AFRL-AFOSR-VA-TR-2023-0090

From atom probe tomography imaging to microstructural quantification: An iterative optimization approach

Marquis, Emmanuelle
REGENTS OF THE UNIVERSITY OF MICHIGAN
503 THOMPSON ST
ANN ARBOR, MI, 48109
USA

10/19/2022
Final Technical Report

Air Force Research Laboratory
Air Force Office of Scientific Research
Arlington, Virginia 22203
Air Force Materiel Command

REPORT DOCUMENTATION PAGE

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ORGANIZATION.

1. REPORT DATE	**2. REPORT TYPE**	**3. DATES COVERED**	
20221019	Final	**START DATE** 20140901	**END DATE** 20200831

4. TITLE AND SUBTITLE
From atom probe tomography imaging to microstructural quantification:
An iterative optimization approach

5a. CONTRACT NUMBER	**5b. GRANT NUMBER** FA9550-14-1-0249	**5c. PROGRAM ELEMENT NUMBER** 61102F
5d. PROJECT NUMBER	**5e. TASK NUMBER**	**5f. WORK UNIT NUMBER**

6. AUTHOR(S)
Emmanuelle Marquis

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
REGENTS OF THE UNIVERSITY OF MICHIGAN
503 THOMPSON ST
ANN ARBOR, MI 48109
USA

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)	**10. SPONSOR/MONITOR'S ACRONYM(S)**	**11. SPONSOR/MONITOR'S REPORT NUMBER(S)**
Air Force Office of Scientific Research 875 N. Randolph St. Room 3112 Arlington, VA 22203	AFRL/AFOSR RTB1	AFRL-AFOSR-VA-TR-2023-0090

12. DISTRIBUTION/AVAILABILITY STATEMENT
A Distribution Unlimited: PB Public Release

13. SUPPLEMENTARY NOTES

14. ABSTRACT
This project explored the limits and uncertainty of current structure analysis in the experimental characterization technique of atom probe tomography (APT) and explored how far the limits can be pushed with new optimized analysis algorithms. We successfully developed new approaches based on morphology, topology, cluster analysis methods, and crystallography. We developed a fully physical forward model that combines with traditional reconstruction for iterative improvement of predicted structures. We first parameterized local evaporation fields from DFT and static tip and combined it with local-field Poisson modeling. We then found the step-one approach to be insufficient and switched to full atomistic dynamics with MD simulations. Initially based on traditional, non-local evaporation fields. Finally, it became clear that including all the underlying physics was necessary, such as charging of surface atoms from the electric field, leading us to rethink a full “ab initio” evaporation physics model. This last step was not completed within this project and was proposed as a future project.

15. SUBJECT TERMS

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**
a. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	UU	38

19a. NAME OF RESPONSIBLE PERSON	**19b. PHONE NUMBER** *(Include area code)*
ALI SAYIR	426-7236

DISTRIBUTION A: Distribution approved for public release.

FA9550-14-1-0249: From atom probe tomography imaging to microstructural quantification: an iterative optimization approach

Final Project Report

Emmanuelle Marquis (PI, University of Michigan)

Wolfgang Windl (Ohio State University)

Funding Period: 09/01/2014-08/31/2019

Funding: $1,390,000

Executive Summary of Proposed Ideas and Outcomes:

This project explored the limits and uncertainty of current structure analysis in the experimental characterization technique of atom probe tomography (APT) and explored how far the limits can be pushed with new optimized analysis algorithms. We successfully developed new approaches based on morphology, topology, cluster analysis methods, and crystallography. We developed a fully physical forward model that combines with traditional reconstruction for iterative improvement of predicted structures. We first parameterized local evaporation fields from DFT and static tip and combined it with local-field Poisson modeling. We then found the step-one approach to be insufficient and switched to full atomistic dynamics with MD simulations. Initially based on traditional, non-local evaporation fields. Finally, it became clear that including all the underlying physics was necessary, such as charging of surface atoms from the electric field, leading us to rethink a full “ab initio” evaporation physics model. This last step was not completed within this project and was proposed as a future project.

Publications for the project period

1. Ghamarian, L.-J. Yu, and E. A. Marquis, Morphological classification of dense objects in atom probe tomography data, Ultramicroscopy 215, 112996 (2020). Codes freely available (DOI:10.5281/zenodo.3724970)
2. Ghamarian, L.-J. Yu, and E. A. Marquis, Quantification of Solute Topology in Atom Probe Tomography Data: Application to the Microstructure of a Proton-Irradiated Alloy 625, Metallurgical and Materials Transactions A 51, 42–50 (2020).
3. Ghamarian and E. A. Marquis, Hierarchical density-based cluster analysis framework for atom probe tomography data, Ultramicroscopy 200, 28-38 (2019). Codes freely available (DOI:10.5281/zenodo.3572572)
4. C. Oberdorfer, T. Withrow, E. A. Marquis, and W. Windl, Atomistic-Simulation Based Modeling of Atom Probe Tomography, Microsc. Microanal. 25 (S2), 284-285 (2019).
5. Y. He, D. E. Perea, S. X. Mao, C. Sang, T. Withrow, C. Oberdorfer, and W. Windl, In situ HR-TEM and Simulation of Si Field Emitter Tips under Field Evaporation, Microsc. Microanal. 25 (S2), 308-309 (2019).
6. C. Oberdorfer, T. Withrow, L. J. Yu, K. Fisher, E. A. Marquis, and W. Windl, Influence of surface relaxation on solute atoms positioning within atom probe tomography reconstructions, Materials Characterization 146, 324-335 (2018).
7. T. Withrow, C. Oberdorfer, W. Windl, and E. A. Marquis, Coupling Molecular Dynamics and Finite Element Simulations to Investigate the Nearest Neighbor Dependence of Field Evaporation, Microsc. Microanal. 23 (S1), 446-447 (2017).
8. L. Yao, A filtering method to reveal crystalline patterns from atom probe microscopy desorption maps, MethodsX 3, 268-273 (2016).
9. H. Hunter, V. Araullo-Peters, M. Gibbons, O. D. Restrepo, S. R. Niezgoda, W. Windl, K. M. Flores, D. C. Hofmann, and E. A. Marquis, Three-dimensional imaging of shear bands in bulk metallic glass composites, J. Microsc. 264, 304-310 (2016).
10. L. Yao, T. Withrow, O. D Restrepo, W. Windl, and E. A. Marquis, Effects of the local structure dependence of evaporation fields on field evaporation behavior, Appl. Phys. Lett. 107, 241602 (2015).
11. T. P. Withrow, Computational Modeling of Atom Probe Tomography, Thesis, The Ohio State University (2018).

Presentations

1. Y. He, D. E. Perea, S. X. Mao, C. Wang, T. Withrow, C. Oberdorfer and W. Windl, In-situ HR-TEM and Simulation of Si Field Emitter Tips under Field Evaporation, Microscopy & Microanalysis Meeting, Portland, August 2019
2. (Invited) C. Oberdorfer, T. Withrow, E. Marquis and W. Windl, Atomistic Simulation Based Modeling of APT, Microscopy & Microanalysis Meeting, Portland, August 2019
3. Ghamarian, E.A. Marquis, Introduction of the DSF software package for solute analysis. Cameca User workshop, Madison, WI, June 2019
4. (Invited) C. Oberdorfer, T. Withrow, L. Yao, EA Marquis, W. Windl. TAPSim Hands-On Tutorial, EU-APT Workshop, Düsseldorf, Germany, August 2019
5. (Invited) C. Oberdorfer, T. Withrow, L. Yao, E.A. Marquis, W. Windl. APT Simulations - Beyond Electrostatics. Opening symposium for advanced S/TEM and APT facilities at the Max-Planck- Institute for iron research, Düsseldorf, Germany, July 2019
6. C. Oberdorfer, T. Withrow, L. Yao, E.A. Marquis, W. Windl. Atomistic Simulation of Field Evaporation from Field Emitter Tips, MS&T 2018, Columbus, October 2018
7. W. Windl, C. Oberdorfer, T. Withrow, L. Yao, E.A. Marquis, Atomistic-Modeling Based Simulation of Field Evaporation Processes, APT&M, Gaithersburg, MD. June 2018

8. C. Oberdorfer, T. Withrow, L. Yao, E.A. Marquis, W. Windl. Biased solute reconstruction due to athermal surface drag, APT&M, Gaithersburg, MD, June 2018
9. W. Windl, C. Oberdorfer, T. Withrow, L. Yao, E.A. Marquis. Advanced APT Simulations by Combining Electrostatics with Molecular Dynamics, TMS, Phoenix, TX, March 2018
10. C. Oberdorfer, T. Withrow, L. Yao, E.A. Marquis, W. Windl.. Non-Diffusive Drag-Effect in APT of AlCu Alloy, TMS, Phoenix, TX, March 2018
11. (Invited) E.A. Marquis, C. Oberdorfer, T. Withrow, L. Yao, W. Windl, Progress in modeling atom probe tomography data, Frontiers of Electron Tomography, National Center for Electron Microscopy, Lawrence Berkeley Laboratory, October 2017
12. T Withrow, C Oberdorfer, E Marquis, W Windl, Coupling Molecular Dynamics and Finite Element Simulations to Investigate the Nearest Neighbor Dependence of Field Evaporation, Microscopy & Microanalysis, St Louis, MO, August, 2017
13. (Invited) Windl, Oberdorfer, Withrow, Yao, Marquis, workshop on High electric Fields in Electrochemistry and in Atom Probe Tomography, Germany, March 2017
14. (Invited) E Marquis. Traditional clustering analysis in APT data. European workshop on atom probe tomography, Oxford UK, September 2016.
15. W Windl, T Withrow, O Restrepo, L Yao, YX Zhang, and E Marquis. From atom probe tomography imaging to microstructural quantification. Seminar, TU Chemnitz, June 2016.
16. (Invited) W Windl, Novel Approaches to Validated Modeling of Defects and Surface Processes. Workshop on Enhancing International Collaborations on Emerging Materials for Defense Applications via Innovative Theory, Simulation, and Experiment, London, UK. June 2016
17. (Invited) E Marquis. Did you say clusters? Bernkastel-Kues, Germany, 23-25 May 2016.
18. E Marquis. A Round Robin Experiment: Analysis of Solute Clustering from Atom Probe Tomography Data. Microscopy & Microanalysis, Columbus OH, July 2016
19. (Invited) E Marquis, L Yao, O Restrepo, W Windl. From atom probe tomography imaging to microstructural quantification I – Modeling field evaporation. Third international congress on 3D materials science (3DMS), St. Charles, IL, July 2016.
20. L. Yao, A. McFarland, E. A. Marquis, T. Withrow, O. D. Restrepo, and W. Windl, DFT Simulations of Atom Probe Tomography, Workshop on Characterization and Modeling, Bernkastel, Germany, May 2015.
21. W. Windl, O. D. Restrepo, T. Withrow, E. A. Marquis, L. Yao, Modeling-Enhanced High-Resolution Microscopy: From 2D Materials to 3D Atom-Probe Tomography, Forschungszentrum Jülich, July 2015.
22. T. Withrow, O. D. Restrepo, L. Yao, A. McFarland, E. A. Marquis, and W. Windl, An Ab Initio Method for Improving Atom-Probe Tomography Reconstructions, Hayes Research Forum, OSU, February 2015.
23. T. Withrow, O. D. Restrepo, W. Windl, L. Yao, A. McFarland, and E. A. Marquis, Ab Initio Simulations to Improve Atom-Probe Tomography Reconstructions, MS&T 2015, Columbus, OH October 2015.

People

- Christian Oberdorfer (Postdoc and Humboldt Fellow, OSU, 2017-2019)
- Travis Withrow (MS May 2018, OSU)
- Lan Yao (Postdoc, UM, 2015-2016)
- Iman Ghamarian (Postdoc, UM, 2017-2019; currently faculty in the Aerospace Engineering Department at the University of Oklahoma)

Project Details

1 - Local evaporation fields parameterized from DFT and static tip, combined with local-field Poisson modeling

Tungsten

We previously showed with aluminum that a simulation which combines the applied field and local structure effects reproduces experimental data better than a model that only takes into account the applied field. A simple nearest neighbor counting model for determining evaporation order has the advantage that it provides more information about the evaporating atom without requiring significantly more computational power than a simulation that only considers the applied field strength. For simple pure FCC metals, neighbor counting provides enough information to determine the structure-dependent zero barrier evaporation field and allows for distinguishing between surface sites in the simulation. Extending this model to materials that are both easier and more interesting to observe in the atom probe is the next step for building a new forward simulation.

The next systems we targeted were pure tungsten evaporation and evaporation of an aluminum copper alloy. Tungsten was chosen for its popularity in atom probe simulations and aluminum copper was chosen as an extension of the already finished aluminum work. For pure tungsten, the aluminum calculations were repeated, and the data was analyzed for a fit to a simple neighbor model.

Calculations were done on (100), (110), (111), and (112) surfaces for tungsten. These surfaces give a range from close-packed to “open” BCC surfaces and at least one common W atom probe tip direction, where more open surfaces have more distance between surface nearest neighbors. Calculation results give BCC zero barrier evaporation fields of 53 ± 3.18 V/nm for the 3+ ion, which matches well with the value from Tsong of 52 V/nm2 and falls within the experimental range of 47-59 V/nm1. From these calculations the likely order of ionization charge for a given evaporation is +3, +2, +4 and finally +1, which matches the order in Tsong[2].

The Müller model[3] is used to calculate the zero barrier evaporation field (ZBEF) for the adatoms. The energy of a single Tungsten atom in vacuum, ionization energy, energy of surface with the adatom, energy of the surface without the adatom, work function and Fermi level are all calculated using VASP. The heat of evaporation, λ, is then: (Energy with adatom) – (Energy without adatom) + (Energy single atom in vacuum). The work function is the difference between the vacuum energy and the Fermi level of the surface. The ionization energy and energy of a single atom in the vacuum are both calculated by placing a single atom (charged and uncharged) in progressively larger cells then fitting the total energy to an equation of the form $E(a) = E_{\infty} + E'/a$ and extrapolating to a = 0. Similar to aluminum, different surface configurations were chosen for the different surface orientations and the zero barrier field calculated for selected atoms.

---

[1] Kellogg, G.L., Measurement of activation energies for field evaporation of tungsten ions as a function of electric field. Physical Review B, 1984. 29(8): p. 4304-4312.
[2] Tsong, T.T., Field-Ion Image-Formation. Surface Science, 1978. 70(1): p. 211-233.
[3] Müller, E.W., Field Desorption. Physical Review, 1956. 102(3): p. 618-624.

Attempting to fit these evaporation values to the characteristics of the perfect structure proved difficult. Initial nearest neighbor only fitting out to 7 nearest neighbors:

$$ZBEF = \sum_{i=1}^{7} a_i nn_i \qquad (1)$$

was unable to reproduce the correct evaporation order. This can be seen in Figure 1, a plot of the DFT evaporations fields for each structure indexed by the evaporation order the fit produces. If the fit reproduces the correct evaporation order the plot should look like a monotonically increasing function. Lower values to the right of higher values indicate a wrong prediction of evaporation order.

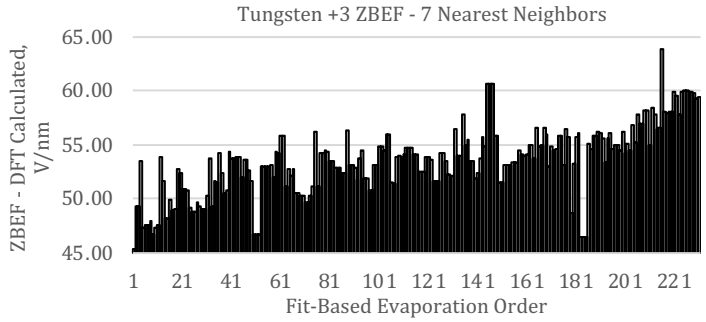


Figure 1: Tungsten evaporations indexed by fit order. Smaller values to the right of larger values indicate the fit chooses evaporated atoms in the wrong order vs the DFT data. Fit is of the form given in Equation 1.

Including angle counting terms improves the fit slightly. The angles that the adatom's nearest neighbors form with one another using the adatom as the center are fixed for BCC, and can be used to add another degree of freedom to distinguish one structure from another. The fit is of the form:

$$ZBEF = \sum_{i=1}^{7} a_i nn_i + \sum_{j=1}^{n} b_j nn_1 a_j \qquad (2)$$

Where the nna term represents the number of 70.5, 109.5, 108 degree angles that the 1nn form with the adatom at the center. Including these angles produces fits that reproduce the evaporation order slightly better but are still not useable. Figures 2 and 3 show the comparison between using only the 1nn angles and the 1nn, 2nn and 3nn angles in a fit.

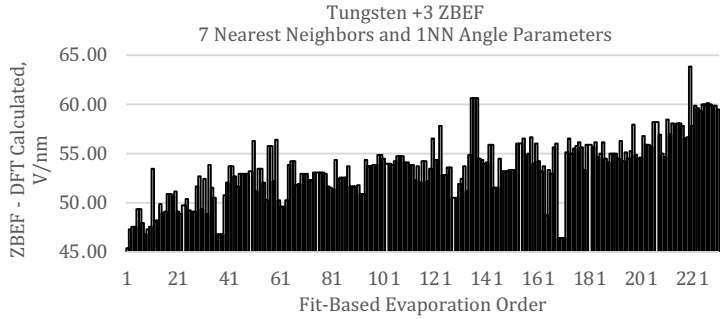


Figure 2: DFT Evaporation ZBEF indexed by fit order. Smaller values to the right of larger values indicate the fit chooses evaporated atoms in the wrong order vs the DFT data. Fit is of the form in Equation 2.

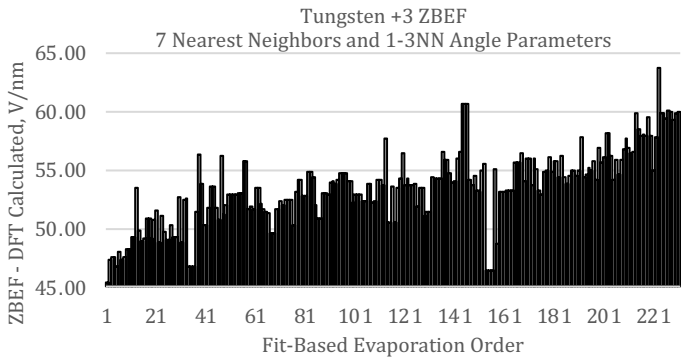
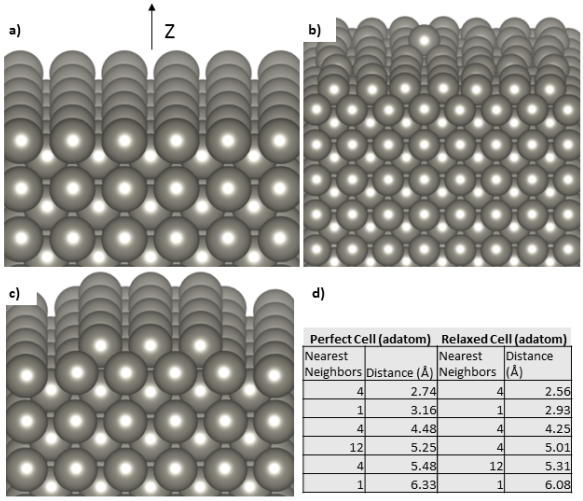


Figure 3: DFT Evaporation ZBEF indexed by fit order. Smaller values to the right of larger values indicate the fit chooses evaporated atoms in the wrong order vs the DFT data. Fit is of the form in Equation 2 but includes terms for nn2a and nn3a.

As can be seen from the plots, neither fit for tungsten ZBEF reproduces the correct evaporation order for most of the DFT calculated evaporations. The likely culprit for this is surface distortion that tungsten undergoes when adatoms are present. The tungsten (100) surface reconstructs only when adatoms are present, documented in the literature for the presence of adsorbed hydrogen. In our calculations the same thing appears to occur where (100) surfaces stay in the crystalline configuration (Figure 4a) but (100) surfaces with absorbed atoms will distort (Figure 4b), sometimes enough to change the nearest neighbor counts of the adatom from the perfect configuration somewhat drastically. This is shown by the (100) adatom nearest neighbor counts, where the 4th and 5th (perfect) nearest neighbors switch in the relaxed cell as a result of the adatom caused surface distortion. Other (100) adatom configurations with greater surface coverage, such as the step cell (Figure 4c), show no distortion. Presumably these structures act like a complete surface rather than surfaces with adatoms. Figure 10 shows the difference in (100) surface distortion and the neighbor switching behavior for the tungsten single adatom 4th and 5th nearest neighbors.



Perfect Cell (adatom)		Relaxed Cell (adatom)	
Nearest Neighbors	Distance (Å)	Nearest Neighbors	Distance (Å)
4	2.74	4	2.56
1	3.16	1	2.93
4	4.48	4	4.25
12	5.25	4	5.01
4	5.48	12	5.31
1	6.33	1	6.08

Figure 4: Tungsten surface distortions and nearest neighbor switching. All cells have been relaxed to their lowest energy configuration by DFT, are periodic in x and y and have 15 Å vacuum in the z direction. a) Perfect (100) tungsten surface exhibits no distortion after relaxation. b) Single adatom causes long range distortion on (100) surface. c) Step structure results in no surface reconstruction d) 4th and 5th nearest neighbor switching for the single adatom after relaxation. The adatom partially buries itself into the surface (resulting in both the 1nn and 2nn distance being closer) which

results in it being closer to the four bulk atoms (5nn in the perfect structure) and surface distortions push the four surface atom (4nn in the perfect structure) further away from the adatom.

In comparison the (110) surface, a more closely packed BCC plane, shows much less distortion. Nearest neighbors do change for less stable adatom configurations, but the change is not as drastic as seen in the (100) case. Figure 5 shows examples of the (110) surface undergoing only slight distortion.

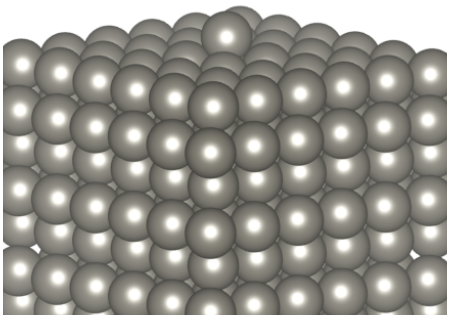


Figure 5: Tungsten (110) relaxed surface with single adatom. The adatom does not cause distortion of the more closely packed surface plane. Vacuum direction is directly up, normal to surface with adatom.

For tungsten the distortion of the surface caused by the adatom configurations makes simple models that do not consider relaxation unreliable and not representative of reality. DFT data can still be used to inform the ZBEF of the relaxed structure adatoms, but surface relaxation of a simulated tip is required to accurately predict the true local structure and therefore the local structure dependent ZBEF of the atoms on the tip.

Aluminum Copper Alloys

Field evaporation of an aluminum alloy was another target for extending our simple neighbor counting model to other systems. Initial calculations show a strong dependence of ZBEF on the local concentration around an adatom. For evaporation of an Al adatom from a pure Al surface the ZBEF is 16.5 V/nm, but 18.1 V/nm when Al is evaporated from a pure Cu surface. Determining the cause of this behavior and creating a model to describe the behavior of Al evaporating near Cu atoms.

To begin with, calculations for AlCu were carried out similarly to both W and Al with everything calculated using DFT and the Müller model used to calculate the ZBEF for the given evaporation. Calculation cells of dilute alloys ranging from 3-10% Cu with Al evaporations were considered. Nearest neighbor counts for both species of atoms were recorded, and an equation of the form:

$$ZBEF = \sum_{i=1}^{13} a_i nn(Al)_i + b_i nn(Cu)_i \quad (3)$$

was fitted to low Cu concentration data. Figure 6 shows the results of that fit.

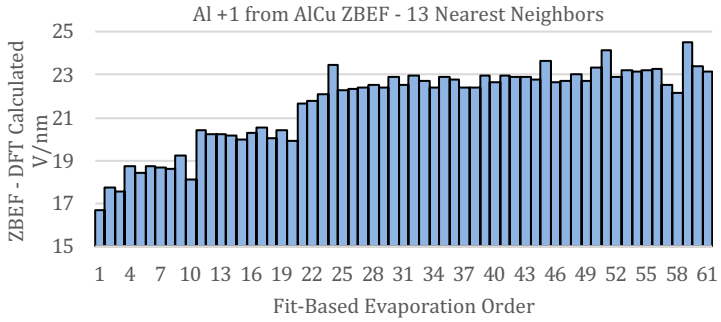


Figure 6: DFT Evaporation ZBEF indexed by fit order. Smaller values to the right of larger values indicate the fit chooses evaporated atoms in the wrong order vs the DFT data. Fit is of the form in Equation 3.

The evaporation order is not preserved well by this fit – even though it considers out to 13 nearest neighbors and their types. The interaction between an Al adatom and Cu in the bulk is more complicated than a simple fit to dilute concentrations can account for. To investigate this relationship further and see if a limited-range neighbor fit was sensible, a maximum interaction radius for an Al adatom needed to be established.

To determine how many Al shells would be required to completely screen an aluminum adatom from a local copper atom in solid solution, the layer dependent behavior of Al ZBEF vs distance from a single Cu atom was calculated for several structures. For these calculations a single Cu atom was placed on an Al lattice site a certain distance away from the adatom. Distances out to 12.15 Å were considered, the maximum possible in a calculation cell that would finish in a reasonable amount of time. Figure 7 shows the setup and calculation data.

These DFT computation cells were not large enough to allow for enough Al to completely isolate the effect of a nearby Cu atom, especially from the effects of a Cu on the surface. None of the DFT calculated ZBEF vs 1st Cu distance converge to the ZBEF of pure aluminum particularly well, so larger cells would be needed to determine the interaction radius. Because of this, we turned to classical molecular dynamics (in LAMMPS)4 to evaluate whether or not an interaction radius could be better defined. Figure 8 uses the same setup as Figure 7, but with a much larger cell. The distance of the second nearest Cu is never closer than 20 Å to the adatom.

For a Cu not on the surface, MD shows a cutoff of around 12 Å for the interaction distance between the adatom and the Cu atom. This is somewhat expected since large distortions do not occur for Cu in bulk Al, and the cutoff radius for the potential is 6.4 Å, so at twice the cutoff radius the adatom and first Cu are no longer interacting. For the case of a Cu on the surface, however, even out to 19 Å the surface Cu still influences the ZBEF of the adatom. The reason for this long-range interaction is a strong surface distortion by the Cu that changes the configuration of the adatom's nearest neighbors. Figure 9 shows how, after relaxation, the Cu distorts surface atoms that are close to the adatom.

4 Plimpton, S., Fast Parallel Algorithms for Short – Range Molecular Dynamics. Journal of Computational Physics, 1995. 117(June 1994): p. 1-19.

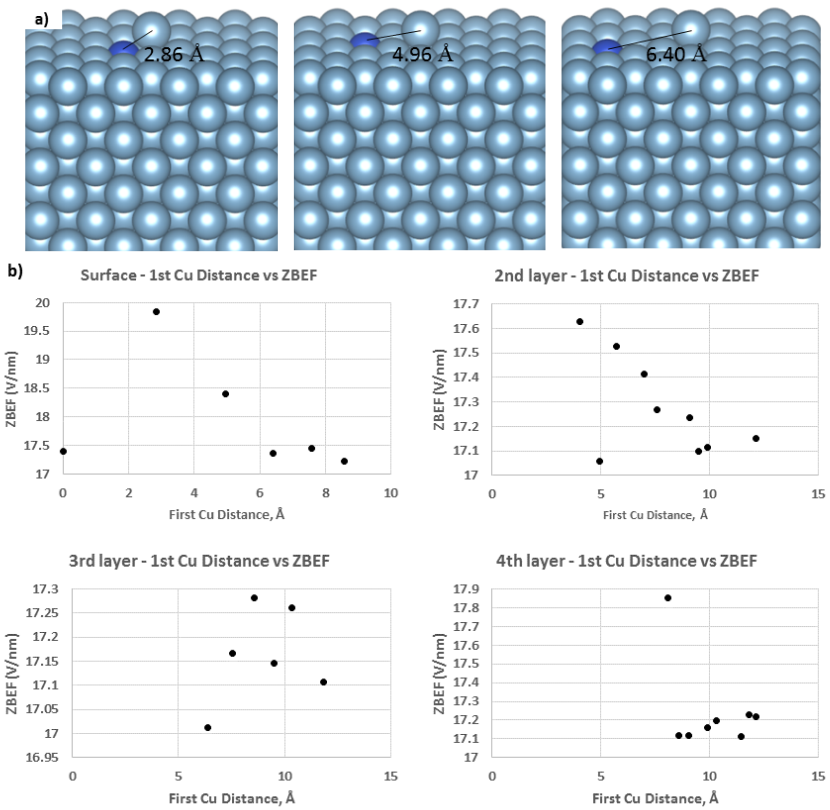


Figure 7: a) Setup for interaction radius calculation. The Cu atom (dark blue) is placed on a lattice site at different distances from the Al adatom and the ZBEF of the adatom is calculated. All cells are periodic in x and y and have 15 Å vacuum in the z, normal to the surface containing the adatom. b) Data from ZBEF calculations as a function of Cu-adatom distance for different layers in the cell. The Surface is the plane of atoms just below the adatom, with the 2nd-4th layers being the consecutive atomic planes below the surface.

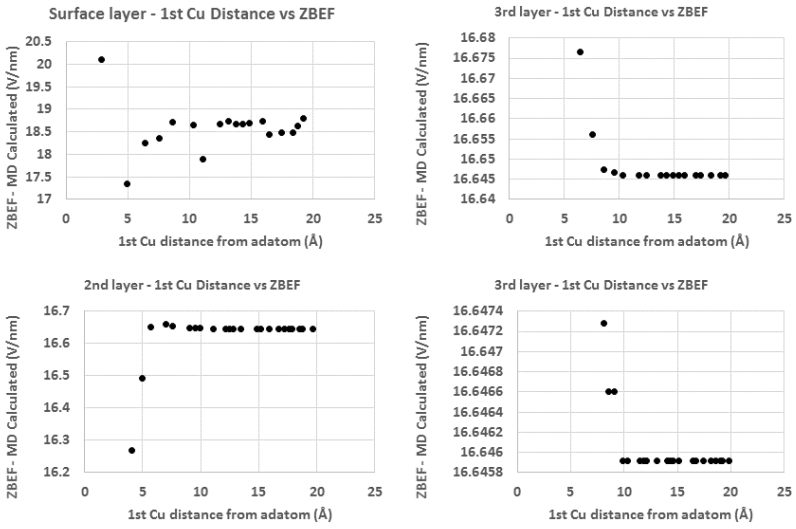


Figure 8: LAMMPS data for ZBEF of an adatom vs distance to first Cu atom. Calculation setup is the same as described in figure 7, but with a larger calculation cell to allow for placing the Cu at further distances from the adatom.

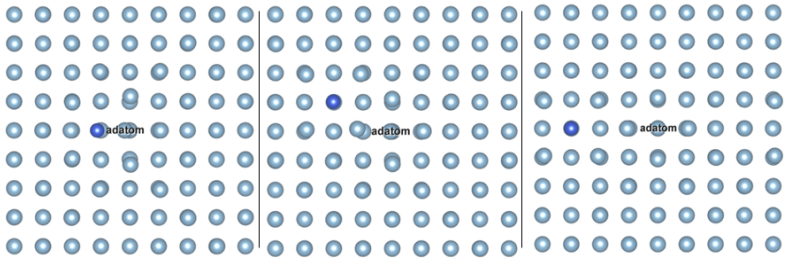


Figure 9: DFT relaxed (100) Al cells with one Cu atom at the surface showing surface distortion caused by the Cu atom (dark blue). View of cells is directly down the z direction; vacuum is along z direction (same cells in Figure 7a. Even when the Cu atom doesn't appear to directly interact with the adatom (right) surface neighbors of the adatom are still distorted by the Cu atom. This causes the changes in the ZBEF seen in Figure 7b.

Just as in tungsten, aluminum copper shows somewhat severe surface distortions that influence the ZBEF for an adatom strongly enough to change its evaporation order. Past the point that the adatom is interacting directly with the Cu atom, the distortion caused by the Cu atom causes surface atoms to pull away from the adatom changing its local environment and thus its evaporation field. This is once again a problem for a simple fit that does not include the effects of surface relaxation. MD can be used to help alleviate this, but the effects of a close copper on the ZBEF are not the same as in DFT. Some of this is due to the second nearest Cu (across the periodic boundary) being much closer in DFT than in MD due to the smaller cell size in DFT.

Accounting for the surface relaxation can be done by introducing explicit relaxation in the simulation of the atom probe tip. With DFT data for the relaxed structures, local structure dependent ZBEF calculations can be done either with MD or, ideally, with a simple fit to the relaxed neighbor data.

Outlook

Coupling explicit MD relaxation with the evaporation simulation is the next step which we are currently implementing. For this, we can draw from the previous experience of Dr. Christian Oberdorfer, a APT and field evaporation modeling expert from the University of Stuttgart, who just joined the Windl group as a postdoc. He is currently working on combining the MD capabilities in LAMMPS with electric field modeling and evaporation simulation codes. This should remedy the current problem that relaxations present in non-fcc elemental single crystals, which is most situations one would be interested in.

In parallel, we are trying to find multicomponent systems where relaxations are small enough that the previously developed neighbor counting model can be applied. For that, we consider alloys with small lattice mismatch and intermetallic compounds that have stable surfaces. Our current candidate materials are CuNi alloys, with a mismatch of less than 3%, and intermetallics such as Ni_3Al.

2 - Molecular Dynamics with Electric Field

As a first milestone on the path to advanced physically sound APT simulations, we extended the traditional finite-element (FE) based electrostatic simulation approach with a molecular dynamics (MD) calculation performed with LAMMPS. We demonstrated the importance of the capability for mechanical relaxation of the atom positions by means of the MD extended simulation approach, which enabled to understand the experimental observation of solute segregation along zone lines through the drag effect.

However, the most important factor, the electric field, is to date not fully included in these simulations. Atoms are removed from the surface by simply postulating that evaporation from a surface site with a high effective field is more likely than from another site with a lower effective field ("effective field" here means the ratio between the applied field in the simulation and the specific material dependent evaporation field which is for most elements between 19 V/nm (Al) and 57 V/nm (W). This approach disregards the mechanical effect of the field in form of the Maxwell force completely.

The second milestone, which we are working on currently, is therefore to include the field-induced forces on the emitter atoms directly in the conducted MD calculations. Figure 10(a) shows as an example the current state of progress on this path. Due to the additionally applied Maxwell force the whole emitter structure is affected rather than just the surface atoms. Eventually, in case the applied field is sufficiently high, the induced force may exceed the local bonding for some surface atoms, which then results in desorption of the atoms from the surface (Figure 10(b)). Our first preliminary simulations showed that the field, which is necessary to trigger the field evaporation, is high enough to cause significant strain in the tip structure, leading to large displacements of the atoms. Numerically, these large displacements push the simulations outside of the validity range of the static FE mesh approach, which we have been using to solve for the field for static tips in non-Maxwell force MD simulations. Specifically, in the full-field simulations, at times the inner part of the mesh that describes the emitter structure overlaps with the outer vacuum part. At this point, the simulation collapses and is no longer meaningful. For that reason, we started to devise an advanced meshing approach which allows for dynamic adaptation of the mesh and fine-grained control of the local mesh size during the simulation in response to the changing emitter under the high mechanical load of the field.

Automated Meshing Approach

In contrast to the situation with the static mesh, the automated meshing approach starts the simulation from the emitter structure as the only input. The atoms in the emitter are defined by their positions in 3D space, atom ids and an additional new tag. This tag marks an atom as either "surface" or "bulk" and allows the meshing of non-convex input shapes (e.g. a surface “dip” at a grain boundary) or even emitter structures with holes which we tried as a test case (Figure 11). The mesh itself is then computed by filling regions outside of the tip bounded by a rectangular box with random vertices following a preset minimum distance until no further vertices can be added.5 Adaptive control of the mesh size is possible by paving the simulation domain with differently sized boxes. Depending on the position of these boxes within the domain and the distance to the emitter structure, respectively, the interior mesh size is adjusted accordingly. During the simulation, once

5 Ebeida et al.; A simple algorithm for maximal Poisson-Disk sampling in high dimensions; Eurographics 31 (2012) pp. 785-794.

the initial field solution has been computed, a finer tiling with such boxes is applied proportional to the changing of the field. Naturally, this results in a small box sizes at the emitter surfaces and larger box sizes farther apart. In the same way the local mesh can be recomputed in response to a changing emitter shape if necessary.

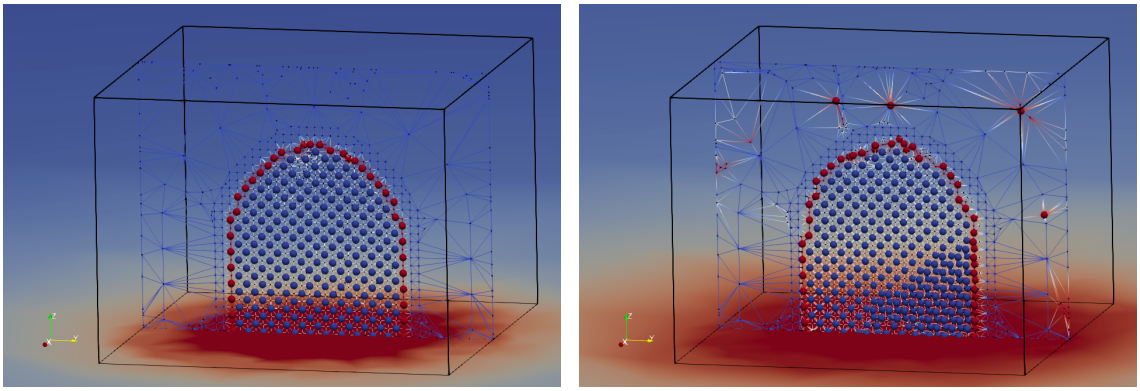


Figure 10 Slice through an APT tip from LAMMPS MD simulations with added field-induced forces. The overlay in the background shows the field solution obtained from the static FEM mesh in the foreground. Spheres depict atoms of the emitter tip (blue bulk, red surface). (a) At very low field the resulting atomic positions are still coincident with the surrounding mesh. (b) At elevated field the emitter lattice is severely strained, and atoms desorb from the surface.

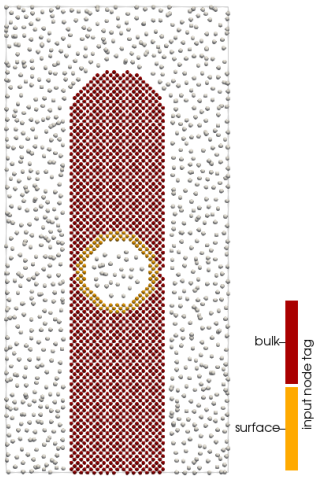


Figure 11: Result from the automated mesh generation approach. In the case of a non-convex shape or a complex emitter structure containing voids, assigning “bulk” and "surface" tags to the atoms at input enables distinction between outer and inner regions. Subsequently, empty space enclosed by surface atoms (yellow spheres) and within a predefined bounding box is filled with additional nodes (light-grey spheres) at random position that follow a preset minimum distance.

Delaunay Refinement for Adaptive Meshing

The first tests of the automated meshing approach as described above were successful. However, these tests immediately revealed that to have control over the numerical

accuracy, additional control over the finite-element quality (the shape of tetrahedron in the basic 3D Delaunay tessellation) is needed. For that reason, current efforts are made to extend the functionality of the devised custom Delaunay mesh generator with the needed mesh refinement abilities.[6]

Site-Specific Evaporation Fields for Alloys

In our first phase of the project, site-specific evaporation fields were calculated with Density-Functional Theory (DFT) first for Al, and in the last year for other elements (Figure 12), and the energies were fitted to surface-neighbor counts to create fast look-up tables suitable for APT simulations. While the resulting fit was reasonably good for Al, for some other elements, and especially for all alloys, that was not the case. To overcome this problem, starting from the idea that in a metal bond strength decreases with the number of neighbors, we fitted the per-bond energy of the studied evaporations as a function of neighbor number to a simple exponential function that yields much better results (Figure 13). The results of these DFT calculations suggest that for some alloyed elements, namely Mg when alloyed in Al, significant reduction of the ZBEF of the solute may occur. The weaker binding energy also leads to longer surface bonds, meaning Mg will be naturally more exposed, leading to a higher probability of undergoing evaporation before reaching a terrace edge. Predicting this type of situation is important so that it is not convoluted with other artifacts, such as the drag effect.

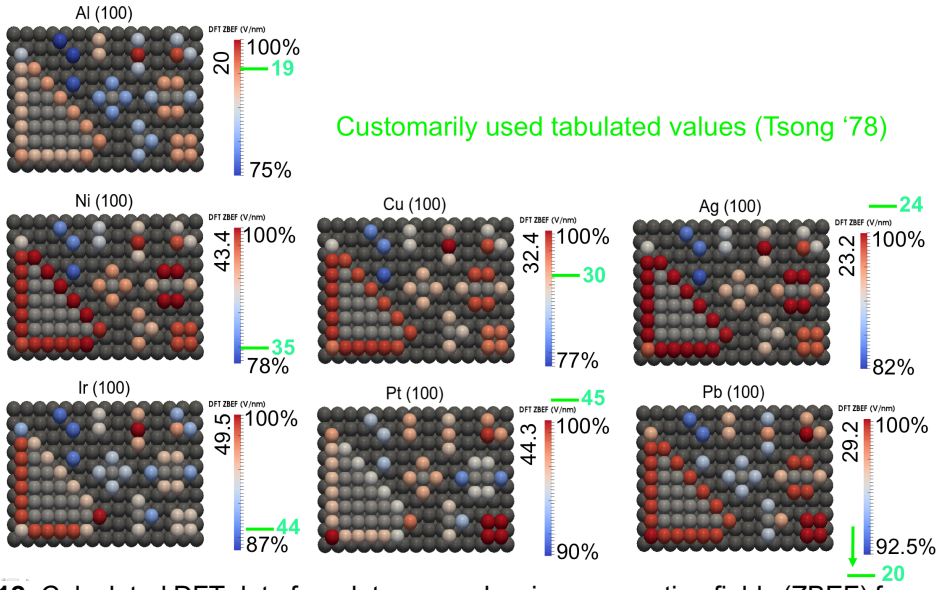


Figure 12: Calculated DFT data for adatom zero-barrier evaporation fields (ZBEF) from different surface sites for all elemental FCC metals considered for study. All values for ZBEF are given in V/nm (vertical numbers), with scaling depending on the range of the actual computed field (range reported in % of the maximum value). In comparison, customarily used values from the literature (Tsong '78) are shown in green, which for some elements are outside of the calculated range. Comparison between surfaces should be done carefully, as colors are not universally scaled.

---

6 R. Shewchuk, Tetrahedral Mesh Generation by Delaunay Refinement, Proceedings of the fourteenth annual symposium on Computational geometry, p. 86-95, Minneapolis, MN (1998).

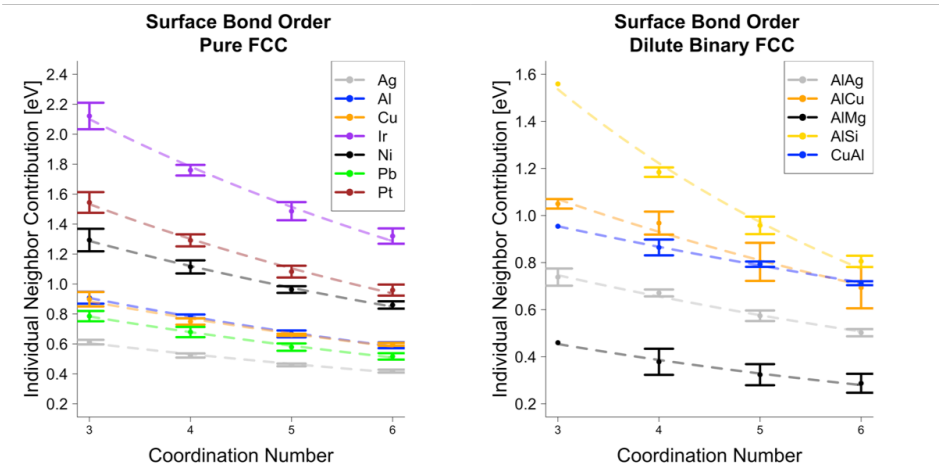


Figure 13: Results of fitting an exponential function to DFT calculated binding energy per bond for the different adatom configurations studied. Results for both pure (left) and dilute binary calculations (right) are shown.

3 - Data quantification

APT data quantification is a multilayered problem involving quantification of reconstruction errors and quantification of the microstructures once reconstructed. We initially focused on the latter in this project. Materials characterization finds its purpose in the understanding and quantification of the relationships between processing parameters, microstructure, and properties. Standardized quantitative methods and algorithms have been developed for many characterization techniques, such as extracting grain size information from metallography data, or precipitate size and shape from electron microscopy imaging. The technique of atom probe tomography (APT), on the other hand, is still waiting for standardized, reproducible, and quantitative methods to be developed and adopted by the APT community. In this project, we focused on the quantification of solute clusters, a long-standing issue within the APT community. The PI had demonstrated the significant limitation in existing cluster detection algorithms as applied to APT datasets, in which the arbitrary nature of user-selected parameters leads to very high variability in reported cluster analysis results.[7]

Hierarchical Density-Based Cluster Analysis Framework for Atom Probe Tomography Data[8]

Atom probe tomography (APT) that uniquely combines high spatial and chemical resolution in three dimensions, has been used extensively to study spatial distributions of solute atoms within solid solution and phase-separating alloys [1]. Whether solute atoms are randomly positioned or follow particular spatial correlations due to chemical ordering, clustering, or precipitation, they can dramatically impact mechanical properties. Past literature provides many examples of APT reconstructions and analyses of solute clusters in a wide range of alloy systems [2], including the early stages of solute clustering in alloys during thermal aging or under irradiation conditions. No other technique can currently surpass the ability of APT to create real space 3D rendering of atomic positions and nanometer-scale features. Therefore, quantification of such solute distributions from APT data is essential to further our understanding and to design optimized alloy structures. Accordingly, a reliable, quantitative cluster-finding method for APT datasets is required. Here, we use the term *cluster* for a group of dense atoms, which in terms of microstructure could mean non-equilibrium solute clusters, GP zones, or thermodynamically stable precipitates.

Identification of clusters in point cloud data is generally not a trivial task. For the analysis of APT data specifically, several methods were developed to quantify solute clusters. These include statistical nearest neighbor analysis [3], Voronoi partitioning [4], frequency distributions [5, 6], pair correlation functions [7, 8], Delaunay tessellation [9], maximum separation method (MSM) [10], envelop method [11], core-linkage [12], iso-concentration surfaces and proximity histograms [13], and Gaussian mixture model [14]. However, most

[7] Atom Probe Tomography Interlaboratory Study on Clustering analysis in experimental data using the maximum separation distance approach, Y Dong, A Etienne, A Frolov, S Fedotova, K Fujii, Koji Fukuya, C Hatzoglou, E Kuleshova, K Lindgren, A London, A Lopez, S Lozano-Perez, Y Miyahara, Y Nagai, K Nishida, B Radiguet, DK Schreiber, N Soneda, M Thuvander, T Toyama, J Wang, F Sefta, P Chou, EA Marquis, Microscopy & Microanalysis. (2019) 25 356-366

[8] Published as: Hierarchical Density-Based Cluster Analysis Framework for Atom Probe Tomography Data, I Ghamarian, EA Marquis. Ultramicroscopy. (2019) 200 28-38. Codes: DOI:10.5281/zenodo.3572572

of these cluster analysis methods are not robust to the complexity and spatial heterogeneities associated with APT data and their applications are limited to very specific microstructures.

The MSM and variants using concentration filtered data [15-17] have been widely used within the APT community for their algorithmic simplicity and availability within commercial software packages such as IVAS^{TM} [16, 18]. Unfortunately, the application of the MSM is not always warranted as the MSM only works well for systems with a large solute concentration contrast between matrix and clusters. Applying this method to microstructures that do not fulfill the associated assumptions can lead to erroneous results [19]. Additionally, even when the assumptions are satisfied, the selection of values for the parameters required by the MSM can be problematic [20-22]. Various approaches have been proposed to select optimized values for the maximum distance (D_{MAX}) defining the required proximity of solutes atoms within a cluster and the minimum number of atoms (N_{MIN}) in a cluster [21, 23-26]. Yet, defining the accuracy of MSM analyses remains an open challenge [27]. More recently, another clustering algorithm was proposed using a Gaussian mixture model to probabilistically differentiate clusters from matrix [14]. While this method can identify clusters more accurately in comparison to the maximum separation method, it is limited to detecting spherical clusters in a very dilute matrix. Beyond the clustering algorithms themselves, the limited detection efficiency of the delay-line detectors used in atom probe tomography instruments, the user interpretation of the mass spectrum [28], various positioning artefacts introduced during data acquisition or reconstruction, e.g. [29-31], add further complexity in the interpretation of the identified spatial clustering and its relation to the original microstructure.

Ideally, a solute cluster-finding algorithm tailored to APT data should maximize the detection ratio of actual clusters to spurious clusters, be independent of cluster morphology, have the ability to quantify the shape of clusters/precipitates, and be able to handle large datasets in a limited amount of time [9]. A robust cluster analysis method should also require as little subjective user input as possible, which means that cluster characteristics should be as implicit as possible. Luckily, such requirements have already been addressed in other scientific fields. Hierarchical clustering representations were developed to minimize user input and facilitate cluster identification through dendrogram plots [32]. Among all hierarchical clustering and classification algorithms [33], that developed by Hartigan determines clusters by finding the maximum connected components of the high density regions associated with the probability density function representing the distribution of points in a dataset [34, 35]. Consequently, clusters are regions in which the ratio between the probability that a group of neighboring atoms forms a cluster and the volume of this cluster is large. The algorithm resolves the issue of estimating an unknown probability density distribution and its relation with clustering, quantifies clustering performance [36] and provides interpretable dendrogram plots.

We proposed to apply hierarchical clustering methods and developed the following workflow. Initially, the Hierarchical Density-based Spatial Clustering of Applications with Noise (HDBSCN) Python package [37] is used to find clusters in a conservative manner (e.g., nearby clusters may be merged), then the Density-Based Clustering (DeBaCl) Python package [35] is used to further analyze each HDBSCAN-detected cluster and if required, prune (i.e., divide) each identified cluster into some smaller, real clusters. While DeBaCl is a faster and more efficient algorithm than HDBSCAN, it requires significant memory in comparison to the HDBSCAN algorithm, and this memory requirement would be intractable for the typical size of APT datasets. This is one of the reasons why

HDBSCN is first used to create subsets of the data that can be further analyzed by DeBaCl. Finally, the k-nearest neighbors (k-NN) algorithm is offered as an option to assign atoms considered as noise by DeBaCl to one of the detected clusters. The developed code for the current study is available online [38].

Methodology

We first briefly review the three algorithms used in the proposed cluster analysis method, and discuss the implementation of these algorithms and the associated parameters. Four synthetic datasets were created to test the performance of the method and compare the outcomes with that of the traditional MSM.

- **Hierarchical Density-based Spatial Clustering of Applications with Noise (HDBSCAN)**

The original Density-based Spatial Clustering of Applications with Noise (DBSCAN) algorithm is a widely-used cluster analysis algorithm due to its robustness to the presence of noise in a dataset [39]. Note that in the context of APT data, solute atoms that are not part of a cluster are called noise. This algorithm requires the subjective selection of a critical parameter (ε) that defines how close points should be to each other so that they are considered members of a cluster. The algorithm is also unable to detect clusters with different densities. These two problems were addressed in the HDBSCAN algorithm [40, 41] that analyzes clusters by applying DBSCAN over a range of ε values and determining the clusters that are stable (persistent) with respect to changes in values of ε. In addition, recent implementations have significantly accelerated the algorithm [37, 42-44], making it attractive for large APT datasets.

Statistical, computational, and topological reviews of HDBSCAN can be found in the literature [37]. Here, only the computational aspects of HDBSCAN are reviewed briefly. The two parameters required in the standard DBSCAN algorithm are a distance scale, ε, and a minimum number of points, k, that expresses the density threshold. Following the terminology used by Campello et al. [41], a point X_i (that is a member of a set of $X=\{X_1,\ldots,X_n\}$ of points within a metric space ($\mathbf{X}$, d)), where d is the Euclidian distance, is called a core point if the number of points within ε is at least equal to k, **Eq. 1**.

$$|B(X_i,\varepsilon)\cap X|\geq k \qquad (1)$$

Here, B is defined as the open ball of radius ε about X_i. Also, two arbitrary points X_i and X_j are considered *ε-reachable* with respect to ε and k if **Eq. 2** and **Eq. 3** are satisfied. These two points are called *density-connected* if they are either directly or by some connections ε-reachable.

$$X_i \in B(X_j,\varepsilon) \qquad (2)$$

$$X_j \in B(X_i,\varepsilon) \qquad (3)$$

A cluster is defined as a non-empty maximal subset of X in which all the pairs of points in the cluster are density-connected. In HDBSCAN, a modified distance metric is introduced [45], such that if $\kappa(X_i)$, also called *core-distance*, represents the distance of X_i to its k^{th} nearest neighbor, the *mutual reachability distance* between X_i and X_j is defined as:

$$d_{mreach}(X_i,X_j)=\begin{Bmatrix} max\{\kappa(X_i),\kappa(X_j),d(X_i,X_j)\} & X_i\neq X_j \\ 0 & X_i=X_j \end{Bmatrix} \qquad (4)$$

The *mutual reachability distance* places outliers farther away from clusters. The calculation process of mutual reachability distance is presented in **Fig. 1**. For a given distribution of points

in a 2D space (**Fig. 1(a)**), the mutual reachability distance of points X_1 and X_2 is equal to their actual distance (**Fig. 1(b)**). Therefore, the mutual reachability distance method does not change the distance between these two points that are inside a relatively dense area of the dataset. In contrast, the mutual reachability distance between points X_1 and X_3 is equal to the core-distance of point X_3 (**Fig. 1(c)**). Therefore, point X_3 that is an outlier is repelled more from the core region.

The hierarchical clustering of X is established by applying the standard Single Linkage Clustering algorithm [46] to the discrete metric space (X, d_{mreach}). Interestingly, the same results are obtained with the hierarchical clustering algorithm for a ε value and the modified DBSCAN for ε and k values [37]. To detect regions with the greatest density within a point cloud, a hierarchical cluster tree based on the variation in density can be used, where the local density at each point is determined by estimating the core-distance value associated with each point, **Eq. 5**.

$$\lambda = \frac{1}{\varepsilon} \qquad (5)$$

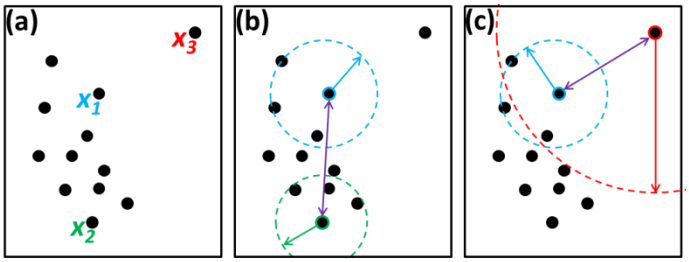


Figure 1*. (a) A 2D distribution of points. Mutual reachability distance for (b) points X_1 and X_2 and (c) points X_1 and X_3. k is set to 4. The core distance for points X_1 and X_2 are depicted by blue and green arrows in (b), respectively. Also, the actual distance between these two points is shown by a double-sided purple arrow. Since the actual distance between these two points is larger than their core-distance values, the mutual reachability distance is equivalent to the real distance between X_1 and X_2.*

The hierarchy of the cluster tree can be simplified by recursively merging some of the clusters. The condensation of the cluster tree is done by considering the minimum acceptable cluster size, m, and only accepting the pruning (i.e. splitting) of a cluster that would not persist against the increment of λ value, into at least two children with sizes larger than m. Based on this approach, the stability, σ, of each cluster is defined by summing the range of λ values for every point of a cluster following **Eq. 6** [37],

$$\sigma(C_i) = \sum_{X_j \in C_i} (\lambda_{\max,C_i}(\mathrm{X}_j) - \lambda_{\min,C_i}(\mathrm{X}_j)) \qquad (6)$$

where $\lambda_{max,Ci}(X_j)$ and $\lambda_{min,Ci}(X_j)$ represent the bounds of λ values over which point X_j is a member of cluster C_i. Large and smaller λ values would lead to pruning or merging and therefore affiliation to a different cluster. Stability is an important quantity that defines real clusters from spurious clusters.

To achieve an optimal clustering attribution among all possible clustering outcomes, the overall persistence score among all selected clusters must be maximized while considering the constraint that there is no overlap between clusters [37]. For this purpose, clusters that maximize the total persistency according to **Eq. 7** are selected,

$$\sum_{i \in I} \sigma(\mathrm{C}_i) \qquad (7)$$

where *I* is a subset of (1, 2, …, *n*) and *n* is the total number of possible clusters. **Equation 7** is subjected to the constraint presented in **Eq. 8** for all *i*, *j*∈*I* with *i*≠*j*.

$$C_i \cap C_j = \varnothing \qquad (8)$$

- **Density-Based Clustering (DeBaCl)**

We selected the DeBaCl algorithm because of its well-interpretable dendrogram plots. The DeBaCl algorithm was developed following the seminal work of Chaudhuri et al. [47, 48] and their definition of Robust Single Linkage. Note that the symbols/terminologies used in this **section 2.2** were chosen in a way to be consistent with those used in the literature describing the DeBaCl package [35]. Therefore, the symbols and terminology used in this section should not be confused with those discussed in the previous section 2.1 describing the HDBSCAN analysis.

In the DeBaCl method, it is first assumed that the acquired data (Xn={x1,…,xn} in R^d) represents an unknown probability density function, *f*. It is possible to establish a hierarchical cluster structure in which a cluster is defined as a connected subset of an *f*-level set (or λ-upper level set of *f* where λ≥0) [34], **Eq. 9**.

$$L_\lambda(f) = \{x \in R^d : f(x) \geq \lambda\} \qquad (9)$$

Since the probability density function is unknown, it is estimated by intersecting the level sets of *f* with the dataset points. The estimation of the probability density function $\hat{f}(x)$ from the dataset is done according to **Eq. 10**,

$$\hat{f}(x_j) = \frac{k}{n.v_d.\mathrm{r}_k^d(x_j)} \qquad (10)$$

where v_d and $r_k(x_j)$ represent the volume of the Euclidean unit ball in R^d and the Euclidean distance from point x_j to its *k*'th closest neighbor, respectively.

The significant advantage is the construction of a level set tree defined as the set of all connected components of $L_\lambda(f)$ for any arbitrary value of λ. Since each pair of clusters can be either the subset of another cluster or be completely independent of the other clusters, these clusters can be depicted by a tree (dendrogram). A pruning point in a dendrogram represents the density level where more than one mode of the probability distribution function (i.e., new clusters) emerges. Vertical lines in a dendrogram that do not prune represent high-density clusters and are called leaves of the level set tree. The dendrogram plots provided by DeBaCl make the interpretation of the cluster analysis results straightforward and an example for a 2D dataset is provided in **Fig. 2**. The dendrogram branches are sorted from right to left by increasing the fraction of the points in each cluster (i.e. empirical mass). Each branch width and the surrounding white space are proportional to the associated cluster empirical mass. The color of each branch is matched with the color of the associated high-density cluster. As shown in **Fig. 2(c)**, by

changing the density value (i.e., λ), the *f*-level sets nest and a cluster tree forms. Each leaf of this tree represents a cluster that exists over a specific range of λ value. Since the density levels depend on the height of the probability density estimate $\hat{f}$, the interpretation of the level set trees based on λ is difficult (e.g., **Fig. 2(c)**). For instance, depending upon the distribution of points in a dataset, λ=1 can be either a high or low density threshold. To prevent this problem, level set trees are indexed according to the probability content associated with the upper level sets following **Eq. 11** [49],

$$\lambda_\alpha = \sup\left\{\lambda : \int_{x \in L_\lambda(f)} f(x)dx \geq \alpha\right\} \qquad (11)$$

where α is a value between 0 and 1, **Fig. 2(d)**. Notably, while some of the nodes are either stretched or compressed, the topology of the dendrogram does not change by expressing the height of the tree using α instead of λ. The interpretation of dendrogram with α is easier because larger α values indicate more condensed and well-disconnected clusters and this parameter is less sensitive to small fluctuations of density estimates. Since α varies between 0 and 1, it is possible to compare level set trees derived from different probability distribution functions. Also, illustrating clusters with large probability content and low density is easiest done using α than λ values. However, the height of a branch associated with a cluster using the α index does not reflect the size of the cluster because the height of a branch depends on both its empirical mass and the empirical mass of the other coexisting branches. To fix this drawback, a κ index was introduced by Kent et al. [35], **Fig. 2(e)**, where κ is the branch-mass. The topology of the dendrogram changes by switching from λ or α to κ because the heights of the leaves in a dendrogram plotted with the κ index are proportional to their empirical mass while the heights of leaves in a dendrogram drawn using α or λ indices depend on the calculated density values.

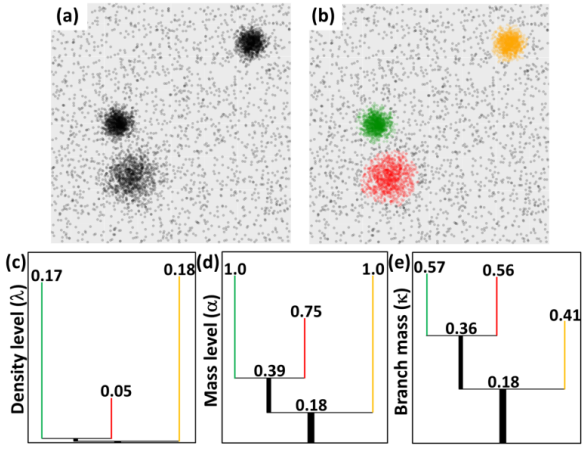


Figure 2. (a) A synthetic 2D dataset. (b) DeBaCl-detected clusters for the associated dataset. For a better visualization of the dense area, a transparency value of 0.125 is applied to each point. (c) Density, (d) mass, and (e) branch-mass dendrogram plots.

- **k-nearest neighbors (k-NN)**

The *k-nearest neighbors* machine learning algorithm is typically used for regression and classification purposes [50]. In k-NN classification, an unclassified object is assigned to the class with which the object has the most common among its k-nearest neighbors. In

the case of a cluster analysis, a point that is not considered as a member of any cluster may be assigned to one of the detected clusters. Initially, a user-defined k value is selected. Subsequently, the first k-nearest neighbor points that are members of clusters are found. The mentioned point is assigned to the cluster that has the greatest number of members neighboring this point.

- **Implementation of the algorithms in an integrated cluster analysis approach**

We combined the three above algorithms to create a workflow enabling robust and reliable density-based cluster identification, outlier detection, and visualization of detected clusters within APT data. The flowchart of the method is presented in **Fig. 3** with inputs and outputs explained in **Table 1** and **Table 2**, respectively. All the computer programs were developed using custom programs written in MATLAB® and Python 2.7. The entire analyses were done on a 4 GB memory MacBook Air machine equipped with 1.8 GHz Intel Corei7 processor. All the Python packages and their versions used for the current study are mentioned in the manual provided for this code. This manual is available on GitHub [38].

HDBSCAN is used first: it needs two mandatory parameters and can divide the dataset into smaller volumes containing clusters that can be more readily handled by DeBaCl. The most important input parameter of HDBSCAN is a rough estimate of the minimum cluster size (MinClusterSizeHDBSCAN) expressed in number of core atoms. For the analysis of APT datasets, it is advised to choose a value of the MinClusterSizeHDBSCAN parameter that is slightly smaller than the expected minimum cluster size because some of the atoms located at the border of each cluster may not be considered as a member of the detected cluster due to the relatively low probability of belonging to that cluster. The next parameter required by HDBSCAN is MinSamplesHDBSCAN. This parameter quantifies how conservative the cluster analysis is. The larger the value of this parameter, the more points are considered as noise and clusters are restricted to progressively denser and denser areas. Alternatively, reducing this parameter leads to more clusters by pruning. For APT data, it is advised to choose relatively large values of this parameter even if the resulting detected clusters are too large because DeBaCl will separate them in the following step. The two parameters, MinClusterSizeHDBSCAN and MinSamplesHDBSCAN, are the only two parameters required to perform the HDBSCAN analysis. Note that in principle, the HDBSCAN package requires only one parameter (i.e., MinClusterSizeHDBSCAN) as the second parameter is set to a default value. However, based on the authors' experience, adjusting the MinSamplesHDBSCAN parameter to a rough value can improve the outcome of the APT cluster analysis.

To make the analysis process more robust for APT datasets, four additional optional parameters are introduced within HDBSCAN. These parameters improve the reliability of identifying real clusters and identifying atoms pertaining to these clusters. The first parameter is the persistency threshold (hdbscanPersistencyThreshold). Only detected clusters that have persistency values (**Eq. 6**) larger than the threshold value are accepted. An appropriate value for this parameter can be selected from a cumulative plot of the number of clusters with respect to the persistency threshold. See **section 3.2** for further discussion on value selection for this parameter. The second parameter (hdbscanProbabilityThreshold) is the probability threshold above which an atom is considered a member of a given cluster. A probability value of being a member of a cluster is assigned to each atom by normalizing the values calculated according to **Eq. 5** within the range of zero and one. This parameter is used to increase the fidelity of the cluster analysis. The next parameter (hdbscanAnalysisTwoTimes) is a Boolean value for

repeating the HDBSCAN analysis. For a large number of microstructures, this parameter is not required, however, for some challenging APT datasets (e.g., datasets in which clusters with considerably different sizes exist), it may be useful to repeat the HDBSCAN analysis. Notably, before conducting the second HDBSCAN analysis, atoms that are considered to be a member of a real cluster are removed from the studied dataset. Therefore, the probability distribution function of the dataset will change. Since the second iteration is for the detection of clusters that were not selected in the first round, the persistency threshold value should be increased compared to its initial value. Here we define as the hdbscanPersistencyThreshold parameter multiplied by the PrefactForPersistency parameter (with values typically between 1 and 5).

The DeBaCl analysis requires two mandatory parameters. The first parameter, kDeBaCl, is used to estimate the probability distribution function according to **Eq. 10**, and is used to set the number of neighbors associated with each atom. The second parameter, gammaDeBaCl, represents the minimum acceptable leaf size below which leaves are recursively merged. The value chosen for this parameter must be either slightly smaller (e.g., 80-90%) or equal to the value chosen for MinClusterSizeHDBSCAN. Consequently, the value of gammaDeBaCl is predetermined by that of MinClusterSizeHDBSCAN.

Again, to improve the outcome of the analysis, two additional parameters are introduced within the DeBaCl analysis. The first one (i.e., IgnorePersisInDeBaClanalysis) is a logical parameter. If this parameter is set to true, initially DeBaCl only analyzes (and may prune) the HDBSCAN-detected clusters that have persistency values greater than hdbscanPersistencyThreshold. Subsequently, the DeBaCl analysis is repeated for all the detected clusters regardless of their persistency values. The second parameter is MassLengthDeBaClThreshold. This parameter is used to merge leaves when the length of each leaf in **Fig. 2(d)** is less than MassLengthDeBaClThreshold value. It is also used to merge recursively leaves when one parent (that can be a part of a cluster) is pruned to more than two children (leaves) with one of the leaves not satisfying the minimum cluster size chosen for DeBaCl analysis (i.e., gammaDeBaCl).

The final step of the proposed cluster analysis framework is conducting k-NN analysis. The purpose of this step is assigning atoms considered as noise by DeBaCl analysis to one of the detected clusters by DeBaCl. The number of closest neighbors considered for k-NN analysis is set by NNNforKNN parameter. Notably, the value of this parameter must be an odd number. Also, to avoid assigning all the atoms detected by HDBSCAN to a cluster by k-NN, atoms which are considered as noise by DeBaCl and have DeBaCl-calculated density values less than a fraction of the minimum DeBaCl density of atoms inside a cluster are removed. This fraction is presented by a parameter called PrefactorDeBaClDensityThreshold.

Table 1. List of input parameters (mandatory inputs for each step are in italic)

Analysis	Parameter	Comment
HDBSCAN	*MinClusterSizeHDBSCAN*	Minimum number of atoms in a cluster
	MinSamplesHDBSCAN	Measure of how conservative the HDBSCAN cluster detection is
	hdbscanPersistencyThreshold	Identified clusters are accepted if they have persistency values larger than this threshold.
	hdbscanProbabilityThreshold	Atoms are considered as members of a cluster if their probability values are greater than this threshold.
	hdbscanAnalysisTwoTimes	A logical parameter

	PrefactForPersistency	Safety factor for the HDBSCAN persistency threshold when the HDSCAN analysis is conducted for a second time.
DeBaCl	*kDeBaCl*	Number of observations to consider as neighbors to a given point
	gammaDeBaCl	Leaf sizes smaller than this value are recursively merged
	IgnorePersisInDeBaClanalysis	A logical parameter
	MassLengthDeBaClThreshold	Survival interval for mass changes of a cluster leaf
k-NN	*NNNforKNN*	Number of nearest neighbors for the k-nearest neighbor analysis. It must be an odd number.
	PrefactorDeBaClDensityThreshold	A correction factor to reduce the minimum density (similar to probability in HDBSCAN) limit required for atoms to be considered inside a cluster

Table 2. List of outputs for each step of the analysis

Outputs_1A & Outputs_1B	Cluster ID for each detected cluster
	Cluster persistency value for each cluster
	Atom positions for each cluster
	Probability for each atom to be considered as a member of a cluster
Outputs_2A & Outputs_2B	New cluster ID for each atom if the HDBSCAN-detected cluster is pruned
	Density (analogous to probability) for each atom
	Dendrogram of mass level
Outputs_3	Updated version of Outputs_2
Outputs_4	Updated version of Outputs_3

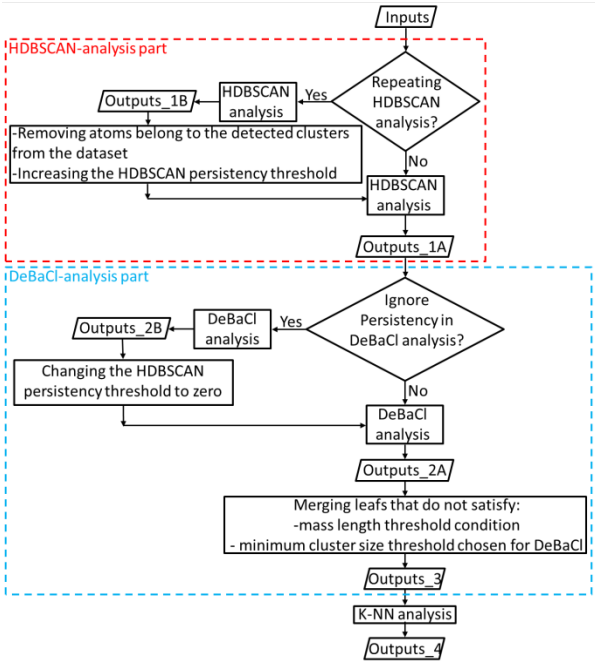


Figure 3. *Flowchart of the proposed framework.*

- **Synthetic APT datasets**

To illustrate the performance of the developed cluster analysis method, four synthetic APT datasets with known ground truth were generated and are shown in **Fig. 4**. The first dataset is used to introduce different aspects of the code through a straightforward analysis. In dataset 2, the detector efficiency, and therefore the overall solute density as well as the concentration of matrix and clusters, changed along the x-axis of the dataset. In dataset 3, some of the clusters were positioned so their boundaries overlap. The interfaces of the clusters in the dataset 4 were set to be diffuse.

Each generated dataset contained a spatial distribution of a given atom type. A fixed number of clusters was randomly dispersed in the volume with the constraint that the cluster positions satisfy a minimum inter-cluster separation distance condition. The radii of the clusters were set by a Gaussian size distribution. An appropriate number of atoms were assigned to each cluster based on a set cluster composition. To make the synthetic datasets more representative of experimental APT data, atoms were moved from their atomic lattice sites by a random Gaussian translation. For the atoms in the matrix, the translation distance was determined by a normal distribution with a variance equal to the lattice parameter in all three directions. To account for the possible trajectory aberrations due to different field evaporation behavior, the variance was two times the lattice parameter for the x and y components of each solute atom position in the clusters. In addition, for dataset 4, the variance of the delocalization was 1.5 x lattice parameter, thereby creating clusters with more diffuse interfaces. To take into account the detection efficiency of APT instruments, 63% of the atoms were randomly removed from the generated datasets. In the case of dataset 2, the detection efficiency was gradually increased from 37% to 55% along the x-axis in 200 steps. Local variations in the atomic densities within the reconstructed APT volumes are commonly observed at crystallographic poles and zone lines, or due to uneven laser heating for example. These density variations can be problematic in the application of the MSM [12]. **Table 3** summarizes the dataset characteristics. The code used to generate the dataset is provided on Github [38]. The proposed hierarchical density-based cluster analysis is in principle independent of the precipitate morphology, unlike the Gaussian mixture model [14]. Therefore, we are only testing near spherical clusters for simplicity. Further tests might explore more complex topologies, in particular the ability to detect dislocation lines decorated by solute atoms.

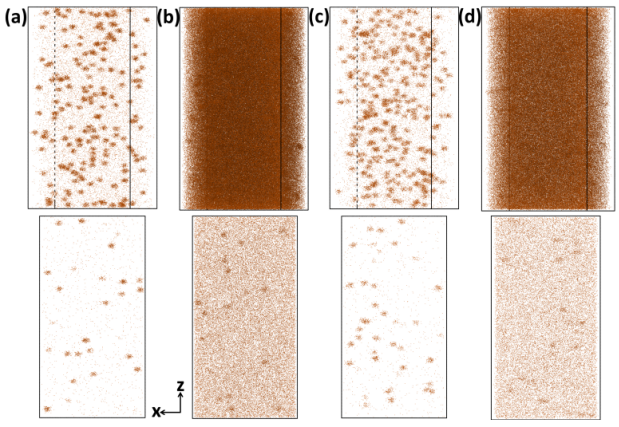


Figure 4. The generated volumes and a 5 nm-thick slice from each box for (a) Dataset 1 with high contrast, uniform density, and spatially distinct clusters, (b) Dataset 2 with low contrast and a

*gradient of detector efficiency along x-axis, (c) Dataset 3 with some clusters in close proximity, and (d) Dataset 4 with low contrast and clusters with diffuse interfaces. The size of each simulated box is 50 nm * 50 nm * 100 nm.*

Table 3. Cluster characteristics for the four synthetic datasets

Dataset	Number of clusters	Cluster radius range before delocalizatio n (nm)	solute concentration in matrix (clusters) (at.%)	Minimum inter-cluster spacing (nm)	Delocalization distance (nm)	Detector efficiency (%)	cluster size range (number of atoms)
1	176	1.1±0.1	0.42 (75)	5.04	0.3515	37	99-212
2	90	1.1±0.1	5 (40)	5.89	0.3515	37-55	58-229
	90	1.3±0.1					
3	275	1.2±0.1	0.41 (40)	2.24	0.3515	37	62-152
4	180	1.1±0.1	4 (40)	5.04	0.5272	37	47-106

- **Assessment of the capability of MSM to identify clusters in the generated datasets**

To compare the capability of the developed approach with a cluster method commonly used within the APT community, namely the maximum separation distance, the same datasets were analyzed by MSM as implemented in the CAMECA® IVAS™ 3.8.0 software [18]. The three important parameters of the MSM, order (O), maximum separation distance (D_{MAX}), and the minimum cluster size (N_{MIN}), for each dataset were determined using nearest neighbor distribution and cluster size distribution functions, as provided by IVAS™ software. The order parameter was adjusted to 10 atoms for all the analyses to increase the algorithm's ability to detect clusters.

In the case of dataset 1, the D_{MAX} and N_{MIN} parameters were set to 0.9 nm and 83 atoms, respectively and the maximum separation method could detect the 176 clusters properly with the cluster size range between 95 and 212 atoms, which is in close agreement with the ground truth. For dataset 2, the D_{MAX} and N_{MIN} parameters were set to 0.7 nm and 40 atoms, respectively. The maximum separation method could identify the location of 165 generated clusters properly, **Fig. 5(a)**. The size of the detected clusters varied between 42 and 351 atoms, which is a wider range than the ground truth (i.e., 58-229 atoms). The MSM analysis for this dataset was sensitive to the parameter selections. For dataset 3, 263 clusters were identified by adjusting D_{MAX} to 0.76 nm and N_{MIN} to 15 atoms, **Fig. 5(b)**. The discrepancy with the actual number of clusters (i.e., 275) is attributed to the inability of this method to prune merged clusters. The cluster size range varied between 15 and 247 atoms, which is far larger than the ground truth (i.e., 62-152 atoms). Finally, for dataset 4, the D_{MAX} and N_{MIN} parameters were set to 0.87 nm and 35 atoms, respectively and 168 clusters were detected correctly, **Fig. 5(c)**. The size of the detected clusters was between 38 and 222 atoms, which is again a larger range in comparison to the ground truth (i.e., 47-106 atoms). The cluster analysis results for this dataset were sensitive to the parameter selection.

Generally, the simplicity of the MSM method is attractive. However, as illustrated with the examples above, it is sensitive to the noise level and proximity of clusters leading to

inaccurate cluster size values and potentially the estimates of cluster density and locations. However, the main drawback is its lack of statistical estimators

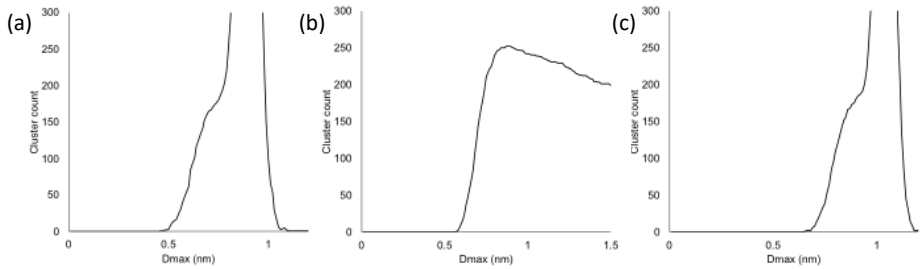


Figure 5. Cluster count distribution determined by maximum separation method for (a) dataset 2, (b) dataset 3 and (c) dataset 4. For the better visualization, cluster count values larger than 300 are not shown.

Results and Discussion

The parameters used to analyze each dataset are summarized in **Table 4**. The output of the analyses in terms of the number of detected clusters and their size are summarized in **Table 5** and **Table 6**, respectively. HDBSCAN_All represents all detected clusters regardless of their persistency value (**Eq. 6**), while HDBSCAN_Selected indicates only those that have persistency values larger than the persistency threshold value.

Table 4. Selected values for the input parameters used in each dataset (mandatory inputs for each step are in italic)

Analysis	Parameter	Dataset 1	Dataset 2	Dataset 3	Dataset 4
HDBSCAN	*MinClusterSizeHDBSCAN*	50	52	35	36
	MinSamplesHDBSCAN	20	12	15	16
	hdbscanPersistencyThreshold	0.04	0.03	0.04	0.03
	hdbscanProbabilityThreshold	0.5	0.85	0.5	0.95
	hdbscanAnalysisTwoTimes	False	False	True	False
	PrefactForPersistency	2	2	2	2
DeBaCl	*kDeBaCl*	7	9	7	9
	gammaDeBaCl	40	45	25	28
	IgnorePersisInDeBaClanalysis	False	False	False	False
	MassLengthDeBaClThreshold	0.35	0.25	0.25	0.3
k-NN	*NNNforKNN*	9	9	9	9
	PrefactorDeBaClDensityThreshold	0.5	0.5	0.5	0.5

Table 5. Number of detected clusters for each analysis steps and MSM

Dataset	HDBSCAN_All	HDBSCAN_selected	DeBaCl	MSM
1	176	176	176 (176)*	176
2	672	180	181 (179)	165
3	218	218	268 (275)	263
4	505	169	185 (178)	168

*The correctly-detected generated clusters by DeBaCl are mentioned in parenthesis.

Table 6. Detected cluster size range in terms of number of atoms

Dataset	HDBSCAN_All	HDBSCAN_Selected	DeBaCl	k-NN	MSM

1	97-233	97-233	97-201	97-233	95-212
2	52-747	108-534	83-272	83-272	42-351
3	58-569	58-569	38-208	43-226	15-247
4	36-394	59-394	28-127	32-137	38-222

- **Dataset 1**

A large minimum inter-cluster spacing as well as a high contrast of the composition between matrix and clusters make the cluster analysis of this dataset relatively straightforward. HDBSCAN detected all the 176 generated clusters that all had persistency values larger than the threshold value. The minimum persistency value of the detected clusters was 0.17, i.e. above the minimum threshold set for the analysis and DeBaCl did not prune any of these clusters. The spatial positions of the detected clusters at each step of the analysis can be visually compared to those of the generated clusters (**Fig. 6)**. However, rather than a qualitative 3D plot, we use a quantitative comparison of the cluster centers, by computing the nearest neighbor distances between clusters for a given cluster distribution or by computing the distances between the detected clusters and their closest generated cluster. Small values of the latter suggest good positioning accuracy, as illustrated in **Fig. 7-1**. However we note that this metric does not account for missed or additional clusters.

The cluster size range of the HDBSCAN-detected clusters is slightly larger than the generated cluster size range. Subsequently, DeBaCl reduced the cluster size range and k-NN relabeled all the atoms considered as noise by DeBaCl (**Table 6)**. Depending upon the parameter values selected for k-NN and the hdbscanProbabilityThreshold parameter, one can relabel all the atoms considered as noise by DeBaCl or only those that have a higher probability to be considered as a member of a DeBaCl-detected cluster.

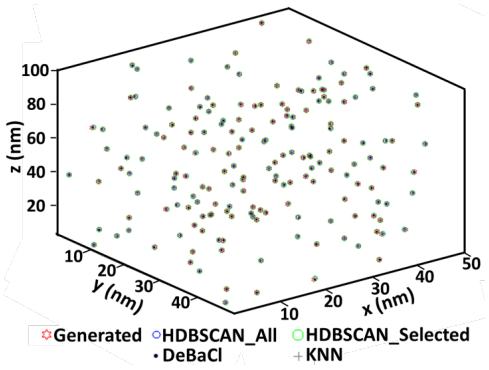


Figure 6*. Spatial distribution of the generated and detected cluster centers for dataset 1 (no challenge dataset).*

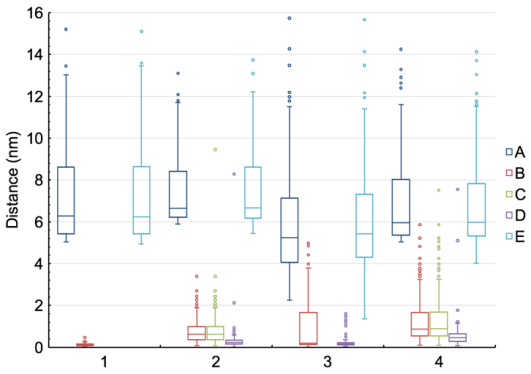


***Figure 7**. Distributions of the inter-cluster distances for each dataset. (A) Distances separating nearest neighbor clusters within the generated cluster distribution; (B) Distances between each cluster in the HDBSCAN_All distribution and the nearest cluster in the generated cluster distributions; (C) Distances between each cluster in the HDBSCAN_Selected distribution and the nearest cluster in the generated cluster distributions; (D) Distances between each cluster in the DebaCl distribution and the nearest cluster in the generated cluster distributions; (E) Distances separating nearest neighbor clusters within the DeBaCl output. The box represent the first quartile, median, and third quartile. The solid line extends from the minimum value to either the maximum value or 1.5 x the inter-quartile range (IQR) of the distributions. Outliers defined as values beyond the solid line are shown as dots. For the differences between distributions, the ground truth cluster positions were used as reference.*

- **Dataset 2**

The fact that 672 clusters were identified by HDBSCAN (**Table 5** and **Fig. 7-2**) casts doubt on whether all the detected clusters are original clusters. The persistency threshold parameter is used to address this question. The cumulative plot of the number of clusters with respect to the HDBSCAN persistency threshold provides a relatively straightforward way to determine an appropriate value for the persistency threshold by identifying a major gap or a sharp change in the slope of the curve (**Fig. 8)**. A persistency threshold of 0.03 resulted in only 180 clusters identified as real clusters (**Table 5**). The DeBaCl analysis enhances the accuracy of finding the location of the generated cluster centers (**Fig. 7-2**), as illustrated in **Fig. 9(a)**. This improvement is a result of trimming the HDBSCAN-detected clusters by DeBaCl through removing points which have the least probability to be considered as a member of the detected cluster. As presented in **Table 6**, the cluster size distribution for the DeBaCl-detected clusters is in good agreement with the generated cluster size range. In general, we expect that for APT dataset with relatively low density (or concentration) contrast between matrix and clusters, HDBSCAN will identify an unusually high number of clusters that can be trimmed using a persistency threshold.

The low concentration contrast between matrix and clusters as well as the random delocalization of atoms can lead to the formation of small, dense volumes. For the most part, these patterns are identified as noise by HDBSCAN since the minimum cluster size parameter that was set relatively high, here at 52 (**Table 4**). However, when these patterns form very close to the original clusters, they may be identified as a part of an actual cluster, leading to inaccuracy in finding the center location for some clusters, **Fig. 9(b)**. In general, for a low contrast dataset, it seems appropriate to increase the probability threshold for considering atoms to be a member of a cluster. It is also wise not to relabel most of the atoms considered as noise by DeBaCl (in the process of trimming HDBSCAN-detected clusters) during k-NN analysis.

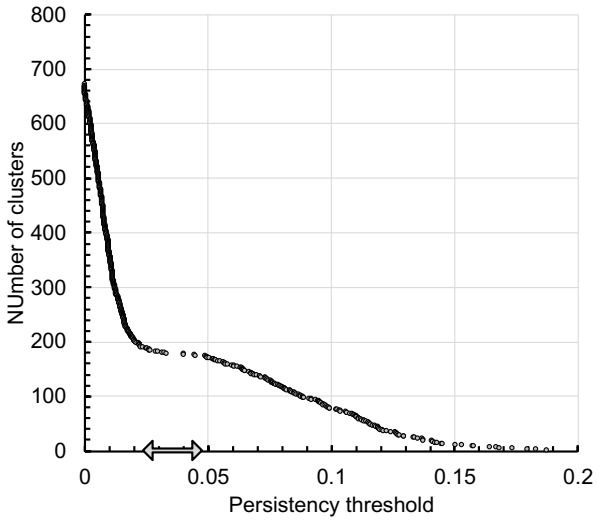


***Figure 8**. Cumulative number of clusters with respect to HDBSCAN persistency threshold for dataset 2. The double-sided arrow represents the interval for an appropriate value of the persistency threshold parameter.*

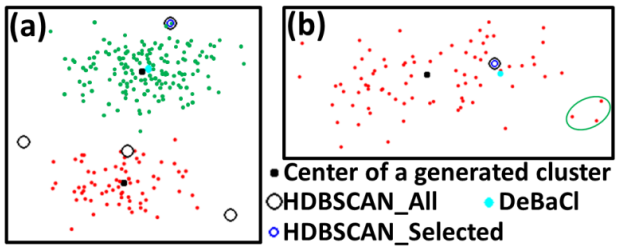


***Figure 9**. (a) A generated cluster considered as noise is presented by red color atoms. DeBaCl could improve the accuracy of finding the center location of the green cluster. (b) DeBaCl could not improve the accuracy of finding the center location of the cluster because a tiny, dense volume (surrounded by the green oval) formed randomly close to this generated cluster. Notably, atoms considered as noise by the developed cluster analysis method are not presented in these plots.*

- **Dataset 3**

HDBSCAN detected 218 very stable clusters with a minimum persistency of 0.11. Notably, 217 of these clusters were initially detected. After removing atoms belonging to these 217 detected clusters and re-doing the analysis, HDBSCAN found one additional cluster. As shown in **Fig. 7-3(b)**, some clusters were not detected accurately, i.e. their minimum distance is further than 2 nm from the center of the generated clusters. Analyzing the HDBSCAN-detected clusters by DeBaCl resulted in the detection of 268 clusters and a significant accuracy improvement in locating the cluster centers, **Fig. 7-3**. Visual inspection shows that HDBSCAN considered some neighboring clusters as one cluster. One of the HDBSCAN-detected cluster centers is presented in **Fig. 10(a)**. This single HDBSCAN cluster contains four generated cluster centers. All the four generated clusters were detected accurately by DeBaCl during the pruning process of the mentioned HDBSCAN-detected cluster, as shown in **Fig. 10(b)**. However a closer look at the clusters that have the largest distances to a generated cluster shows that some of them did not satisfy the two conditions required for pruning (see **Fig. 3**) (i.e., gammaDeBaCl larger than 25 and MassLengthDeBaClThreshold greater than 0.25) and therefore were not pruned,

in other word remained as the combination of two clusters. Two of these clusters are shown in **Fig. 10(c)** by blue and red color atoms.

The cluster size distributions presented in **Table 6** indicate a larger distribution range in comparison to the generated dataset, as expected from the lack of pruning of some of the clusters. Since the concentration contrast between clusters and matrix is high, a wide range of hdbscanProbabilityThreshold values (e.g., 0.5) can be set without a significant impact on the final cluster size ranges after DeBaCl and k-NN analyses.

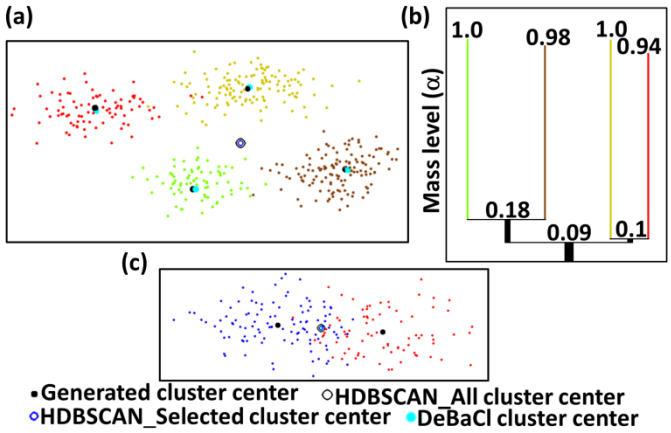


Figure 10. *(a) An HDBSCAN-detected cluster containing four generated clusters and (b) the corresponding mass plot. The colors of leaves are matched with the associated clusters in plot (a). (c) An HDBSCAN-detected cluster which is not pruned by DeBaCl. Noteworthy, noise atoms are not shown in plots (a) and (c).*

- **Dataset 4**

Because of the low density contrast, HDBSCAN detected a far greater number of clusters (i.e., 505) than the generated clusters (i.e., 180). The number of detected clusters reduced to 169 by applying the persistency threshold of 0.03 following the procedure mentioned in **section 3.2**. As presented in **Fig. 7-4(c)**, ~20% of the generated clusters were not detected properly or were considered as noise by HDBSCAN. DeBaCl could markedly improve the accuracy of the cluster detection by finding 178 generated clusters out of the 185 detected clusters in this dataset, **Table 5** and **Fig. 7-4(d)**. Two of the generated clusters were not detected due to their too low persistency values. These two missing clusters are pointed out by black arrows in **Fig. 11**. Notably, seven of the DeBaCl-identified clusters do not have any overlap with the generated clusters. One of these clusters is shown by the red arrow in **Fig. 11**. This cluster was detected initially by HDBSCAN and due to its high persistency value (i.e., 0.05), it was considered as a real cluster. This issue illustrates how the relatively large delocalization distance used for simulating positioning errors within the APT dataset and small cluster sizes create uncertainty in the detection of clusters. The lower bound of the cluster sizes detected by DeBaCl is significantly smaller than the lower bound of the cluster size distribution of the generated clusters. This difference can be assigned to the large probability threshold used in the HDBSCAN analysis (**Table 4**).

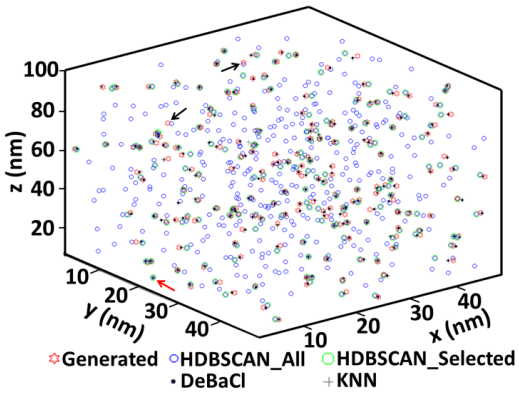


Figure 11. *The spatial distribution of the generated and detected cluster centers for dataset 4 (diffuse interfaces).*

- **Parameter Selection and Sensitivity analysis**

Compared to the maximum separation method, the present method is less sensitive to parameter selection. For cases in which the composition contrast of the matrix and clusters is high enough (e.g., dataset 1 and dataset 3), parameter selection is relatively straightforward. However, more caution is required for the cases with small composition contrasts between matrix and clusters (e.g., dataset 2 and dataset 4).

The outcome of a HDBSCAN analysis is not very sensitive to the minimum cluster size parameter as long as this parameter is chosen small enough to cover all the cluster sizes in the dataset and large enough to make a meaningful difference between clusters and noise. For instance, a value less than 10 to 15 may lead to considering some of the noise atoms as clusters (especially in low composition contrast APT datasets) due to the random formation of tiny, dense volumes by noise atoms.

The purpose of the HDBSCN step is to conservatively identify clustered regions. Indeed, when HDBSCAN merges close real clusters into one, DeBaCl analysis can prune the HDBSCAN-detected cluster properly (e.g., HDBSCAN_Selected and DeBaCl values for the merged clusters dataset in **Table 5** and **Table 6**). Therefore, choosing a relatively large value for the MinSamplesHDBSCAN parameter makes the cluster analysis results more reliable.

The selection of the persistency threshold parameter was already explained in **section 3.2**. As presented in **Table 4**, no major change was made in the persistency threshold value for analyzing different datasets. For dataset 2, changing the persistency threshold value between 0.03 and 0.05 does not significantly change the output of the HDBSCAN cluster analysis. Moreover, the persistency threshold parameter can be ignored in the large concentration contrast APT datasets. Notably, the option of not considering the persistency threshold for the cluster analysis is provided in the developed code for the current study. However, it is recommended not to ignore persistency values for the DeBaCl analysis if the composition contrast between matrix and clusters is low.

For low composition contrast APT datasets, it is wise to use larger values of the probability threshold for the HDBSCAN analysis. The major impact of this parameter is in determining the size of each detected cluster. As presented in **Table 4**, for low composition contrast datasets (i.e., dataset 2 and dataset 4), a larger value of the probability threshold was set in comparison to dataset 1 and dataset 3. Generally, in the case of high concentration

contrast datasets, atoms inside the actual clusters have a relatively large probability value and the value of hdbscanProbabilityThreshold parameter does not have a major impact on the size of the detected clusters.
Although it is mentioned that the kDeBaCl parameter is a mandatory parameter for the DeBaCl analysis, a fixed value (e.g., a value between 7 to 12) can be used for analyzing APT datasets. The values of this parameter were varied randomly for the four current APT datasets to show that it does not dramatically affect the performance of DeBaCl analyses.

The minimum cluster size (gammaDeBaCl) can be determined based on the value selected for the minimum cluster size in the HDBSCAN analysis step (see **section 2.4** and **Table 7**). This parameter in combination with MassLenghtDeBaClThreshold parameter determines the pruning condition for a cluster. As presented in **Table 4**, selecting a value between 0.2 and 0.4 should be meaningful for the MassLenghtDeBaClThreshold parameter.

Since k-NN analysis is an optional part of the developed cluster analysis code, it depends mainly on the user to either relabel a fraction of atoms considered as noise by DeBaCl or consider the cluster sizes determined by DeBaCl as the final cluster size values for the detected clusters. As shown in **Table 4**, no major change was applied to the k-NN parameters for all the four datasets. Since the probability values assigned to atoms inside the clusters is far larger than the noise atoms, the accuracy of k-NN analysis would be high even if we relabel all or most of the atoms considered as noise by DeBaCl for the large concentration contrast. However, in the case of low concentration difference between matrix and clusters, more caution is required. As stated previously, a large value should be chosen for the hdbscanProbabilityThreshold parameter if the concentration contrast is low.

To quantitatively investigate the sensitivity of the cluster analysis associated with the proposed method in the current study, dataset 2 was studied with different values of the four mandatory parameters required for running the code. All the other parameters were kept the same as those presented in **Table 4**. There was no major variation in the number of DeBaCl-detected clusters. The accuracy of all the conditions in finding the generated clusters, the cluster centers and the approximate size of the clusters was admirable (**Table 7)**. For instance, the maximum error in finding the number of real clusters among the 11 different input parameters was ~ 6%. Finally, we note that a cluster analysis is valid if its outcome is not sensitive to the variation of the four mandatory input parameters. Alternatively, if outputs are very sensitive to input parameter changes, we hypothesize that (1) the selected values for the parameters are very far from the appropriate values (e.g., choosing a value smaller than 5 for MinClusterSizeHDBSCAN) and/or (2) the APT dataset is not appropriate for density-based cluster analysis because there is no significant difference in the density of clusters and matrix.

For microstructures similar to the ones generated for the current study, possible values of hdbscanPersistencyThreshold, MassLengthDeBaClThreshold, and PrefactorDeBaClDensityThreshold may be found between 0.02 and 0.05, 0.2 and 0.4, and 0.5 to 0.75, respectively. These values may only be appropriate for the present data and further investigations are required to find the optimum values for other microstructures.

***Table 7**. Sensitivity analysis for four mandatory parameters of dataset 2 with 180 clusters and the cluster size range of 58-229 atoms*

	Parameter		Condition number										
			1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
HDBSCAN	*MinClusterSizeHDBSCAN*		52	45	40	35	30	40	40	40	40	35	36
	MinSamplesHDBSCAN		12	12	12	12	12	14	17	20	23	22	14
DeBaCl	*kDeBaCl*		9	10	8	9	10	8	9	10	8	9	10
	gammaDeBaCl		45	39	34	29	25	34	34	34	34	30	31
Outputs	Number of cluster	HDBSCAN_All	672	789	883	1004	1177	782	672	629	576	666	868
		HDBSCAN_ Selected	180	182	184	184	190	181	180	179	179	180	181
		DeBaCl	181	183	185	187	192	181	180	179	179	180	181
		Correct detection	179	180	180	180	180	180	180	179	179	180	180
	DeBaCl/k-NN cluster size range (number of atoms)		83-272	66-257	66-207	47-182	47-177	61-166	68-153	68-164	70-167	65-174	61-161

- **Limitations of the developed method**

In the case of low composition contrast between matrix and clusters, the performance of the developed hierarchical density-based clustering method is remarkably sensitive to the size of generated clusters. As an example, for a synthetic dataset in which the compositions of the matrix and clusters were 10% and 30%, respectively, almost all the generated clusters (with the cluster size of 1.4±0.1 nm (82-156 atoms)) were detected properly. However, by reducing the cluster size to 1.1±0.1 nm (34-89 atoms), the proposed method failed in identifying the generated clusters. This failure can be assigned to the formation of low persistency clusters due to the larger delocalization applied to atoms inside the clusters in comparison to the matrix (as stated in section 2.5). Notably, by applying the same delocalization to the matrix and cluster atoms, it was possible to detect the generated clusters in the mentioned synthetic dataset because the relative persistency of the actual clusters in comparison to the noise (i.e., matrix atoms) increased. However, reducing the cluster size to 0.9±0.1 nm (14-48 atoms) and applying the same delocalization to both matrix and cluster atoms led to the formation of low persistency clusters. While the present algorithm could initially detect all the generated clusters (HDBSCAN-analysis part presented in **Fig. 3**), 15%-25% of the detected clusters were rejected due to the low persistency of the random solute clusters. Further reduction of the cluster size in this low composition contrast dataset dramatically increased the possibility of interpreting noise (i.e., matrix atoms) as actual clusters. As a result, HDBSCAN detected far more clusters than the generated number of clusters in this type of datasets and a great number of these detected clusters were not rejected because of their relatively large persistency values. Therefore, the proposed cluster analysis method is not suitable for finding extremely small clusters (e.g., less than 0.8 nm) in low composition contrast APT datasets.

Conclusions

In this study, a hierarchical density-based cluster analysis for identifying clusters in APT datasets was introduced. The capabilities of this method and comparison with the maximum separation method were illustrated using four synthetic APT datasets. The maximum separation method partially find the true clusters in three of the datasets, while the proposed approach performed well in terms of number of detected clusters with respect to the generated clusters, the accuracy of finding the location of the generated cluster centers and the cluster sizes for these datasets. In particular, the new method

enables the analysis of microstructures with lower density contrast between matrix and clusters, neighboring clusters, or varying density variations, which are often observed in experimental datasets but remained problematic for the MSM method [12].

Four mandatory inputs are required for the proposed cluster analysis method: two parameters for the HDBSCAN analysis and two parameters for the DeBaCl analysis. Fortunately, not all four parameters have to be individually set and ranges of values are discussed. Specifically, the DeBaCl parameter (kDeBaCl) can be a constant value for all APT cluster analyses. The DeBaCl minimum number of cluster/leaf size can be set based on the value chosen for the HDBSCAN minimum cluster size. A rough estimate of the HDBSCAN minimum cluster size is sufficient as long as the selected value is small enough to cover all the cluster sizes and large enough not to consider random matrix fluctuations. The second mandatory HDBSCAN parameter (MinSamplesHDBSCAN) is also forgiving. A large value might lead to some HDBSCAN-detected clusters to contain more than one real cluster, which is then addressed by pruning during DeBaCl analysis. We also found that for the four synthetic datasets the values of this parameter changed within a relatively small range (i.e., between 12 and 20).

Further studies are required to evaluate the performance of the proposed method in identifying clusters associated with more complex topologies such as significantly large cluster size range and far smaller minimum inter-cluster spacing. The analysis code, test datasets, and code to generate datasets are available online and it can be readily optimized depending upon the requirements of the cluster analyses specifically for APT datasets.

References

[1] B. Gault, M.P. Moody, J.M. Cairney, S.P. Ringer, Atom probe microscopy, Springer Science & Business Media2012.
[2] E.A. Marquis, J.M. Hyde, Applications of atom-probe tomography to the characterisation of solute behaviours, Materials Science and Engineering: R: Reports 69(4) (2010) 37-62.
[3] T. Philippe, F. De Geuser, S. Duguay, W. Lefebvre, O. Cojocaru-Mirédin, G. Da Costa, D. Blavette, Clustering and nearest neighbour distances in atom-probe tomography, Ultramicroscopy 109(10) (2009) 1304-1309.
[4] P. Felfer, A.V. Ceguerra, S.P. Ringer, J.M. Cairney, Detecting and extracting clusters in atom probe data: A simple, automated method using Voronoi cells, Ultramicroscopy 150 (2015) 30-36.
[5] M.P. Moody, L.T. Stephenson, A.V. Ceguerra, S.P. Ringer, Quantitative binomial distribution analyses of nanoscale like-solute atom clustering and segregation in atom probe tomography data, Microscopy research and technique 71(7) (2008) 542-550.
[6] M.P. Moody, L.T. Stephenson, P.V. Liddicoat, S.P. Ringer, Contingency table techniques for three dimensional atom probe tomography, Microscopy research and technique 70(3) (2007) 258-268.
[7] F. De Geuser, W. Lefebvre, D. Blavette, 3D atom probe study of solute atoms clustering during natural ageing and pre-ageing of an Al-Mg-Si alloy, Philosophical Magazine Letters 86(04) (2006) 227-234.
[8] L. Couturier, F. De Geuser, A. Deschamps, Direct comparison of Fe-Cr unmixing characterization by atom probe tomography and small angle scattering, Materials Characterization 121 (2016) 61-67.
[9] W. Lefebvre, T. Philippe, F. Vurpillot, Application of Delaunay tessellation for the characterization of solute-rich clusters in atom probe tomography, Ultramicroscopy 111(3) (2011) 200-206.
[10] J.M. Hyde, C.A. English, Microstructual processes in irradiated materials, in: Lucas RGE, Snead L, Kirk MAJ, Elliman RG (Eds.) MRS 2000 Fall Meeting Symposium, Boston, MA, 2001, pp. 27-29.

[11] M.K. Miller, E.A. Kenik, Atom probe tomography: A technique for nanoscale characterization, Microsc. microanal. 10(3) (2004) 336-341.
[12] L.T. Stephenson, M.P. Moody, P.V. Liddicoat, S.P. Ringer, New techniques for the analysis of fine-scaled clustering phenomena within atom probe tomography (APT) data, Microsc. microanal. 13(6) (2007) 448-463.
[13] O.C. Hellman, J.A. Vandenbroucke, J. Rüsing, D. Isheim, D.N. Seidman, Analysis of three-dimensional atom-probe data by the proximity histogram, Microsc. microanal. 6(5) (2000) 437-444.
[14] J. Zelenty, A. Dahl, J. Hyde, G.D. Smith, M.P. Moody, Detecting Clusters in Atom Probe Data with Gaussian Mixture Models, Microsc. microanal. 23(2) (2017) 269-278.
[15] Y. Chen, P.H. Chou, E.A. Marquis, Quantitative atom probe tomography characterization of microstructures in a proton irradiated 304 stainless steel, Journal of Nuclear Materials 451(1–3) (2014) 130-136.
[16] W. Lefebvre, F. Vurpillot, X. Sauvage, Atom Probe Tomography, Put Theory into practice, Elsevier2016.
[17] J.M. Hyde, G. DaCosta, C. Hatzoglou, H. Weekes, B. Radiguet, P.D. Styman, F. Vurpillot, C. Pareige, A. Etienne, G. Bonny, N. Castin, L. Malerba, P. Pareige, Analysis of Radiation Damage in Light Water Reactors: Comparison of Cluster Analysis Methods for the Analysis of Atom Probe Data, Microscopy & Microanalysis 23(2) (2017) 366-375.
[18] D.J. Larson, T.J. Prosa, R.M. Ulfig, B.P. Geiser, T.F. Kelly, Local electrode atom probe tomography, New York, US: Springer Science (2013).
[19] E.A. Marquis, V. Araullo-Peters, Y. Dong, A. Etienne, S. Fedotova, K. Fujii, K. Fukuya, E. Kuleshova, A. Lopez, A. London, On the Use of Density-Based Algorithms for the Analysis of Solute Clustering in Atom Probe Tomography Data, Environmental Degradation of Materials in Nuclear Power Systems, Springer, 2017, pp. 881-897.
[20] J. Hyde, E. Marquis, K. Wilford, T. Williams, A sensitivity analysis of the maximum separation method for the characterisation of solute clusters, Ultramicroscopy 111(6) (2011) 440-447.
[21] P.D. Styman, J.M. Hyde, K. Wilford, G.D. Smith, Quantitative methods for the APT analysis of thermally aged RPV steels, Ultramicroscopy 132 (2013) 258-264.
[22] E.A. Jägle, P.-P. Choi, D. Raabe, The Maximum Separation Cluster Analysis Algorithm for Atom-Probe Tomography: Parameter Determination and Accuracy, Microsc. microanal. 20(6) (2014) 1662-1671.
[23] R.K.W. Marceau, L.T. Stephenson, C.R. Hutchinson, S.P. Ringer, Quantitative atom probe analysis of nanostructure containing clusters and precipitates with multiple length scales, Ultramicroscopy 111(6) (2011) 738-742.
[24] R.P. Kolli, D.N. Seidman, Comparison of Compositional and Morphological Atom-Probe Tomography Analyses for a Multicomponent Fe-Cu Steel, Microsc. microanal. 13(4) (2007) 272-284.
[25] A. Cerezo, L. Davin, Aspects of the observation of clusters in the 3-dimensional atom probe, Surface and Interface Analysis 39(2-3) (2007) 184-188.
[26] C.A. Williams, D. Haley, E.A. Marquis, G.D.W. Smith, M.P. Moody, Defining clusters in APT reconstructions of ODS steels, Ultramicroscopy 132 (2013) 271-278.
[27] Y. Dong, A. Etienne, A. Frolov, S. Fedotova, K. Fujii, K. Fukuya, C. Hatzoglou, E. Kuleshova, K. Lindgren, A. London, A. Lopez, ,, S. Lozano-Perez, Y. Miyahara, Y. Nagai, K. Nishida, B. Radiguet, D.K. K. Schreiber, N. Soneda, M. Thuvander, T. Toyama, J. Wang, F. Sefta, P. Chou, E.A. Marquis, Atom Probe Tomography Interlaboratory Study on Clustering analysis in experimental data using the maximum separation distance approach, Microscopy & microanalysis (in review) (2018).
[28] D. Hudson, G.D.W. Smith, B. Gault, Optimisation of mass ranging for atom probe microanalysis and application to the corrosion processes in Zr alloys, Ultramicroscopy 111(6) (2011) 480-486.
[29] B. Gault, M.P. Moody, F. De Geuser, A. La Fontaine, L.T. Stephenson, D. Haley, S.P. Ringer, Spatial resolution in atom probe tomography, Microsc. microanal. 16(1) (2010) 99-110.
[30] E.A. Marquis, F. Vurpillot, Chromatic aberrations in the field evaporation behavior of small precipitates, Microsc. microanal. 14(6) (2008) 561-570.
[31] F. Vurpillot, A. Bostel, D. Blavette, Trajectory overlaps and local magnification in three-dimensional atom probe, Applied Physics Letters 76(21) (2000) 3127-3129.

[32] J. Sander, X. Qin, Z. Lu, N. Niu, A. Kovarsky, Automatic extraction of clusters from hierarchical clustering representations, Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2003, pp. 75-87.
[33] P. Arabie, G. De Soete, Clustering and classification, World Scientific1996.
[34] J.A. Hartigan, Clustering algorithms (probability & mathematical statistics), John Wiley & Sons Inc New York, 1975.
[35] B.P. Kent, A. Rinaldo, T. Verstynen, DeBaCl: A Python package for interactive DEnsity-BAsed CLustering, arXiv preprint arXiv:1307.8136 (2013).
[36] J.A. Hartigan, Consistency of single linkage for high-density clusters, Journal of the American Statistical Association 76(374) (1981) 388-394.
[37] L. McInnes, J. Healy, S. Astels, hdbscan: Hierarchical density based clustering, The Journal of Open Source Software 2(11) (2017) 205.
[38] E.A. Marquis, APT cluster analysis - open tools. <https://github.com/emarq/ >, 2018.
[39] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, Kdd, 1996, pp. 226-231.
[40] R.J. Campello, D. Moulavi, J. Sander, Density-based clustering based on hierarchical density estimates, Pacific-Asia conference on knowledge discovery and data mining, Springer, 2013, pp. 160-172.
[41] R.J. Campello, D. Moulavi, A. Zimek, J. Sander, Hierarchical density estimates for data clustering, visualization, and outlier detection, ACM Transactions on Knowledge Discovery from Data (TKDD) 10(1) (2015) 5.
[42] R.R. Curtin, Faster dual-tree traversal for nearest neighbor search, International Conference on Similarity Search and Applications, Springer, 2015, pp. 77-89.
[43] R.R. Curtin, W.B. March, P. Ram, D.V. Anderson, A.G. Gray, C.L. Isbell Jr, Tree-independent dual-tree algorithms, arXiv preprint arXiv:1304.4327 (2013).
[44] W.B. March, P. Ram, A.G. Gray, Fast euclidean minimum spanning tree: algorithm, analysis, and applications, Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2010, pp. 603-612.
[45] J. Eldridge, M. Belkin, Y. Wang, Beyond hartigan consistency: Merge distortion metric for hierarchical clustering, Conference on Learning Theory, 2015, pp. 588-606.
[46] R. Sibson, SLINK: an optimally efficient algorithm for the single-link cluster method, The computer journal 16(1) (1973) 30-34.
[47] K. Chaudhuri, S. Dasgupta, Rates of convergence for the cluster tree, Advances in Neural Information Processing Systems, 2010, pp. 343-351.
[48] K. Chaudhuri, S. Dasgupta, S. Kpotufe, U.v. Luxburg, Consistent Procedures for Cluster Tree Estimation and Pruning, IEEE Transactions on Information Theory 60(12) (2014) 7900-7912.
[49] A. Rinaldo, A. Singh, R. Nugent, L. Wasserman, Stability of density-based clustering, Journal of Machine Learning Research 13(Apr) (2012) 905-948.
[50] M. Kuhn, K. Johnson, Applied predictive modeling, Springer2013.

Morphological classification of dense objects in APT data[9]

In addition to solute clustering, solute interactions with crystallographic defects are also of technological significance. Solute segregation to grain boundaries can affect grain boundary cohesion as well as grain growth, and quantification methods to assess grain boundary excess from APT data are already available. Solute segregation to dislocations can affect mechanical properties, such as hardening and strain aging. While APT is uniquely suited to quantify segregation to dislocations, quantitative methods remain severely limited. Hyde et al. adopted a manual approach to define the locations of

9 Published in Morphological classification of dense objects in atom probe tomography data, I Ghamarian, EA Marquis. Ultramicroscopy (2020) 215 112996 Codes:
DOI:10.5281/zenodo.3724970

dislocation lines.[10] Felfer et al. illustrated the use of density-based methods with Voronoi Tesselation to find and define isolated dislocation lines and analyze segregation for a given dislocation.[11] The automated identification of dislocation lines within 3D APT volumes, however, has not been described in the literature.

Our approach extends the CHD method described above. The CHD algorithm is first used to identify dense objects. A skeleton finder algorithm is then applied to the individual identified objects, and the length of the skeleton quantified (Figure 1).

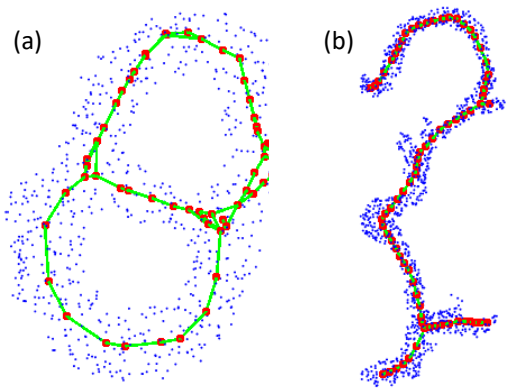
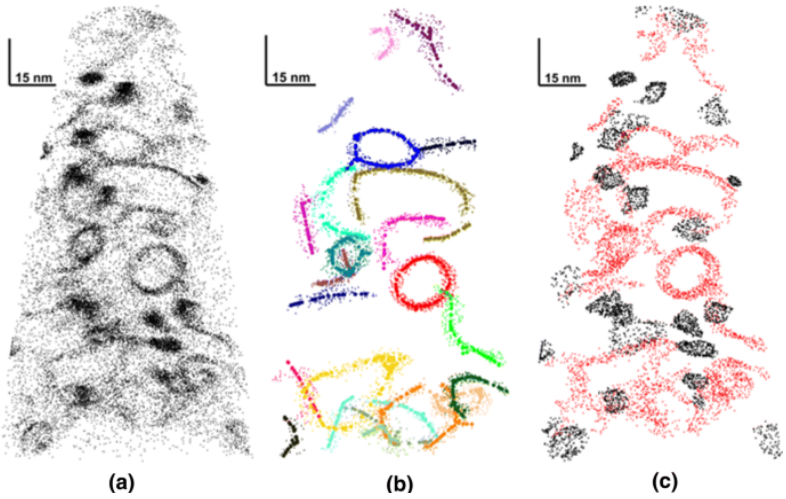


Figure 1: original dense point clouds and their identified skeletons.

After determining the length of the dense objects, a morphological classification is applied to separate the dense objects and assign them to different morphological features, i.e. clusters and dislocation lines or loops. The workflow was successfully demonstrated on the irradiated microstructure of a Ni-based alloy as shown in Figure 2. All codes are freely available online.



[10] C.A. Williams, J.M. Hyde, G.D.W. Smith, E.A. Marquis, Journal of Nuclear Materials, 2011, vol 412, pp.100-105
[11] P. Felfer, A. Ceguerra, S. Ringer, J. Cairney, Ultramicroscopy, 2013, vol. 132, pp. 100-106

Figure 2: Quantification of the microstructure of ion-irradiated Alloy 625 as determined by APT (a) reconstructed volume showing Si ions only, (b) dislocations lines detected using the workflow, and (c) dislocation loops (red) and Si clusters (black) illustrating the automated approach and ability to separate dense Si clusters by their morphology.