ±70º, with a 20º step and including the 0º projection. This small number of projections was conditioned by the large times required to register a complete X-EDS SI for each one of the projection angles acquired. Nevertheless, the panel in **Figure 6.20** (a) showcases (qualitatively) the high accuracy level of the ADF reconstruction recovered by the TVM methodology (described in the first chapter of this work). An excellent agreement between the ADF reconstruction volume projected at different angles and the original ADF acquired projections used for the ET reconstruction can be observed in this panel, despite the severe angular undersampling for the experiment. The penalty parameter μ for the TVAL3 algorithm used in the reconstruction of the ADF signal was set as 2^7, in order to promote the reconstruction of fine details[21].

The panels in **Figure 6.20** (b-d) show the elemental volume renders reconstructed from the acquired X-EDS SI. The projections for each one of the elements were the elemental mappings extracted from the modelled signal of

[21] More information on importance of setting the right values for the penalty parameters can be found in chapter 1.

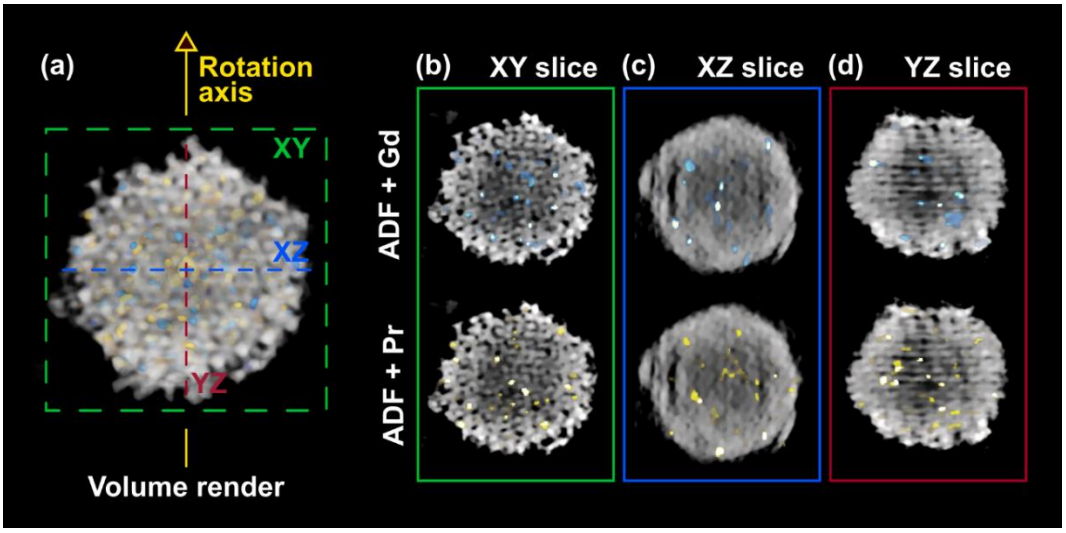


Figure 6.21. (a) Volume rendering representation overlaying the ADF morphological ET reconstruction and the elemental Pr and Gd analytical ET reconstructions. The slices through the volumes reconstructed, overlaying the ADF signal and the Gd (blue) and Pr (yellow) signals, are shown for the XY planes in (b), the XZ planes in (c) and the YZ planes in (d).

the X-EDS SI. An example of such model is shown in **Figure 6.20** (e), where the main lines used to calculate the mappings for each one of the elements is highlighted with the same colour scheme of the volume renders. For these analytical reconstructions, a penalty parameter μ of 2^4 was selected to mitigate the undersampling effect of the noisy mapping conditions. Again, more information on importance of setting the right values for the penalty parameters can be found in chapter 1 of this work.

Finally, some slices of XY planes (i.e., the family of planes parallel to the 0º projection) through the element volumes reconstructed are shown in **Figure 6.20** (e). From them, it becomes apparent that the Ce follows the general structure recovered by the ADF reconstruction (i.e., it the Ce reconstruction retains the morphology of the D-gyroid mesoporous structure). Meanwhile, the Gd and Pr tend to form clusters of higher concentration in between the nodal junctions for the CeO_2 grains that form the D-gyroid of the mesoporous structure.

The generalized segregation of dopants (i.e., formation of higher concentration regions) towards junctions in the mesoporous structure is further explored in **Figure 6.21**. The different slices through the overlay of ADF and Pr-Gd signal volume reconstructions (XY planes in **Figure 6.21** (b), XZ planes in **Figure 6.21** (c), and YZ planes in **Figure 6.21** (d)) confirm what was already demonstrated through the analysis of **Figure 6.20** (i.e., the dopant segregation).

Overall, the lateral resolution of the reconstructions is much lower than that of the EELS analysis from the previous subsections. However, the ET experiments allowed the generalization of the dopant segregation hypothesis to the whole 3D volume of the material through a qualitative analysis of the obtained results. As bonus result, the mesoporous structure was also proven to retain the D-gyroidal inner structure of channels from the KIT-6 silica template through the analysis of the slices of the ADF morphological reconstruction.

6.4. CGO and CGO(HF). Quantitative analysis.

The last section of this chapter is devoted to the quantitative analysis of a second round of EELS experiments conducted for the CGO mesoporous structures. It also includes the results for the quantitative analysis of the EELS experiments for the Co decorated - HF post-processed CGO.

6.4.1. Quantitative analysis of the CGO sample.

The initial qualitative analysis for the CGO samples revealed a tendency towards a beam-induced cation reduction during the long exposure times for the acquisition of EELS SI. As such, the study of localized changes in the oxidation state of the Ce in the Gd-doped CeO_2 mesoporous materials was hindered.

The second run to test these effects begun with the acquisition of several EELS time series for different illumination conditions. A time series experiment is

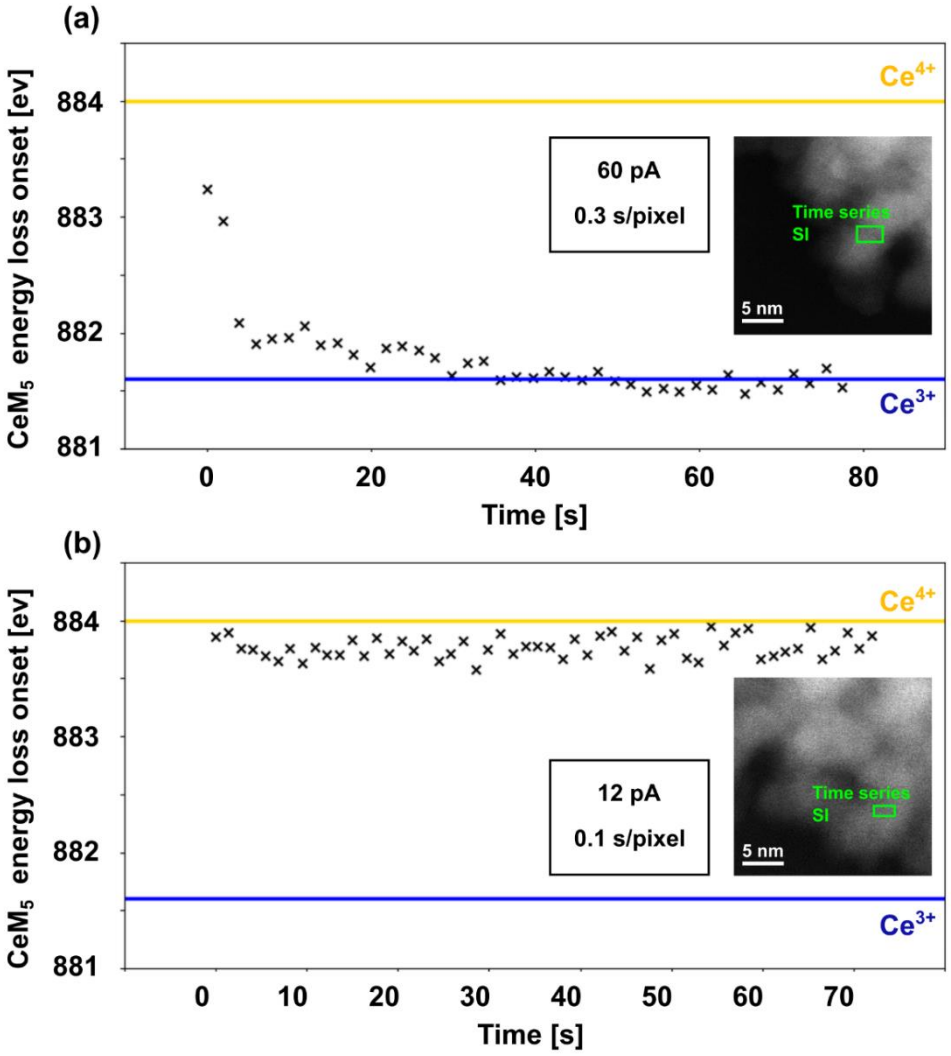


Figure 6.22 Scatter plots for the CeM_5 WL onset position analysed during a time series acquisition. (a) Results for the case with Ce reduction under the electron beam. (b) Results for the case without Ce reduction under the electron beam.

an acquisition mode that consists of a continuous registration of EELS spectra with a fixed acquisition time (in this case, 0.1 and 0.3 spectra per second, plus a small flying time between them controlled by the acquisition software) during a total period of time undefined (i.e., the acquisition was manually stopped after a large number of spectra were acquired, typically after a couple of minutes of continuous registration at fixed time intervals). The spectra are acquired in random positions inside a ROI area delimited by the user over the sample. The results for two of these experiments carried out for the CGO

material are shown in **Figure 6.22**. Both ROI areas (inset of the scatter plots) were placed above the bulk regions of granular structures in the outer corona parts of the CGO mesoporous material, and dangling over the vacuum of a grid hole. The electron current conditions on the first experiment (**Figure 6.22** (a)) prompted a fast degradation of the material, which is translated to the clear Ce reduction process shown in the scatter plot through the rapid displacement of the CeM5 onset towards lower energy loss values. Meanwhile, the second set of conditions over a similar area in the sample (with a much lower current intensity and register time per pixel, see (**Figure 6.22** (b)) showed clear signs of stability in the oxidation state of the Ce cation (Ce^{4+}) for the whole duration of the experiment. Hence, the following EELS SI acquired under this illumination conditions were not conditioned by the electron beam damage artefacts previously observed for samples of this same composition.

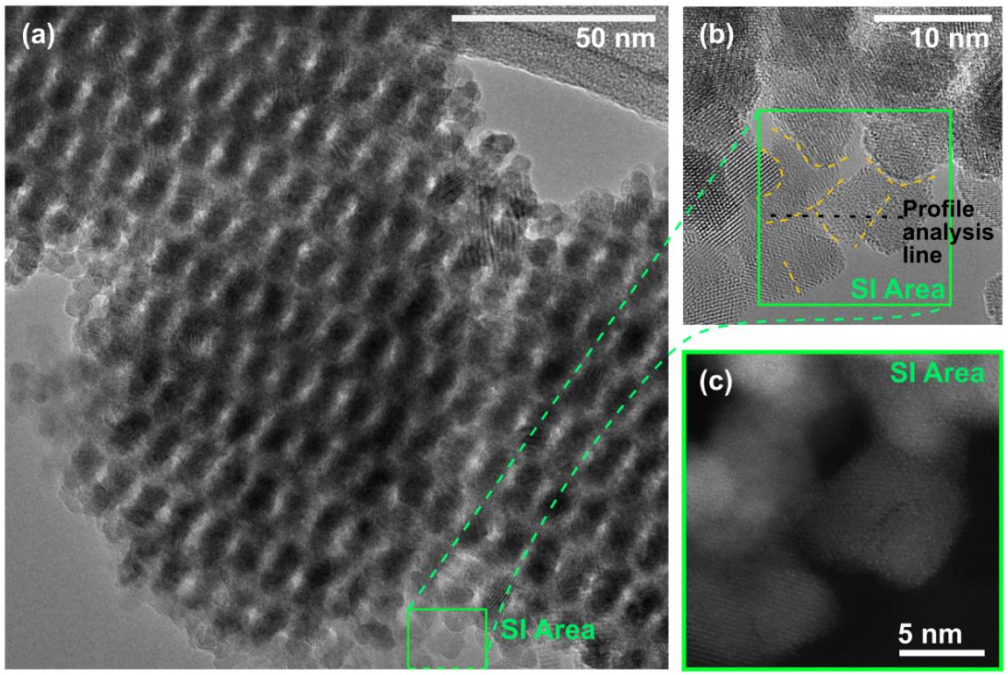


Figure 6.23. (a) Low magnification HRTEM image of the mesoporous structures of CGO where the ROI area for the EELS SI acquisition was located. (b) Zoomed HRTEM image of the actual ROI area. The lines in yellow are guidelines marking several GBs. (c) High resolution HAADF image of the ROI area for the EELS SI.

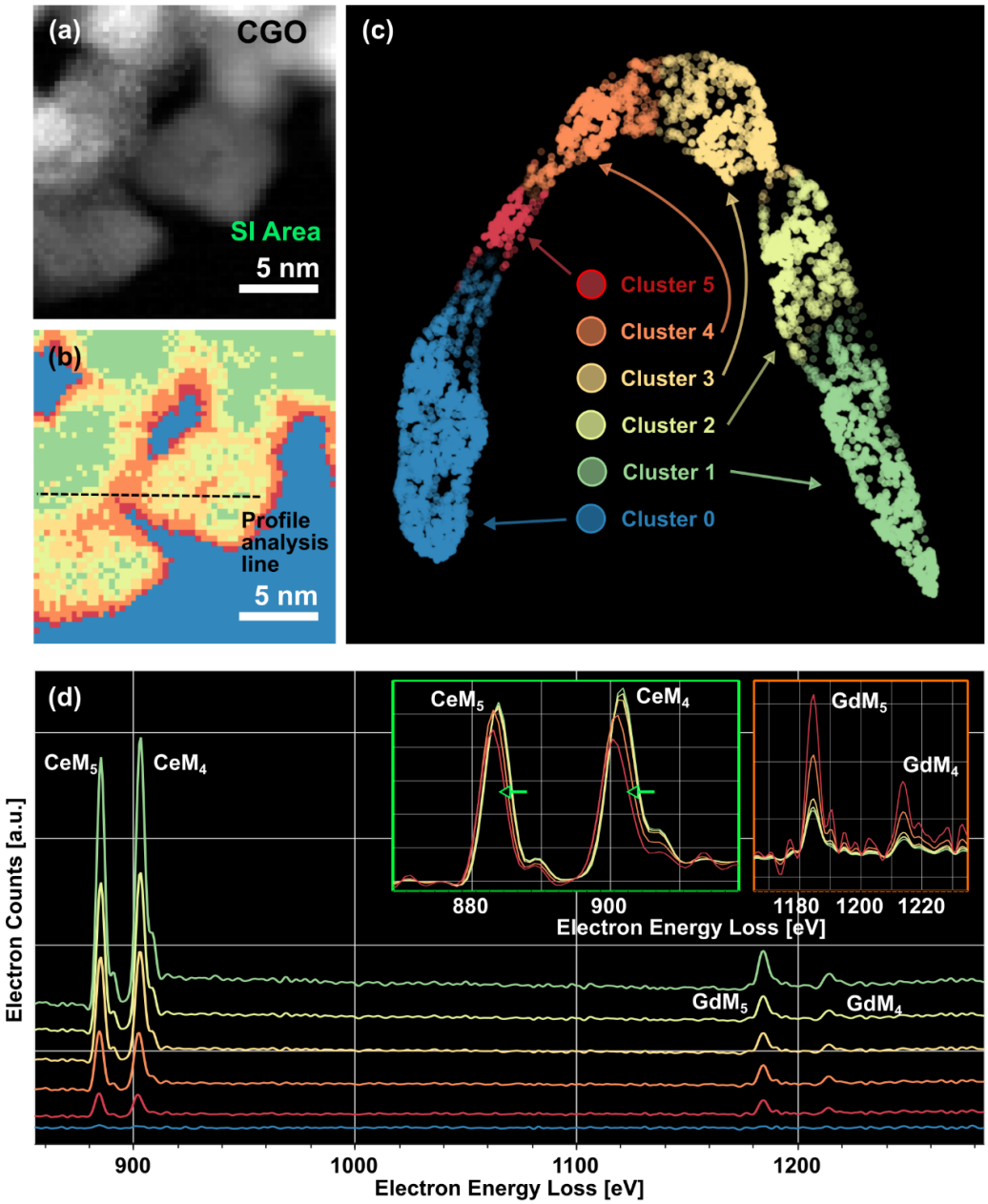


Figure 6.24. (a) Co-acquired HAADF image for the EELS SI of the CGO mesoporous material. (b) HDBSCAN clustering map resolved for the UMAP projection shown in (c). (d) Staggered average signals for each one of the clusters in (b). The normalized average signals for the energy loss ranges in the proximity of the Ce and Gd edges are shown in the green and orange inset graphs, respectively.

The ROI area for the acquired EELS SI is shown in **Figure 6.23**. Much like in the second experiment for the CGPO (previous section) a large ROI area in a region with an intricate structure of grains and several GB was selected.

The initial clustering result for the EELS SI are shown in **Figure 6.24**. As the raw spectra presented a fairly low sNR (remember that the illumination conditions were tuned down to avoid sample damage), the same PCA denoising process explained in the previous section for the CGPO material was conducted in this case. The background signal was also removed previous to the clustering segmentation, given that the final intention was to carry out a quantitative analysis using the **WhatEELS** software solution. This time, an extra step involving a Fourier ratio deconvolution is included as well, given that the spectra were acquired in a dual EELS mode (the low-loss and core-loss regions of the spectra were registered at the same time).

UMAP produced a series of well separated clusters of datapoints (i.e., the pixels in the EEL SI), that were easily identified by HDBSCAN as the natural clusters from the projection map **Figure 6.24** (c). The translation of these clusters back to the image space produced the cluster segmentation image shown in **Figure 6.24** (b). The qualitative analysis of the average signals for each one of the clusters resolved (**Figure 6.24** (d)) already identifies the clear Ce reduction towards GBs and GSs (see the changes in the CeM_{54} WLs in the normalized average signals in the green inset area). This is concomitantly accompanied by a segregation of Gd (observe the clear increment of the relative intensity of the GdM_{54} WLs in the clusters containing the GBs and surfaces shown in the orange inset).

The quantitative analysis results shown in **Figure 6.25** were extracted from the NLLS routine in WhatEELS, using the clustering classification showcased in **Figure 6.24**. The mappings for the Ce M_5 centre position (**Figure 6.25** (a)) and the Ce M_5/M_4 WL ratio (**Figure 6.25** (c)) are consistent with a Ce reduction towards the GSs and GBs. The quantification analysis in **Figure 6.25** (e) indicates a clear segregation of Gd towards these same GB and

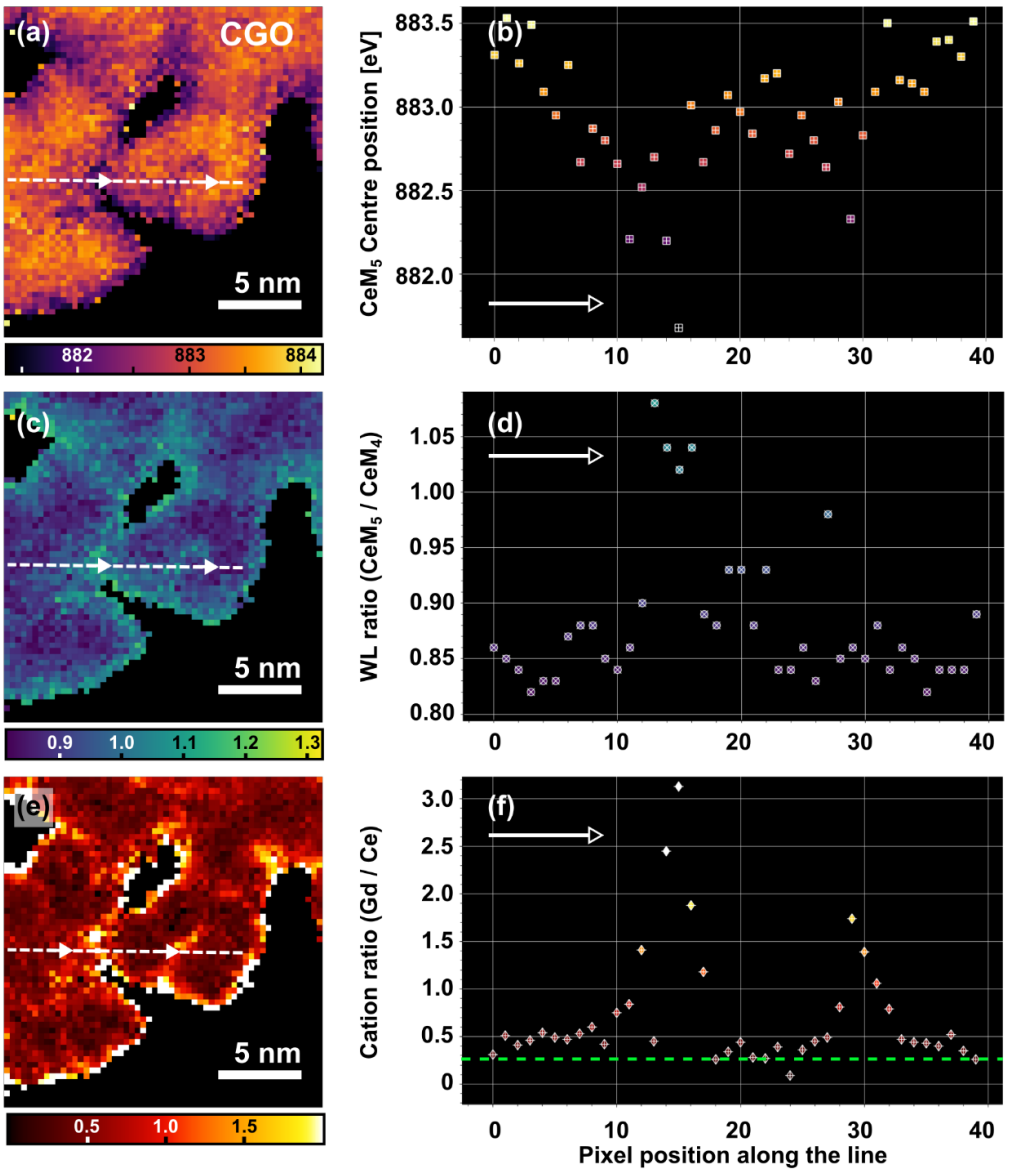


Figure 6.25. Quantitative results for the CGO mesoporous material (a) Mapping of the centre position energy loss value for the CeM_5 WL component resolved by **WhatEELS**. (b) Scatter plot for the centre positions of the CeM_5 WL along the pixels covered by the dashed line in (a).(c) Mapping of the Ce M_5/M_4 WL ratio variations from the component resolved by **WhatEELS**. (d) Scatter plot for the Ce M_5/M_4 WL ratio variations along the dashed line highlighted in (c). (e) Gd/Ce ratio quantification mapping extracted from **WhatEELS**. (f) Gd/Ce ratio quantification for the pixels along the line highlighted in (e).

GS. The profile lines for these 3 quantities further confirm this hypothesis of Ce reduction and Gd segregation towards GBs and GSs. Furthermore, the quantification scatter plot in **Figure 6.25** (f) indicates that the Gd/Ce is very close to the stoichiometric values provided (green line) for the regions traversing the bulk of the grains.

6.4.2. Quantitative analysis of the CGO(HF) sample.

In this final section the results for the last sample analysed for these set of mesoporous materials are presented. It corresponds to the CGO post-process with an etching bath in HF and decorated with atomic Co in its surface.

The focus in this sample was set in testing the hypothesis that the HF bath effectively removed the excess of Gd dopant segregation towards the GS of the

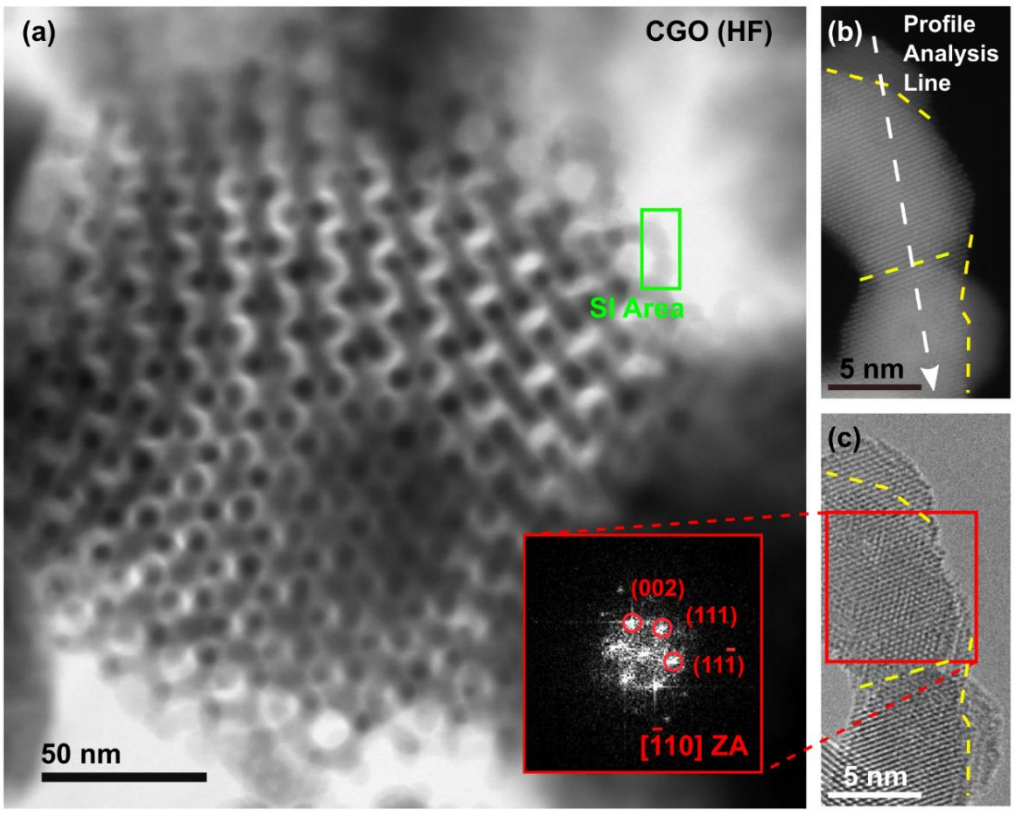


Figure 6.26. (a) Low magnification BF image of the CGO(HF) mesoporous material where the ROI area for the EELS SI acquisition was located. Zoomed high resolution HAADF image (b) and (c) HRTEM images of the actual ROI area for the EELS acquisition. The dashed lines in yellow are visual guidelines marking several GBs in the field of view.

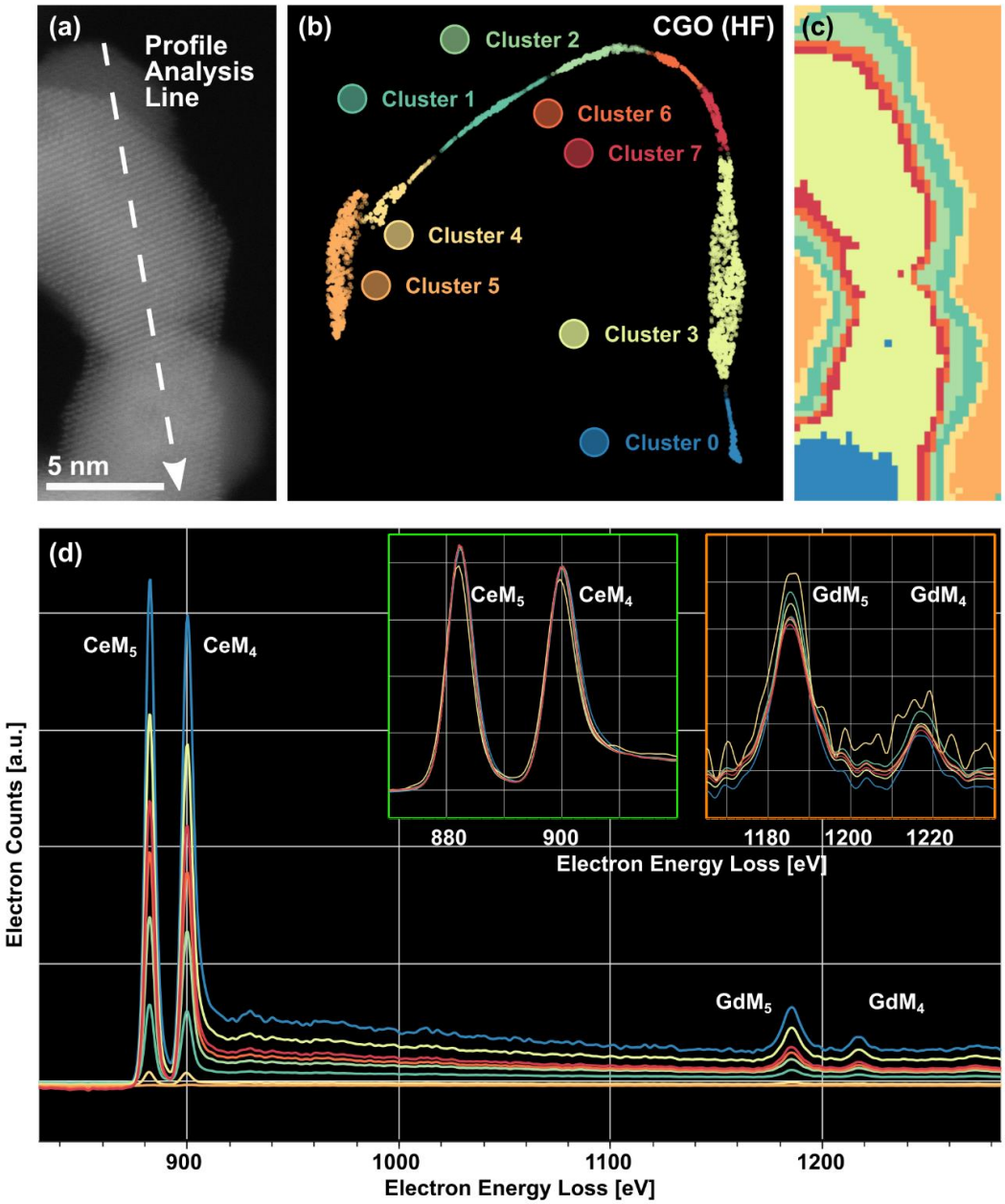


Figure 6.27. (a) Co-acquired HAADF image for the EELS SI of the CGO(HF) mesoporous material. (c) HDBSCAN clustering map resolved for the UMAP projection shown in (b). (d) Average signals for each one of the clusters in (c). The normalized average signals for the energy loss ranges in the proximity of the Ce and Gd edges are shown in the green and orange inset graphs, respectively.

CGO mesoporous material. Also, the initial intention was to resolve the position of the Co decoration within the mesoporous structure of the material. There was an interest in knowing if the Co segregated forming high concentration clusters, as it was the case for the Gd and Pr dopants in the other samples of this set.

Unfortunately, the percentage of Co added to the samples fell below the detection limits of the EELS experiment, even after raising the beam current again up to 110pA. This increment of the electron current intensity hindered the possibility of studying the Ce oxidation state changes in the material, as the beam is expected to reduce the Ce of the sample under these illumination conditions, as previously evidenced.

The ROI area for the EELS SI acquisition is shown in **Figure 6.26**. The clustering results for this dataset are shown in **Figure 6.27**. Once again, the segmentation map (label-map) resolved by HDBSCAN (**Figure 6.27** (c)) on the UMAP projection (**Figure 6.27** (b)) is used by WhatEELS to perform the NLLS fitting and retrieve the quantitative results.

The qualitative analysis of the average signals for the clusters resolved by HDBSCAN (**Figure 6.27** (d)) shows a steady signal increment towards thicker areas of the sample. However, no signs of changes in the Ce oxidation state or significant Gd dopant segregation clusters appear on this sample. The former is probably a consequence of the sample reduction under the beam conditions set for the experiment, as already mentioned before. The latter would be an initial qualitative proof that hypothesis of the surface cleansing of excessive Gd dopant cluster formation by the HF etching process was correct. A slight increment of the GdM_{54} with respect to the CeM_{54} WLs can be observed for the cluster 4 (outermost part of the grains, or GSs) in the normalized spectra shown in the inset of **Figure 6.27** (d). Notwithstanding, this effect is small compared to the GdM_{54} increment observed for the other samples.

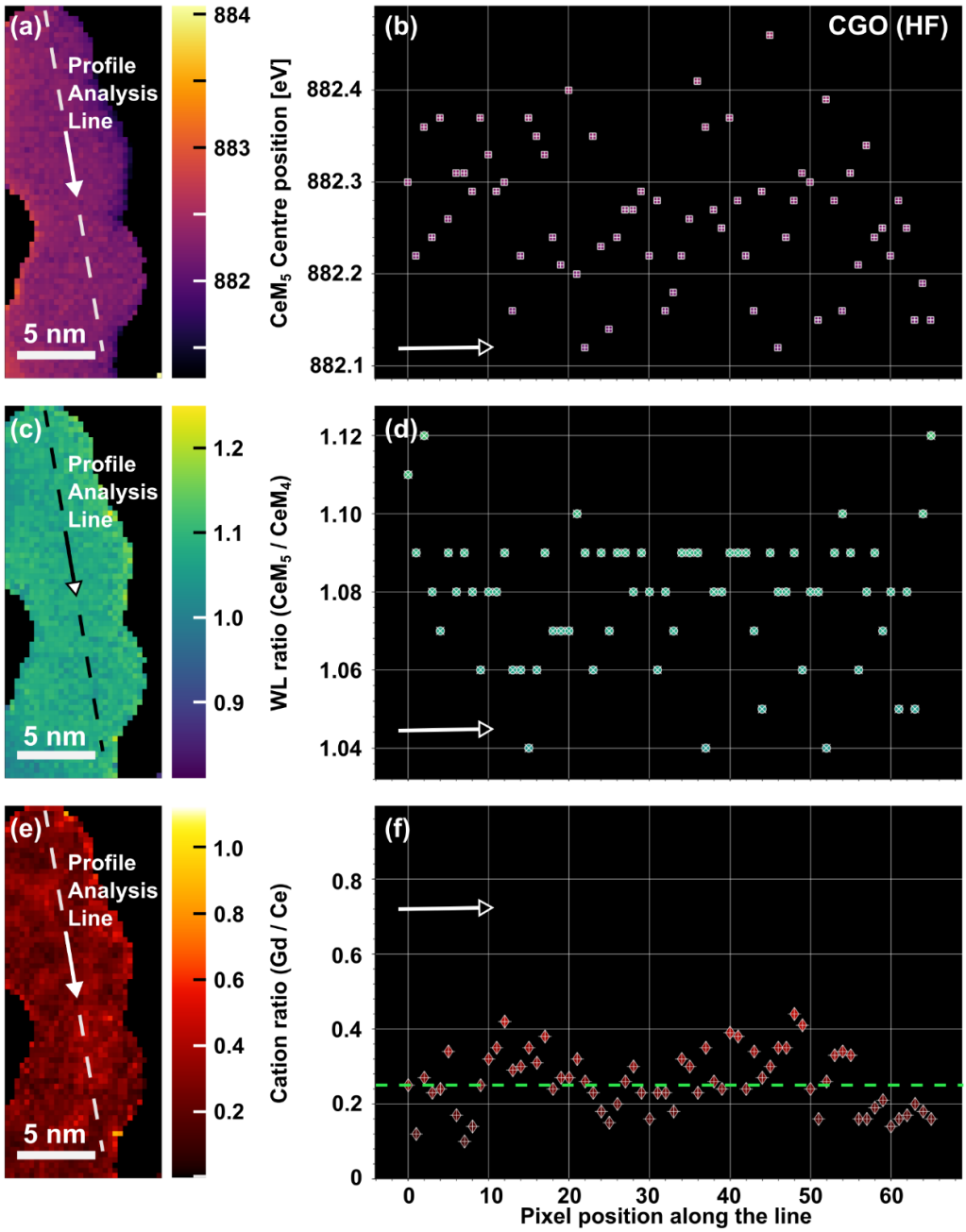


Figure 6.28. Quantitative results for the CGO (etched with HF) mesoporous material (a) Mapping of the centre position energy loss value for the CeM_5 WL component resolved by **WhatEELS**. (b) Scatter plot for the centre positions of the CeM_5 WL along the pixels covered by the dashed line in (a).(c) Mapping of the Ce M_5/M_4 WL ratio variations from the component resolved by **WhatEELS**. (d) Scatter plot for the Ce M_5/M_4 WL ratio variations along the dashed line highlighted in (c). (e) Gd/Ce ratio quantification mapping extracted from **WhatEELS**. (f) Gd/Ce ratio quantification for the pixels along the line highlighted in (e).

To clarify this last point, WhatEELS was used to carry out the Ce/Gd quantification and check the complete reduction of the Ce cation across the whole SI surface.

These quantitative results are shown in **Figure 6.28**. The CeM_5 centre positions and M5/M4 WL ratio (along with the profiles shown crossing a GB in **Figure 6.28** (b) and (d)) show a total Ce3+ reduction in the whole area of the SI. The Gd/Ce ratio quantification results are consistent with the hypothesis of an overall homogeneous distribution of Gd dopant (**Figure 6.28** (e)), close to the stoichiometric ratio of the formula, and holding in the event of traversing a GB (**Figure 6.28** (f)).

6.5. Conclusions.

The analysis presented in this chapter for the set of doped CeO_2 mesoporous materials used the whole range of characterization tools described in the previous chapters of this work.

As a matter of fact, the completion of this set of experiments, and the lessons learned during the characterization of these materials, were the major driving forces behind the development of the **WhatEELS** software solution and the exploration of new routes for the inclusion of advanced clustering techniques for EELS data analysis.

Overall, the results obtained for each one of the materials under analysis successfully prove the hypothesized dopant segregation towards grain boundaries and grain surfaces (both Pr and Gd). Also, the cerium reduction from Ce^{4+} to Ce^{3+} towards the same regions in the mesoporous structures (GBs and GSs) was successfully demonstrated for the CGPO and the CGO materials.

Finally, a 3D reconstruction of the mesoporous structure using the advanced TVM methodology for electron tomography was also successfully obtained. Besides the morphological reconstruction, the analytical counterpart (i.e., analytical ET) further demonstrated the tendency of the dopant elements (Gd

and Pr) to form clusters of higher concentrations located in the junctions between grains across the whole volume of the mesoporous CGPO.

These TEM characterization results played a major role in a larger study of the electrochemical performance for devices synthesized with these types of doped mesoporous ceramic materials. They were a key element in the explanation of charge exchange (ionic and electronic) physical processes at the nanoscale, that ultimately governed the macroscopically measured properties (e.g., conductivity)[20].

6.6. References.

[1] X. Deng, K. Chen, H. Tüysüz, Protocol for the Nanocasting Method: Preparation of Ordered Mesoporous Metal Oxides, Chem. Mater. 29 (2017) 40–52. https://doi.org/10.1021/acs.chemmater.6b02645.

[2] W.J. Bowman, J. Zhu, R. Sharma, P.A. Crozier, Electrical conductivity and grain boundary composition of Gd-doped and Gd/Pr co-doped ceria, Solid State Ionics. 272 (2015) 9–17. https://doi.org/10.1016/j.ssi.2014.12.006.

[3] S.R. Bishop, T.S. Stefanik, H.L. Tuller, Electrical conductivity and defect equilibria of Pr0.1Ce 0.9O2-δ, Phys. Chem. Chem. Phys. 13 (2011) 10165–10173. https://doi.org/10.1039/c0cp02920c.

[4] T. Holstein, Studies of polaron motion: Part I. The molecular-crystal model, Ann. Phys. (N. Y). 8 (1959) 325–342. https://doi.org/https://doi.org/10.1016/0003-4916(59)90002-8.

[5] L. Friedman, T. Holstein, Studies of polaron motion: Part III: The Hall mobility of the small polaron, Ann. Phys. (N. Y). 21 (1963) 494–549. https://doi.org/10.1016/0003-4916(63)90130-1.

[6] H.L. Tuller, A.S. Nowick, Small polaron electron transport in reduced CeO2 single crystals, J. Phys. Chem. Solids. 38 (1977) 859–867. https://doi.org/10.1016/0022-3697(77)90124-X.

[7] J.J. Plata, A.M. Márquez, J.F. Sanz, Electron mobility via polaron hopping in bulk ceria: A first-principles study, J. Phys. Chem. C. 117 (2013) 14502–14509. https://doi.org/10.1021/jp402594x.

[8] E.P. Francisco de la Peña Vidar Tonaas Fauske, Pierre Burdet, Petras Jokubauskas, Magnus Nord, … Andreas Garmannslund, hyperspy/hyperspy: HyperSpy v1.5.2, (2019). https://doi.org/10.5281/zenodo.3396791.

[9] P. Torruella, M. Estrader, A. López-Ortega, M.D. Baró, M. Varela, F. Peiró, S. Estradé, Clustering analysis strategies for electron energy loss spectroscopy (EELS), Ultramicroscopy. 185 (2018) 42–48. https://doi.org/10.1016/J.ULTRAMIC.2017.11.010.

[10] L. Mcinnes, J. Healy, J. Melville, UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 2020. https://arxiv.org/abs/1802.03426v3.

[11] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering, J. Open Source Softw. 2 (2017) 205. https://doi.org/10.21105/joss.00205.

[12] J. Blanco-portals, F. Peiró, S. Estradé, Strategies for EELS Data Analysis . Introducing UMAP and HDBSCAN for Dimensionality Reduction and Clustering, Microsc. Microanal. (2021) 1–14. https://doi.org/10.1017/S1431927621013696.

[13] M. López-Haro, M. Tinoco, S. Fernández-Garcia, X. Chen, A.B. Hungria, M.Á. Cauqui, J.J. Calvino, A Macroscopically Relevant 3D-Metrology Approach for Nanocatalysis Research, Part. Part. Syst. Charact. 35 (2018) 1700343. https://doi.org/10.1002/ppsc.201700343.

[14] J.M. Muñoz-ocaña, A. Bouziane, F. Sakina, R.T. Baker, A.B. Hungría, J.J. Calvino, A.M. Rodríguez-chía, M. López-haro, Optimization of STEM-HAADF Electron Tomography Reconstructions by Parameter Selection in Compressed Sensing Total Variation Minimization-Based Algorithms, Part. Part. Syst. Charact. 37 (2020) 2000070. https://doi.org/10.1002/ppsc.202000070.

[15] C. Li, W. Yin, H. Jiang, Y. Zhang, An efficient augmented Lagrangian method with applications to total variation minimization, Comput. Optim. Appl. 56 (2013) 507–530. https://doi.org/10.1007/s10589-013-9576-1.

[16] Y.X. Wang, Y.J. Zhang, Nonnegative matrix factorization: A comprehensive review, IEEE Trans. Knowl. Data Eng. 25 (2013) 1336–1353. https://doi.org/10.1109/TKDE.2012.51.

[17] R.F.F. Egerton, Electron Energy-Loss Spectroscopy in the Electron Microscope, Springer Science, Boston, MA, 2011. https://doi.org/10.1007/978-1-4419-9583-4.

[18] J. Blanco-Portals, P. Torruella, F. Baiutti, S. Anelli, M. Torrel, A. Tarancón, F. Peiró, S. Estradé, WhatEELS . A new python-based interactive software solution for ELNES analysis combining clustering and NLLS, Ultramicroscopy. 232 (2021) 113403. https://doi.org/10.1016/j.ultramic.2021.113403.

[19] S. Lichtert, J. Verbeeck, Statistical consequences of applying a PCA noise filter on EELS spectrum images, Ultramicroscopy. 125 (2013) 35–42. https://doi.org/10.1016/j.ultramic.2012.10.001.

[20] F. Baiutti, J. Blanco-Portals, S. Anelli, P. Torruella, M. López-Haro, J.J. Calvino, S. Estradé, M. Torrell, F. Peiró, A. Tarancón, J.B. Portals, S. Anelli, P. Torruella, J.J. Calvino, Tailoring the Transport Properties of Mesoporous Doped Cerium Oxide for Energy Applications, J. Phys. Chem. C. 125 (2021) 16451–16463. https://doi.org/10.1021/acs.jpcc.1c04861.

[21] L.A.J. Garvie, P.R. Buseck, Determination of Ce4+/Ce3+ in electron-beam-damaged CeO2 by electron energy-loss spectroscopy, J. Phys. Chem. Solids. 60 (1999) 1943–1947. https://doi.org/10.1016/S0022-3697(99)00218-8.

[22] J.P. Winterstein, C.B. Carter, Electron-beam damage and point defects near grain boundaries in cerium oxide, J. Eur. Ceram. Soc. 34 (2014) 3007–3018. https://doi.org/10.1016/j.jeurceramsoc.2014.02.017.

Chapter 7

General conclusions.

This PhD thesis has been mainly devoted to the exploration and practical implementation of advanced characterization methods for electron energy loss spectroscopy (EELS) and electron tomography (ET). Regarding the analysis of EELS datasets, the focus was set on the incorporation of new machine learning algorithms for clustering and dimensionality reduction. Furthermore, the combination of these newly explored algorithms with the non-linear least squares (NLLS) fitting process was proven advantageous for the characterization of the energy loss near-edge structures (ELNES), especially during the analysis of large spectral datasets composed by a rich mixture of signals from samples with a large number of different elements. Regarding the field of ET and analytical-ET, the focus was set on the translation and successful implementation in Python of an algorithm based on the compressed sensing mathematical theory for the reconstruction process, TVAL3.

Besides the theoretical revision of the foundations and the description of the practical implementation for every single one of these new methodologies, their actual capabilities were tested through the resolution of a series of experimental problems, that effectively served as the real case scenarios demonstrators.

Finally, an important aspect of this thesis was the clear intention to incentivise the widespread use of all these new characterization routes, an effort exemplified by the development of the first version of a free access software solution known as **WhatEELS,** targeted to the ELNES analysis of transition metal and rare earth oxides.

This manuscript opened in chapter 1 with a revision of the state-of-the-art for the TEM characterization techniques of EELS and ET, as well as the most prominent of their potential applications. This summary allowed to put in perspective the work done during the thesis regarding the development of new analysis methodologies.

The general objectives were grouped together and presented in chapter 2. The specific work done towards the completion of every single one of these objectives is extensively described in chapters 3 to 5 for the implementation of the new techniques of analysis, and in chapter 6 for the description of their practical use in order to resolve a challenging characterization problem involving a set of Pr-Gd doped CeO_2 mesoporous materials.

Each one of the chapters 3-5 worked as a single and individual unit, although their interconnection is later revealed in the experimental analysis described in chapter 6. The conclusions, major findings and successes that can be extracted from each one of these chapters are described in the following lines.

Chapter 3 was mainly devoted to the description of the performance of TVAL3, a total variation minimization (TVM) solver, as the reconstruction algorithm for an ET (and analytical ET) experiment.

- The chapter began with a theoretical revision of the reconstruction problem in ET as a TEM characterization technique in section 3.2, opening with the classic iterative reconstruction methods (e.g., SIRT) and working its way up to the more advanced compressed sensing algorithms (i.e., TVAL3). This description put in perspective the potential value of the translation of the TVAL3 algorithm from its original scripting language (MATLAB) to a free access platform (Python). The successful translation and the practical implementation of the TVM will certainly play an important role for the future of the ET (and analytical-ET) reconstructions in our group (LENS).

- The TVM methodology (using TVAL3) was successfully used as a reconstruction technique for a controlled synthetic dataset in section

3.3. In it, the TVAL3 showed an incontestable superiority over the SIRT algorithm though the quantitative results obtained for of a series of tests devised to analyse the effects of several of the most common undersampling problems encountered during the reconstruction of experimental samples (i.e., limited number of projections, low SNR and image erosion).

- In section 3.4.1., a series of morphological ET reconstructions were successfully carried out on two different sets of hollow and quasi-spherical γ-Fe_2O_3 NPs with different shell thicknesses (with average values of 4.26 nm and 4.49 nm), using the TVM method. The semi-quantitative analysis done using an approximate model for the morphological structure, and the volumes measured for several of the reconstructed NPs, allowed a statistical approximation of the differences in the shell thicknesses for both samples.
- In section 3.4.2, the X-EDS analytical-ET done for another quasi-spherical hollow NP with a mixed composition of γ-Fe_2O_3 and NiO confirmed the results of a conventional EELS SI characterization. These results indicated that both oxides formed an alloy throughout the whole structure of the NPs, with occasional pure NiO extrusions towards the surface, instead of the hypothesized double-shell morphology with the NiO located in the outermost part of the NPs.

Chapter 4 was devoted to the implementation of machine learning algorithms for clustering and dimensionality reduction tasks as part of the characterization processes for electron energy loss spectroscopy (EELS).

- The mathematical foundations of the linear matrix factorization solvers (PCA, sPCA and NMF) used for EELS SI analysis as dimensionality reduction methods (DRM) were revised in section 4.3.1. This revision provided a unified vision of these algorithms as solutions with slight variations in the constraints set for the same problem (the extraction of a low rank model for an EELS SI dataset through the minimization of

a matrix factorization formulation). Also, the results of a series of tests conducted for each one of them as DRM, using the same sample of a Fe-Mn oxide core-shell NP of known composition and elemental distribution, showed some of the common pitfalls and potential problems arising from their widespread use.

- The mathematical foundations of the new non-linear graph-based approaches t-SNE and UMAP, for DRM in EELS datasets, were laid out in section 4.3.2. These DRM were applied to the same Fe-Mn oxide core-shell NP, to exemplify how these techniques produce a projection image of the hypothetical manifold in which the datapoints (collection of single spectra) of an EELS SI live, in their native hyperdimensional space. These results demonstrated that, even for such high-dimensional datasets, these non-linear DRM were potentially capable of retaining structural information after the projection to the low dimensional embedding.

- Section 4.4 revised the mathematical foundations of several of the already known clustering algorithms used for EELS analysis (K-means, hierarchical agglomerative clustering, DBSCAN and OPTICS), and introduced the state-of-the-art hierarchical density-based algorithm of HDBSCAN. The potential effects of using these algorithms for a segmentation of an EELS SI (without any previous DRM employed) were explored using the same Fe-Mn oxide NP dataset.

  At the same time, a small revision of the most commonly used metrics by these algorithms for the computations of distances between datapoints was provided, stablishing a clear parallelism between the so-called cosine 'metric' and an effective L_2 pre-normalization of the dataset followed by any computation using the Euclidean metric.

- Section 4.5 showed the promising results of using the combination of the new DRM and clustering algorithms (UMAP and HDBSCAN), through the analysis of the Fe-Mn oxide NP. Not only they produced the

expected segmentation results (already proven by quantitative methods in previous analysis of the sample using the Oxide Wizard software solution for the Digital Micrograph, see section 4.2), but they also did so by following a complete data-driven methodology, as no intervention at any point in the analysis procedure was required from the user.

- The results from section 4.6.1 proved the superior performance capabilities of the UMAP-HDBSCAN combination using a synthetic dataset to conduct quantitative measures.

  This combination was able to identify small changes in the energy loss position of WLs (i.e., changes in the oxidation state of the materials reflected in the ELNES), even when these changes were modelled by a continuous and slowly varying gradient. At the same time, it was also successful in the detection of small changes of the elemental composition, for elements included in areas of the dataset with a weight below the 5% mark (also in cases modelled with a continuously varying function). On top of that, the synthetic dataset was devised to mimic a real case scenario in which with a gradient modulation for the background signal arising from an underlying thickness variation in the sample permeates the whole SI.

  Finally, these different composition regions and areas with small changes in the oxidation state presented a diverse range of sizes, a fact that posed a particular difficult problem to the other combinations of DRM and non-density-based clustering algorithms.

- The new DRM and clustering algorithms were proven capable of dealing with cases of low SNR in section 4.6.2. Once again, a synthetic dataset was devised to test the detection limits for small EELS signals diluted in the background through a series of examples heavily polluted with noise. The outstanding capabilities of UMAP to produce a faithfully representation of the manifold underlying the EELS SI structure, as well

as the capability of HDBSCAN to separate smaller but denser regions within a larger distribution of datapoints, were proven again as the key aspects of this methodology combining both algorithms that allowed the detection of such outlier regions included in the synthetic dataset.

Chapter 5 introduced **WhatEELS**, the integral software solution developed during this PhD thesis to combine clustering and NLLS and that includes several of the most popular analysis tools for the characterization of transition metal and rare earth oxides.

- The revision of the theory behind the elemental quantification process using the Egerton method for EELS was shown in section 5.2. A clear view is provided for the importance of including the experimental limiting effects of the apertures on the convergence and collection angles for the computation of the elemental partial cross sections, and as an extension, for the elemental quantification process itself. The computations carried out by **WhatEELS** for the integration of the partial cross sections were proven to be aligned with those provided by Digital Micrograph when the geometric correction factor was set in place.

- The architecture and inner workings of the software, as well as a series of visual examples for the succession of actions required to complete an analysis, were successfully described in section 5.3. The software was divided between the peripheral modules (i.e., comprised by the general oscillator strength surface visualizer and the clustering analysis tool) and the core modules (i.e., the tools to create the components of the model and carry out the NLLS fitting, as well as those in charge of the analysis of the results).

- The results of using **WhatEELS** for a quantitative analysis of the Fe-Mn oxide NP calibration sample of known composition were showcased in section 5.4. The software solution was proven to provide an accurate picture of the elemental distribution, as well as an accurate description

of the areas with different manganese oxide phases within the shell (i.e., different Mn oxidation states).

Finally, chapter 6 describes the results for the analysis for a series of samples of CeO_2 mesoporous materials doped with different quantities of Gd and Pr, and using all the new analysis techniques described in chapters 3 to 5. The adoption of these types of ceramic materials for high temperature catalysis and thermochemical fuel production has been highlighted in the literature previously. The actual samples analysed in this PhD thesis were devised to be included as the backbone layer for the electrode in an energy conversion device operating at high temperature (solid oxide fuel cell, SOFC), given their promising performances as ionic and mixed ionic-electronic conductors. The results of our analysis played a key role in understanding the macroscopic electrochemical performance from the nanoscopic physical phenomena supported by the characterization of the structure of these materials at the nanoscale.

- The qualitative analysis conducted through the combination of UMAP-HDBSCAN for the $CeGd_{0.2}O_{0.8}$ (or CGO), $CePr_{0.2}O_{0.8}$ (or CPO) and $CeGd_{0.1}Pr_{0.1}O_{0.9}$ (or CGPO) samples in section 6.2. showed a clear tendency of the Gd and Pr dopants to segregate towards the main CeO_2 structures grain boundaries (GB) and grain surfaces (GS).

- Section 6.3.1. showed the quantitative analysis of the CGPO sample. It was carried out by the combination of the dataset segmentation via UMAP-HDBSCAN and the NLLS fitting via **WhatEELS.** The analysis of the central positions through the SI for the gaussian component assigned to CeM_5, as well as the analysis of the white-line (WL) ratio variations for the CeM_{54} edge was in good agreement with the hypothesis of cerium reduction localized in the GB and GS. At the same time, the elemental quantification showed the same tendency of Pr and Gd dopant segregation towards the same GB and GS where the Ce was reduced.

- The results for the ET and X-EDS analytical-ET reconstructions (using the TVM method) for the CGPO sample were shown in section 6.3.2. They qualitatively described the formation of high-density concentration of dopants in small clusters towards the boundaries between grains throughout the whole volume of the mesoporous structure.

- The results of the time-series experiments for the CGO sample, shown in section 6.4.1, showcased the importance of the precise control of the electron dosage over the sample to avoid biased results for the measurements of the Ce oxidation state in the materials of this composition.

- The quantitative results for the CGO sample shown in section 6.4.1. (using UMAP-HDBSCAN and **WhatEELS**), showed the same tendency of Ce reduction and dopant segregation towards GB and GS than that of the CGPO.

- The quantitative results for the CGO(HF) sample (using UMAP-HDBSCAN and **WhatEELS**), shown in section 6.4.2., demonstrated the effectivity of the HF post-processing etching procedure in eliminating the excess of dopant in the GB and GS of the material, while maintaining a constant dopant ratio close to the stoichiometric values throughout the whole structure of the CeO_2 mesoporous sample.

To conclude this manuscript, the intention of these final lines is to put in perspective the importance of the work done, revisiting once again each one of the topics covered above.

Beginning with the field of electron tomography, the algorithm chosen (TVAL3) to bring the compressed sensing reconstruction capabilities to our group (LENS), while it has already been showcased in the literature for some time, it still remains as one of the forefront options in the field. The morphological and analytical volume reconstructions carried out during the thesis were often done for samples that presented some sort of limitation to

the maximum electron dosage allowed (i.e., a limitation on the maximum illumination time and, thus, on the maximum number of projections acquired). Even under these severe undersampling conditions, that in some cases were further degraded by a low SNR (e.g., the analytical sets acquired for shorter periods of time such as fast X-EDS mappings in analytical-ET), the TVM method for the reconstructions showed an outstanding performance overall.

Another important point to be made is the translation of this algorithm (TVAL3) to Python. Although freely distributed by their owners, it was only available in a proprietary language (MATLAB), and for a long time this hindered its proper use within our group.

Following this thread of insufficient available tools or inefficient available ones, precisely these factors were the main drive behind the development of **WhatEELS**. Long periods of time were invested on developing homemade tools for ELNES analysis, in order to tackle some of the more challenging characterization problems encountered during the thesis (e.g., the problem described for the set of mesoporous materials). **WhatEELS** is the culmination of the intention to bring together all these new solutions under a common roof. Hopefully this tool will also serve as a solid ground for future expansions, probably focused on the implementation of new machine learning solutions for EELS and ELNES analysis.

Precisely some of these future machine learning solutions for EELS analysis were also explored on this thesis. The state-of-the-art algorithms for dimensionality reduction and clustering, UMAP and HDBSCAN respectively, demonstrated to be valuable assets for the clustering segmentation of large spectral datasets. They even showed that a fully data-driven analysis is possible while still beating the performance of all the previous candidates by a good margin.

The potential of using all these advanced techniques (i.e., TVM, **WhatEELS** and UMAP-HDBSCAN) for the resolution of complex problems was finally showcased with the set of Pr-Gd doped CeO_2 mesoporous materials. Again, the

tools available at the beginning of this PhD only allowed for a qualitative approach to the problem. However, by the end of it the only qualitative analysis remaining was that done on the volumes reconstructed for the mesoporous structures (and only prompted by the limited number of projection angles acquired and the registering times for the X-EDS maps). Both the advanced clustering techniques combined with NLLS and the homemade results analysis tools included in **WhatEELS** provided the resources to resolve quantitatively the problems of the dopant segregation and local changes in the Ce oxidation state.

And the most important part, these tools are now available for free to be used by anyone, to resolve all sorts of new and exciting problems.

Appendix A

Instrumentation list.

A list of all the equipment used to prepare the samples and acquire the experimental datasets analysed along the development of this thesis is presented in the following lines.

CCIT – Universitat de Barcelona

JEOL JEM 2100, with a LaB_6 thermionic electron gun operating at 200 kV, equipped with an Oxford Instruments EDX system, and a set of bright field/dark field (HAADF) STEM detectors.

JEOL JEM 2010F, with a field emission gun (FEG) operating at 200 kV, equipped with a Gatan GIF for the acquisition of EELS and a set of bright field/dark field (HAADF) STEM detectors.

Ion milling machines from Fischione (model 1010), and GATAN (PIPS)

Single grinder polisher machine from Pace Technologies, with a 12'' rotating disc. A polishing tripod was used to support the sample.

Plasma cleaner from Fischione, model 1020.

CNBM – Adam Mickiewitcz University

Probe corrected JEOL JEM ARM200F operating at 200 kV, equipped with a Gatan Quantum GIF EELS detector, JEOL EDX detector and bright field/dark field (HAADF) STEM detectors.

DME – Universidad de Cádiz

Probe and image corrected ThermoFisher TITAN3 Themis 60-300, with a XFEG operated at 80 and 200 keV, equipped with a Gatan Quantum ERS 966 for EELS, a Super-X Chemi-STEM for X-EDS, and a CMOS ceta 4k camera for HRTEM. It was used with both Fischione Model 2020 y 2021 holders for electron tomography.

Appendix B

Scientific curriculum.

Education

2021 – 2016	PhD in Nanoscience Universitat de Barcelona (UB)
2015 – 2016	Master's degree in Nanoscience Universitat de Barcelona (UB)
2010 – 2015	Bachelor's degree in Physics Universidad de Santiago de Compostela (USC)

Grants awarded

2018	Ajuts per a la contractació de personal investigador predoctoral en formació (FI-2018). Grant 2018FI_B_00360 from Comissió Executiva d'Ajuts de Recerca de l'Agència de Gestió d'Ajuts Universitaris i de Recerca (AGAUR) de la Generalitat de Catalunya (Spain)

Journal publications

1. **Blanco-Portals, Javier**; Peiró, Francesca & Estradé, Sònia. "Strategies for EELS Data Analysis. Introducing UMAP and HDBSCAN for Dimensionality Reduction and Clustering". ***Microscopy and Microanalysis*** (2021) pp. 1 – 14.
   DOI: doi.org/10.1017/S1431927621013696

2. **Blanco-Portals, Javier**; Torruella, Pau; Baiutti, Federico; Anelli, Simone; Torrell, Marc; Tarancón, Albert; Peiró, Francesca & Estradé, Sònia. "WhatEELS. A python-based interactive software solution for ELNES analysis combining clustering and NLLS". ***Ultramicroscopy*** (2022) 232, pp. 113403. DOI: doi.org/10.1016/j.ultramic.2021.113403

3. Baiutti, Federico; **Blanco-Portals, Javier**; Anelli, Simone; Torruella, Pau; López-Haro, Miguel; Calvino, José; Estradé, Sònia; Torrell, Marc; Peiró, Francesca; & Albert Tarancón. "Tailoring the Transport Properties of Mesoporous Doped Cerium Oxide for Energy Applications". ***The Journal of Physical Chemistry C*** (2021) 125-30, pp. 16451 – 16463. DOI: doi-org.sire.ub.edu/10.1021/acs.jpcc.1c04861

4. Valenti, Marco; Bleiji, Yorick; **Blanco-Portals, Javier**; Muscarella, Loreta A.; Aarts, Mark; Peiró, Francesca; Estrade, Sònia & Alarcón-Lladó, Esther. "Grain size control of crystalline III–V semiconductors at ambient conditions using electrochemically mediated growth" ***Journal of Materials Chemistry A*** (2020) 8 – 5, pp. 2752-2759.
   DOI: doi.org/10.1039/c9ta07410d

5. Fonoll-Rubio, Robert; Andrade-Arvizu, Jacob; **Blanco-Portals, Javier**; Becerril-Romero, Ignacio; Guc, Maxim; Saucedo, Edgardo; Peiró, Francesca; Calvo-Barrio, Lorenzo; Ritzer, Maurizio; Schnohr, Claudia S.; Placidi, Marcel; Estradé, Sònia; Izquierdo-Roca, Victor & Pérez-Rodríguez, Alejandro. "Insights into interface and bulk defects in a high efficiency kesterite-based device". ***Energy & Environmental Science*** (2021) 14 – 1, pp. 507 – 523.
DOI: doi.org/10.1039/D0EE02004D

6. López-Vidrier, Julià; Gutsch, Sebastian; Blázquez, Oriol; Valenta, Jan; Hiller, Daniel; Laube, Jan; **Blanco-Portals, Javier**; López-Conesa, Lluís; Estradé, Sònia; Peiró, Francesca; Garrido, Blas; Hernández, Sergi & Zacharias, Margit. "Effect of Si_3N_4-mediated inversion layer on the electroluminescence properties of silicon nanocrystal superlattices". ***Advanced Electronic Materials*** (2018) 4 – 5, pp. 1700666. DOI: doi.org/10.1002/aelm.201700666

7. Berestok, Taisiia; Guardia, Pablo; Du, Ruifeng; **Blanco-Portals, Javier**; Colombo, Massimo; Estradé, Sònia; Peiró, Francesca; Brock, Stephanie L. & Cabot, Andreu. "Metal oxide aerogels with controlled crystallinity and faceting from the epoxide-driven cross-linking of colloidal nanocrystals". ***ACS applied materials & interfaces*** (2018) 10 – 18, pp. 16041-16048. DOI: doi.org/10.1021/acsami.8b03754

8. Berestok, Taisiia; Guardia, Pablo; **Blanco-Portals, Javier**; Estradé, Sònia; Llorca, Jordi; Peiró, Francesca; Cabot, Andreu & Brock, Stephanie L. "Metal oxide aerogels with controlled crystallinity and faceting from the epoxide-driven cross-linking of colloidal nanocrystals". ***ACS applied materials & interfaces*** (2018) 10 – 18, pp. 16041-16048. DOI: doi.org/10.1021/acsami.8b03754

9. Torruella, Pau; Ruiz-Caridad, Alicia; Walls, Michael; Roca, Alejandro G.; López-Ortega, Alberto; **Blanco-Portals, Javier**; López-Conesa, Lluís; Nogués, Josep; Peiró, Francesca & Estradé, Sònia. "Atomic-scale determination of cation inversion in spinel-based oxide nanoparticles". ***Nano Letters*** (2018) 19 – 9, pp. 5854-5861. DOI: doi.org/10.1021/acs.nanolett.8b02524

10. Dalmases, Mariona; Torruella, Pau; **Blanco-Portals, Javier**; Vidal, Albert; Lopez-Haro, Miguel; Calvino, José; Estradé, Sònia, Peiró, Francesca & Figuerola, Albert "Gradual transformation of Ag_2S to Au2S nanoparticles by sequential cation exchange reactions: binary, ternary, and hybrid compositions". ***Chemistry of Materials*** (2018) 30 – 19, pp. 6893-6902. DOI: doi.org/10.1021/acs.chemmater.8b03208

11. Berestok, Taisiia; Guardia, Pablo; **Blanco-Portals, Javier**; Nafria, Raquel; Torruella, Pau; López-Conesa, Lluís; Estradé, Sònia; Ibáñez, María; de Roo, Jonathan; Luo, Zhishan; Cadavid, Doris; Martins, José C.; Kovalenko, Maksym V.; Peiró, Francesca & Cabot, Andreu. "Tuning branching in ceria nanocrystals". ***Chemistry of Materials*** (2017) 20 – 10, pp. 4418-4424. DOI: doi.org/10.1021/acs.chemmater.7b00896

12. **Blanco-Portals, Javier**; Berestok, Taisiia; Torruella, Pau; Coll, Catalina; López-Conesa1, Lluís; Guardia, Pablo; Coy, Emerson L.; Cabot, Andreu; Estradé, Sònia & Peiró, Francesca. "Atomistic modelling and high-resolution electron microscopy simulations of CeO2 nanoparticles". ***Applied Physics Letters*** (2017) 111 – 22, pp. 223107. DOI: https://doi.org/10.1063/1.5005499

Book chapters

1. Torruella, Pau; **Blanco-Portals, Javier**; Yedra, Lluís; López-Conesa, Lluís; Rebled, Josep M.; Peiró, Francesca & Estradé, Sònia. "Electron Tomography." ***New Trends in Nanoparticle Magnetism***. Springer, Cham, (2021).pp 257 – 283
   DOI: doi.org/10.1007/978-3-030-60473-8_11

Congress contributions

1. "Bulk and interfacial limiting factors on Sb_2Se_3 thin film solar cells." Vidal-Fuentes, B.P.; Blanco-Portals, J.; Becerril-Romero, I.; Guc, M.; Peiró, F.; Saucedo, E.; Pérez-Rodríguez, A.; Estradé, S.; Izquierdo-Roca, V.; **E-MRS Spring Meeting** (2021). Oral.

2. "WhatEELS. Clustering and NLLS in a New Software Solution for ELNES Analysis." **Blanco-Portals, J**.; Torruella, P.; Baiutti, F.; Anelli, S.; Torrell, M.; Tarancón, A.; Peiró, F.; Estradé, S.; **(MFS) Microscopy at the Frontiers of Science** (2021). Oral (*speaker*).

3. "WhatEELS. A complete software solution for ELNES analysis that combines clustering and NLLS." **Blanco-Portals, J**.; Torruella, P.; Baiutti, F.; Anelli, S.; Torrell, M.; Tarancón, A.; Peiró, F.; Estradé, S.; **XXXVI Trobades Científiques de la Mediterrània Josep Miquel Vidal - Microscòpies i espectroscòpies: accedint a la nanoescal**,(2021). Oral.(*speaker*)

4. "Electron energy loss spectroscopy and machine learning tools for unveiling oxidation state and inversion parameter at atomic resolution in Mn_xO_y-Fe_xO_y spinel oxides core-shell nanoparticles." Torruella, P.; Ruiz-Caridad, A.; Walls, M.; Roca, A.G.; Lopez-Ortega, A.; Blanco-Portals, J.; Lopez-Conesa, L.; Nogues, J.; Estradé, S. Peiró, F..; **7th Annual Conference on AnalytiX** (2019). Oral.

5. "TEM techniques applied to advanced materials for fuel cells: from electron diffraction to analytical tomography." Blanco Portals, J.; López-Conesa, L.; Estradé, S.; Peiró, F.; **3rd ELECMI International Workshop** (2019). Invited / keynote talk.

6. "Hierarchical clustering algorithms applied to large EELS datasets." Peiró, F.; Torruella, P.; López-Conesa, L.; Blanco-Portals, J.; Estradé, S.; **Nanotech Poland 2019 and Nanotechnology and Innovation in the Baltic sea Region**,(2019). Oral.

7. "Determination of the inversion parameter in spinel oxides core-shell nanoparticles at atomic resolution by STEM-EELS techniques." Torruella, P.; Ruiz-Caridad, A.; Walls, M.; Roca, A.G.; López-Ortega, A.; Blanco-Portals, López-Conesa, L.; J. Nogués, J.; Estradé, S.; Peiró, F.; **2nd Global Conference on Magnetism and Magnetic Materials**,(2019). Oral.

8. "EELS plasmon mapping of gold nanodomes assessed by BEM simulations." Coll, C.; Bicket, I.; Sepúlveda, B.; Blanco-Portals, J.; López Haro, M.; Nogués, J.; Botton, G.; Estradé, S.;
Peiró, F.; **4ª REUNIÓN DE LA RED de EXCELENCIA IMAGINE**, (2019). Oral.

9. "OxiSpy: a hyperspy-based multifitting tool for large EELS-SI in transition metals and rare earth oxides." **Blanco-Portals, J**.; Baiutti, F.; Anelli, S.; Torrell, M.; López-Haro, M.; Tarancón, A.; Estradé, S.; Calvino, J.J.; Peiró, F.; **4ª REUNIÓN DE LA RED de EXCELENCIA IMAGINE,** (2019). Oral (*speaker*).

10. "Advanced EELS and X-EDS electron tomography techniques applied to the resolution of doped CeO_2 mesoporous structures." **Blanco-Portals, J.**; Baiutti, F.; Anelli, S.; Torrell, M.; López-Haro, M.; Estradé, S., Tarancón, A.; Calvino, J.J.; Peiró, F.; **Annual Meeting of the Institute of Nanoscience and Nanotechnology of the UB (IN2UB)** (2019). Poster (*presenter*).

11. "New generation mesoporous oxygen electrode for solid oxide cell application." Baiutti, F.; Anelli, S.; Torrell, M.; Blanco Portals, J.; López-Conesa, L., Estradé, S.; Peiró, F.; Tarancón, A.; **22nd International Conference on Solid State Ionics (SSI-22)**, (2019). Invited / keynote talk.

12. "Improved mesoporous oxygen electrode for Solid Oxide Electrolyser Cells application." Baiutti, F.; Anelli, S.; Torrell, M.; Blanco Portals, J.; López-Conesa, L., Estradé, S.; Peiró, F.; Tarancón, A.; **IEEE International Symposium on the Applications of Ferroelectrics (ISAF), International Conference on Electroceramics (ICE); European Meeting on Ferroelectricity (EMF); International Workshop on PiezoMEMS (IWPM); Piezoresponse Force Microscopy Workshop (PFM). ISAF/EMF/ICE/IWPM/PFM Joint Conference (F2cπ2)**,(2019). Invited / keynote talk.

13. "Doped cerium oxide mesoporous structure: a complete analytical TEM characterization based on MVA, clustering and X-EDS electron tomography." **Blanco-Portals, J.**; Baiutti, F.; Anelli S.; Torrell, M.; López-Haro, M.; Estradé, S.; Tarancón, A., Calvino, J.J.; Peiró, F.; **(MFS) Microscopy at the Frontiers of Science** (2019). Oral (*speaker*).

14. "High-Resolution Electron Microscopy and High-Angle Annular Dark-Field Image Simulation of CeO_2 Nanoparticles." **Blanco-Portals, J.**; Berestok, T.; Torruella, P.; Coll, C.; López-Conesa, L.; Guardia, P.; Coy, E.; Cabot, A.; Estradé, S.; Peiró, F.; **IN2UB Annual Meeting** (2018). Poster (*presenter*)

15. "Atomic resolution spatially-resolved inversion parameter in spinel oxides." Torruella, P.; Ruiz-Caridad, A.; Walls, M.; Roca, A.G.; Lopez-Ortega, A.; Blanco-Portals, J.; Lopez-Conesa, L.; Nogues, J.; Peiro, F.; Estrade, S.; **19th International Microscopy Congress - IMC19**, (2018). Oral.

16. "Machine Learning and Big Data Analysis Techniques in Electron Tomography: Towards a non-biased and Automatized Segmentation of the Spectrum Volume." **Blanco-Portals, J.**; López-Haro, M.; Estradé, S.; Calvino, J.J.; Peiró, F.; **2ª REUNIÓN DE LA RED de EXCELENCIA IMAGINE**,(2018). Oral (speaker).

17. “Elucidation of the structure of octapod branched ceria nanocrystals by transmission electron microscopy: contrast simulations and electron tomography.” Blanco-Portals, J.; Berestok, T.; Torruella-Besa, P.; Coll, C.; López-Conesa, L.; Guardia, P.; Cabot, A.; Estradé, S.; Peiró, F.; **NanoTech Poland International Conference and Exhibition**, (2017). Invited / keynote talk

18. “Atomistic Modelling and High-Resolution Electron Microscopy Simulations of CeO_2 Octapods.” **Blanco-Portals, J.**; Berestok, T.; Torruella, P.; Coll, C.; López-Conesa, L.; Guardia, P.; Coy, E.; Cabot, A.; Estradé S.; Peiró, F.; **(MFS) Microscopy at the Frontiers of Science** (2017). Oral (*speaker*).

19. “Assessment of octopod morphology of branched ceria nanocrystals.” Blanco-Portals, J.; Berestok, T.; Torruella, P.; Coll, C.; López-Conesa, L.; Guardia, P.; Coy, E.; Cabot, A.;
Estradé S.; Peiró, F.; **5th Nano Today Conference** (2017). Poster.

20. “Assembly of sulphide-based nanoparticles into gels and aerogels” Berestok, T.; Guardia, P.; Meyns, M.; Blanco, J.; López-Conesa, L.; Estrade, S.; Brock, S.L.; Peiro F.; Cabot, A..; **EUROMAT European Congress and Exhibition in Advanced Materials and Processes**, (2017). Oral.

Workshops and schools

2017 The TEM-UCA Summer Workshop 'Transmission Electron Microscopy of Nanomaterials'. Dep. de Ciencia de los Materiales e Ingeniería Metalúrgica y Química Inorgánica, Universidad de Cádiz. July.

2017 Quantitative Electron Microscopy QEM2017. CNRS - French national research agency. June.

Supervision of research projects.

2019 Reconstruction algorithms for electron tomography: application to AuAgS nanorods. **Bachelor Thesis**. David Mañá González.

2018 Big data analysis applied to EEL spectroscopy. **Bachelor Thesis.** Joan Sospedra.

2018 TEM characterization of advanced kesterite structures. **Master's Thesis.** Robert Fonoll.

Appendix C

Resumen en castellano

Esta tesis se ha dedicado principalmente a la exploración e implementación de nuevas herramientas y técnicas de análisis computacional para la caracterización de nanomateriales y dispositivos mediante la microscopía de transmisión de electrones (TEM). En concreto, los campos de la espectroscopia de pérdida de energía de electrones (EELS) y de la tomografía electrónica (TE) centraron la atención de la mayor parte del trabajo realizado, ambos a la vanguardia de las técnicas de caracterización a la nanoescala en el campo de la ciencia de materiales.

Empezando por EELS, esta técnica da acceso al estudio de una amplia gama de diferentes propiedades y fenómenos físicos en materiales a escalas nanoscópicas, destacando entre otras técnicas por su alta resolución espacial y energética. Típicamente, el EELS es capaz de resolver detalles en una escala espacial de Ángstroms(**Å**, es decir, atómica) y a una escala energética menor a 1eV (en torno a 0.7 eV para un cañón de electrones FEG, e incluso en torno a 100 meV con un monocromador estándar).

Algunos de los fenómenos físicos a la nanoescala que pueden investigarse mediante espectroscopía EELS serían la medición de los valores de energía de banda prohibida para distintos tipos de nanomateriales, o la determinación de la respuesta de los plasmones de superficie y sus modos resonantes en dispositivos para aplicaciones ópticas a la nanoescala. El estudio de fonones mediante espectroscopía EELS también ha sido recientemente desbloqueado, gracias en parte al incremento del brillo de las fuentes de electrones y los nuevos monocromadores para el haz.

De manera más común, y especialmente en el contexto de este doctorado, el EELS se ha utilizado para el análisis cuantitativo y cualitativo de distribuciones de elementos a la nanoescala. En estos casos, el EELS destaca por su capacidad para la detección de elementos ligeros en la tabla periódica (p. Ej., oxígeno, nitrógeno y carbono). A mayores, el EELS permite el estudio de los estados de oxidación y de los números de coordinación atómica con una resolución atómica. Para ello, se recurre al análisis de la estructura fina del espectro en las regiones cercanas al borde principal de pérdida de energía (ELNES) en los espectros EELS.

Los avances en la instrumentación TEM, incluido el desarrollo de mejores espectrómetros EELS (por ejemplo, cámaras de detección directa) y software para una adquisición más rápida, están detrás del aumento de tamaño y complejidad en los datos espectrales adquiridos. Siguiendo esta línea, el requisito de adquirir espectros con una gran resolución en energía y un buen ratio señal / ruido (SNR) para el análisis ELNES ha sido paulatinamente sustituido en la práctica por la adquisición de un número mucho mayor de imágenes de espectros, aun a coste de una menor calidad. Este nuevo paradigma se fundamenta en la hipótesis de que una muestra estadística mayor, aunque ruidosa, proporciona acceso al mismo nivel de detalle en el proceso de caracterización que una muestra espectral más pequeña y limpia, con la ventaja adicional de requerir habitualmente tiempos de adquisición más cortos. Esta práctica es ahora común para la caracterización de materiales sensibles al haz de electrones o para experimentos centrados en el estudio de los estados de oxidación de materiales propensos a reducirse/oxidarse bajo el haz de electrones.

Este aumento en el tamaño de los conjuntos de datos y los desafíos de lidiar con mayores niveles de ruido, han propiciado la incorporación de múltiples técnicas de análisis originarias del campo del aprendizaje automatizado (‘machine learning’ en inglés), una tendencia que se espera continúe en el futuro.

Algunos algoritmos, como el de análisis de componentes principales (PCA) o la factorización matricial no negativa (NMF), llevan tiempo siendo empleados para el análisis de datos EELS. Recientemente el interés se ha enfocado en la incorporación de nuevos métodos de agrupamiento ('clustering' en inglés) para resolver problemas de segmentación de imágenes de espectros EELS de manera no supervisada. Una de las principales ventajas de estos métodos es que los representantes espectrales de cada uno de los grupos resueltos son siempre físicamente significativos, ya que son calculados mediante el promedio de señales. Esto no siempre sucedía con PCA y NMF.

Un ejemplo de la utilización de estos nuevos algoritmos sería el uso de K-means para extraer los representantes espectrales de un set de datos, que luego son utilizados en un ajuste de mínimos cuadrados de múltiples componentes (MLLS) para mejorar la exactitud de los resultados obtenidos. Otro ejemplo sería el del uso combinado de técnicas no-lineales de reducción dimensional (como t-SNE) y algoritmos de agrupamiento basados en la densidad de puntos (OPTICS) para la segmentación de imagen de espectros mediante una metodología puramente basada en los datos (reduciendo posibles prejuicios en la interpretación de resultados) .

Una parte muy importante del trabajo realizado en esta tesis se ha centrado en la elaboración de una revisión sistemática de estos nuevos métodos de agrupamiento, de los métodos de reducción dimensional ya establecidos (PCA y NMF), y de algunas de las posibles combinaciones entre ambos tipos de técnicas. Esta revisión se llevó a cabo a través de una serie de análisis cualitativos y cuantitativos del rendimiento de dichos algoritmos, así como de sus ventajas y de sus mayores deficiencias. Además, se introdujo una nueva metodología mediante el uso combinado de algoritmos de última generación para la reducción dimensional no-lineal (UMAP) y algoritmos jerárquicos de agrupamiento basados en la densidad de puntos (HDBSCAN). Hemos demostrado que el rendimiento de esta combinación supera con creces el de cualquier otra técnica de segmentación estudiada para imágenes de espectro EELS.

Además, el uso combinado técnicas de ajuste no-lineales por mínimos cuadrados (NLLS) y de algoritmos de agrupamiento ha resultado ser una solución prometedora para mejorar la estabilidad en el análisis de ELNES. Por lo tanto, esta combinación representa un activo importante para el análisis de datos en experimentos que involucran cambios de estado de oxidación (por ejemplo, la reducción localizada de cationes en óxidos de metales de transición y óxidos de tierras raras).

Junto al estudio e implementación de toda esta serie de nuevas técnicas para el análisis de EELS, un objetivo prioritario de este trabajo de doctorado ha sido el proporcionar una solución de software completa que las incluyese. De esta forma, cualquier usuario futuro podrá reproducir y aplicar estas metodologías a nuevos experimentos, sin requerir de una alta competencia en lenguajes de programación. Este objetivo desembocó en el desarrollo de la herramienta de software denominada **WhatEELS**.

Pasando ahora al campo de la reconstrucción tridimensional por tomografía de electrones (TE), esta técnica supone uno de los conjuntos de herramientas más versátiles para la caracterización a nanoescala, ya que permite la reconstrucción de volúmenes a partir de un conjunto limitado de proyecciones adquiridas.

Dichas reconstrucciones pueden orientarse hacia la recuperación de la morfología en 3D (TE morfológica o, simplemente, TE) o hacia la obtención de la composición química en 3D (TE-analítica).

La TE morfológica utiliza para el proceso de reconstrucción los datos adquiridos mediante los modos de imagen TEM (es decir, imágenes de campo oscuro HAADF o DF, o incluso imágenes de campo brillante, BF). Por lo general, la TE se centra en la reconstrucción de volúmenes a la 'mesoescala', en la que se pueden investigar cuantitativamente la distribución de tamaños y dispersión de aglomerados de NP, resolver las facetas expuestas en la superficie de una NP u obtener la estructura interna de cavidades. También existe la posibilidad de reconstruir la estructura cristalina atómica en 3D

mediante un experimento de TE. Para ello, en la literatura se han descrito experimentos que recurren a conjuntos de proyecciones ADF (o HAADF) con resolución atómica donde la muestra es orientada en varios ejes de zona. También se han descrito experimentos que recurren a una combinación de adquisición de imágenes HAADF de alta resolución y procesos de 'conteo de átomos' (es decir, procesos modelado atómico de las imágenes HAADF que usan la relación entre el contraste y la profundidad de la columna atómica orientada).

Con respecto a la TE analítica, las reconstrucciones se llevan a cabo con proyecciones constituidas por mapas químicos, que a su vez son extraídos de datos espectrales adquiridos mediante cualquiera de las técnicas analíticas en un TEM (es decir, X-EDS o EELS). En algunos casos, ambas señales son adquiridas y reconstruidas simultáneamente. Generalmente, el objetivo de estas reconstrucciones está en la obtención de las distribuciones en 3D de los diferentes elementos presentes en los nanomateriales. Sin embargo, también se ha demostrado que la TE-analítica es capaz de abordar problemas con un mayor nivel de complejidad, como la reconstrucción en 3D de la distribución de los estados de oxidación de los elementos químicos del material.

Estos últimos se basan principalmente en el análisis ELNES de una serie de imágenes de espectros EELS, ya que estas estructuras características proporcionan un fácil acceso a la identificación de estados de oxidación para cationes en óxidos de metales de transición y óxidos de tierras raras. La extracción de los mapas para esta distribución de los estados de oxidación en cada proyección puede abordarse de diversas maneras. Por un lado, estaría el análisis de componentes independientes (ICA) para el set de datos de EELS. Una vez realizada la descomposición ICA, se utilizarían los pesos de las componentes ELNES asociadas a los estados de oxidación de interés como si se tratasen de las proyecciones para el proceso de reconstrucción mediante una TE. Otra opción sería el cálculo de los mapas de pesos para un ajuste MLLS de las señales de referencia ELNES de los diferentes estados de oxidación de interés en el material bajo análisis.

En general, los avances en TE (también aplicable a TE-analítica) se pueden dividir en dos subcategorías diferentes. La primera estaría relacionada con el proceso de adquisición en sí mismo. Esto incluiría tanto la implementación de nuevos modos y herramientas de software para la adquisición como los avances en la instrumentación TEM específicamente enfocada a la adquisición experimental de series tomográficas.

Por ejemplo, recientemente, una gran parte de la investigación en el campo de la TE se ha dedicado a la implementación de rutinas de adquisición más rápidas y eficientes. Estas evitarían, en principio, la sobreexposición de las muestras al haz de electrones, sin comprometer el nivel de detalle reconstruido (TE rápido o fast-ET en inglés). Uno de los avances instrumentales más relevantes para el campo de la TE (tanto morfológica como analítica) ha sido el desarrollo de las cámaras y sensores de detección directa, con una sensibilidad mejorada y respuesta más rápida. Estos instrumentos han facilitado la obtención de un mayor número de proyecciones durante los experimentos de TE, evitando el asociado aumento del tiempo total para la adquisición. A su vez, también permiten una reducción drástica de los voltajes y corrientes en el haz de electrones, manteniendo aun así unos niveles aceptables de SNR.

El desarrollo del crio-TEM también se encuentra a la vanguardia de los avances para TE, dado que permite la implementación de esta técnica de caracterización en muestras que de otra manera serían demasiado sensibles al haz de electrones. El número de publicaciones en este campo en particular (crio-TEM en general, y crio-TEM TE en particular) se ha visto dramáticamente incrementado en los últimos tiempos, apoyado por los avances en la técnica. La relevancia de dichos avances queda perfectamente ejemplificada a través de la concesión del Premio Nobel de Química 2017 a Jacques Dubochet, Joachim Frank y Richard Henderson por su trabajo en el desarrollo de crio-TEM.

La segunda subcategoría de avances en el campo de la TE se centraría la mejora y el desarrollo de algoritmos para el proceso de reconstrucción. Sin embargo,

en la literatura todavía se puede encontrar de forma común el uso algoritmos de reconstrucción antiguos. Un ejemplo sería la reciente reconstrucción de la estructura 3D de la maquinaria molecular para el virus SARS-CoV-2 a partir de una serie de imágenes adquiridas en un crio-TEM, utilizando el algoritmo algebraico iterativo SIRT. Aunque potencialmente exitosos, estos algoritmos más antiguos requieren por lo general un gran número de proyecciones para lograr una reconstrucción precisa. Incluso entonces, los resultados pueden mostrar algunos de los artefactos de reconstrucción más persistentes (por ejemplo, artefactos de líneas, o 'streaking artefacts' en inglés).

La inclusión de información conocida a priori acerca de las muestras estudiadas en el proceso de reconstrucción, como el uso de algoritmos formulados a partir de la teoría matemática de muestreo disperso ('compressed sensing' en inglés), ha sido probado con un alto nivel éxito en la literatura. En general, esas técnicas avanzadas son capaces de lograr un alto nivel de calidad en las reconstrucciones a partir de un conjunto muy limitado de proyecciones.

En el contexto de esta tesis doctoral, el trabajo se ha centrado precisamente en la implementación de algoritmos avanzados para la reconstrucción tomográfica. El foco de atención ha sido el algoritmo TVAL3, una solución matemática para el problema de la minimización de variación total (TVM) cuyos fundamentos teóricos se encuentran en el campo del muestreo disperso. El uso de este algoritmo específico para la TE ya había sido probado con anterioridad para procesos de reconstrucción en la literatura. A pesar de ello no se disponía de una implementación funcional que no requiriese el uso de lenguajes patentados. Por ello, se fijó como objetivo la traducción de dicho algoritmo desde su formulación inicial en MATLAB a Python (libre acceso). En líneas generales, TVAL3 requiere muchas menos proyecciones para una reconstrucción precisa y muestra una mayor tolerancia a condiciones ruidosas que los algoritmos clásicos (SIRT).

Finalmente, esta metodología basada en el algoritmo TVAL3 se ha utilizado a lo largo de la tesis para la caracterización experimental de una amplia variedad de nanomateriales diferentes

La estructura de este manuscrito, y la estructura de esta tesis doctoral en su conjunto, se basa en tres pilares principales que incluyen parte de las técnicas del estado-del-arte descritas en las líneas anteriores: (1) el uso de algoritmos de muestreo disperso para la reconstrucción en TE, (2) la exploración de técnicas de agrupación y de reducción de dimensión no lineal avanzadas para el análisis de imágenes de espectroscopía EELS, y (3) el desarrollo de una nueva herramienta de software para el análisis de ELNES.

Cada uno de estos pilares, que conlleva en sí mismo la resolución de un desafío técnico, constituye un capítulo de esta tesis doctoral. Se incluye un cuarto y último capítulo que describe los resultados de un experimento clave realizado para una serie de materiales mesoporosos de ceria (CeO_2).

Algoritmo para la minimización de la variación total: Optimización de las reconstrucciones morfológicas y analíticas mediante la tomografía de electrones (Capítulo 3).

> En este capítulo se presentó una revisión histórica sobre el desarrollo de la teoría detrás de los algoritmos de reconstrucción en TE, así como de su implementación. Se incluyeron SIRT y la adaptación de TVAL3 (metodología TVM) hecha durante su traducción desde MATLAB a Python. Se demostró satisfactoriamente, y mediante el uso de sets de datos sintéticos, la ventaja que el método TVM tiene sobre SIRT cuando se trata de lidiar con experimentos con un número muy pequeño de proyecciones o con problemas patológicos de muestreo insuficiente (erosión de imágenes) y SNR bajo.
>
> El capítulo incluye la exposición de los resultados experimentales de las reconstrucciones morfológicas (mediante imágenes HAADF) y analíticas (mapas X-EDS) de un caso de estudio de una serie de haces

de nanopartículas huecas de óxidos de hierro y óxidos mixtos de hierro y níquel. Los resultados indicaron los diferentes tamaños promedio de las cortezas de estas NPs, así como la formación de segregados de Ni.

Aprendizaje automatizado para análisis de datos EELS: métodos de reducción dimensional y agrupamiento (Capítulo 4).

En este capítulo se revisaron las técnicas de reducción dimensional (DRM) más relevantes usadas para el análisis de datos EELS, incluidos los métodos clásicos de factorización matricial PCA y NMF, y los enfoques más modernos basados en el cómputo no lineal de gráficas de conectores t-SNE y UMAP. También se llevó a cabo una revisión de los algoritmos de agrupamiento más relevantes utilizados para la segmentación de imágenes de espectros EELS, incluyendo K-Means, la agrupación jerárquica por aglomeración, DBSCAN, OPTICS y HDBSCAN (este último, la adición más reciente a los métodos probados para la agrupación de datos EELS).

Esta revisión se llevó a cabo tanto desde una perspectiva teórica como desde un punto de vista experimental, probando los efectos de las diferentes combinaciones posibles de DRM y técnicas de agrupamiento. El análisis cualitativo y cuantitativo del rendimiento de dichas combinaciones resultó en la declaración de UMAP-HDBSCAN como la combinación más efectiva para la segmentación de imágenes de espectros de forma no supervisada y puramente basada en los propios datos.

Dicho análisis se hizo tanto sobre una muestra sintética de composición fija (para la evaluación de puntuaciones para cada clúster separado y comprobar la resistencia de los métodos a condiciones de ruido extremas), como sobre una muestra experimental de una nanopartícula ‘core-shell’ de óxido de hierro y óxido de manganeso. En esta última, el análisis basado en HDBSCAN y UMAP consiguió resolver de forma satisfactoria la separación de elementos entre el núcleo (de óxido de

hierro) y la corteza (óxido de manganeso), así como la aparición de 2 fases distintas de óxido de manganeso en la corteza(MnO y Mn_3O_4).

Análisis ELNES combinando la agrupación en clústeres y NLLS: la herramienta de software WhatEELS (Capítulo 5).

Este capítulo proporciona una revisión de la arquitectura del software y las principales capacidades de la herramienta de análisis programada durante la tesis y que es conocida como **WhatEELS**. Con esta solución de software se proporciona una plataforma común para usar las técnicas de agrupamiento de clústeres y los ajuste mediante NLLS, con la finalidad de llevar a cabo análisis características de ELNES en espectros de óxidos de metales de transición y tierras raras.

A parte de la descripción de la arquitectura, el capítulo incluye una serie de demostraciones paso a paso de cómo debe ser usada la herramienta, así como una descripción de todas las posibilidades de análisis de resultados a posteriori mediante el uso de los módulos incluidos (esto es, usando los módulos de análisis incluidos por defecto tras el ajuste de NLLS). La muestra de control es la misma nanopartícula de óxido de hierro y manganeso que se analizó en el capítulo anterior. Los resultados cuantitativos confirmaron lo que ya se había resuelto entonces, la separación de elementos entre núcleo y corteza, y la formación de 2 fases para el óxido de manganeso en esta última.

Análisis de la segregación de dopantes de gadolinio y praseodimio, y análisis del estado de oxidación del cerio, en ceria mesoporosa (Capítulo 6).

Este último capítulo presenta los principales resultados de caracterización de TEM para una serie de experimentos que involucran materiales mesoporosos de CeO_2 dopados con praseodimio (Pr) y gadolinio (Gd). La adopción de este tipo de materiales cerámicos para catálisis a alta temperatura y producción de combustible termoquímico

se ha demostrado con anterioridad en la literatura. Los materiales analizados en esta tesis doctoral fueron diseñados para configurar electrodos en un dispositivo de conversión de energía operando a alta temperatura (en celdas de combustible de óxidos sólidos, SOFC), dado su prometedor rendimiento como conductores iónicos y mixtos iónico-electrónicos.

Los resultados de nuestro análisis jugaron un papel clave en la comprensión del comportamiento electroquímico macroscópico a través de los fenómenos físicos nanoscópicos apoyándonos en la caracterización de la estructura de estos materiales a nanoescala.

La atención se centró en la investigación de la posible segregación de dopantes y cambios en el estado de oxidación catiónica del Ce hacia las fronteras de grano y superficies de las estructuras granulares en la ceria mesoporosa.

Estos experimentos requirieron el uso de todas las técnicas de análisis introducidas en los capítulos anteriores. De hecho, jugaron un papel central como motor para la implementación y uso de la metodología TVM para las reconstrucciones ET, la exploración de nuevos algoritmos de agrupación de clústeres, y el desarrollo de la herramienta de software WhatEELS.

Gracias a la combinación de todas ellas, se pudo resolver de forma cuantitativa los cambios de estado de oxidación del cerio localizados en las fronteras de grano y la segregación de dopantes (tanto Pr como Gd) en las mismas regiones. Todo ello, a pesar de la clara tendencia de Ce a reducirse bajo el haz (lo cual forzó la adquisición de espectros con muy bajo SNR) y a pesar del solapamiento de señales asociadas al Ce, Pr y Gd que complicaron la cuantificación elemental.

Cerramos este resumen, poniendo en contexto la importancia del trabajo realizado en esta tesis. En líneas generales, la implementación de todas estas técnicas de análisis respondió a la necesidad ineludible de nuevas y más

potentes herramientas para abordar los retos de caracterización a la nanoescala que se presentaron a lo largo del camino. Por ejemplo, el estudio de los mesoporosos de ceria puso de manifiesto que las herramientas disponibles en un principio eran insuficientes para el tipo de análisis requerido. El haber creado una herramienta gratuita y modular como **WhatEELS** permitirá en un futuro la potencial expansión del uso de estas técnicas por la comunidad científica especializada, para el análisis de nuevos e interesantes resultados experimentales.