

ARTICLE OPEN



On strong-scaling and open-source tools for analyzing atom probe tomography data

Markus Kühbach¹✉, Priyanshu Bajaj^{1,2}, Huan Zhao¹, Murat H. Çelik³, Eric A. Jägle^{1,4} and Baptiste Gault^{1,5}

The development of strong-scaling computational tools for high-throughput methods with an open-source code and transparent metadata standards has successfully transformed many computational materials science communities. While such tools are mature already in the condensed-matter physics community, the situation is still very different for many experimentalists. Atom probe tomography (APT) is one example. This microscopy and microanalysis technique has matured into a versatile nano-analytical characterization tool with applications that range from materials science to geology and possibly beyond. Here, data science tools are required for extracting chemo-structural spatial correlations from the reconstructed point cloud. For APT and other high-end analysis techniques, post-processing is mostly executed with proprietary software tools, which are opaque in their execution and have often limited performance. Software development by members of the scientific community has improved the situation but compared to the sophistication in the field of computational materials science several gaps remain. This is particularly the case for open-source tools that support scientific computing hardware, tools which enable high-throughput workflows, and open well-documented metadata standards to align experimental research better with the fair data stewardship principles. To this end, we introduce *paraprobe*, an open-source tool for scientific computing and high-throughput studying of point cloud data, here exemplified with APT. We show how to quantify uncertainties while applying several computational geometry, spatial statistics, and clustering tasks for post-processing APT datasets as large as two billion ions. These tools work well in concert with Python and HDF5 to enable several orders of magnitude performance gain, automation, and reproducibility.

npj Computational Materials (2021)7:21 | <https://doi.org/10.1038/s41524-020-00486-1>

INTRODUCTION

Precise and accurate quantification of uncertainties and the significance of scientific results is essential for every study in computational and experimental materials science. However, such quantification is not only tedious but frequently also difficult in practice and theory. Software tools accomplish here a critical task: they encode theory and practice into methods, which ideally everybody can use for analyzing computer simulations and experiments. Data science tools, ideally automated, performant, and linked into high-throughput workflows, have successfully increased the complexity and type of materials science problems, which a specific community can study. Prominent examples are documented in the condensed-matter physics^{1–6}. In contrast, many software tools in experimental materials science are still at an earlier stage of development, which is in part due to the commercial interests of instrument vendors that lock data into proprietary formats and that use often opaque data-processing routines. In effect, this restricts the capabilities of such software tools for performing uncertainty quantification and reduces the effectiveness when exchanging results between computational and experimental materials scientists.

Atom probe tomography (APT) is one such example where open software tools and performant high-throughput analyses would be valuable. APT is a destructive microscopy and microanalysis technique, which allows the characterization of specific microstructural features with near-atomic resolution in three dimensions. Using either controlled laser or high-voltage pulses superimposed on a DC high-voltage, APT relies on the process of field evaporation to remove individual atoms from a

needle-shaped specimen in the form of ions. These are collected by a position-sensitive time-resolved detector system^{7–10}. The time-of-flight of each ion allows for elemental identification with isotopic resolution. The association of a range of mass-to-charge-state-ratio values to a single element is usually referred to as ranging^{11,12}. Following elemental identification, a combination of a reverse projection and a sequential depth-increment computation allows to reconstruct a point cloud; and thereby reveal the original atomic arrangement of the specimen¹³. The capability to resolve the atomic positions makes it possible to couple such experiments with computer simulations at the atomic scale¹⁴.

Improvements in instrumentation and experimental protocols in the past decade have made multi-million, as well as for some materials even billion, atom datasets accessible. Inspecting for instance the joint database for all APT microscopes of the MPIE yields a list of 743 datasets (measured between January, 2016 and February, 2020), which have all at least 100×10^6 ions collected. Combining APT with other microscopy techniques, in particular transmission electron microscopy¹⁵, results in a uniquely powerful tool for advanced materials characterization. The range of applications spans fields as diverse as physical metallurgy^{16,17}, geology and planetary chronology^{18–21}, solar energy harvesting²², biology^{23,24}, or semiconductors^{25–27}. Specimens in these fields range from single-crystalline, single-phase chemistry to complex multinary polycrystals with ten or more elements or amorphous phases^{28–31}. The range of materials amenable to APT analysis will keep expanding in the coming years with new cryo-preparation and transfer protocols being explored^{32–34}.

¹Max-Planck-Institut für Eisenforschung GmbH (MPIE), Düsseldorf, Germany. ²m4p material solutions GmbH, Magdeburg, Germany. ³Institute for Advanced Simulation (IAS), Jülich Supercomputing Centre (JSC), Jülich, Germany. ⁴Universität der Bundeswehr München, Neubiberg, Germany. ⁵Department of Materials, Imperial College London, Royal School of Mines, London, UK. ✉email: m.kuehbach@mpie.de

Post-processing of these APT data is a critical step in every study. Examples include tasks like reconstructing a dataset from the time-of-flight detector hit sequence⁷, characterizing spatial statistics³⁵, analyzing concentration fields and profiles³⁶, characterizing second-phase precipitates^{37–40}, or reconstructing microstructural features with methods from computational geometry^{41–43}.

A de facto near-monopolistic APT instrument landscape means that the Integrated Visualization and Analysis Software (IVAS), which is in the process of being replaced by APSuite, is an almost mandatory starting point for most practitioners^{44,45}. Benefits of such commercial software packages are their clear integration into the data-acquisition software of the instrument and the functionally-rich graphical user interface (GUI). It should be mentioned that many of the functions behind this GUI were implemented based on methods which were developed by members of the APT community. As a prototypic example of the challenges, which experimentalists face, though, IVAS has two key limitations: first, the raw data (of an APT measurement) are stored in proprietary container files. Second, the source code is closed. Therefore, quantifying eventual methodological uncertainties or interfacing IVAS with other community tools remains tricky; although the situation recently improved via scripting options within IVAS⁴⁶.

The situation motivated efforts by the APT community to develop complementary scripts and software tools^{38,41–43,47–56}. That these tools are open is an advantage because it enables peer-reviewing and continuous development by the scientific community. Most of these community tools constitute proof-of-concept implementations of algorithms or ad hoc developed patches of functionalities that are missing in commercial software. With a prime focus on serving as supplementary tools to support particular research efforts, though, these tools provide usually for sequential execution only.

In effect, practitioners are typically reluctant in setting up high-throughput post-processing workflows. Consequently, gaps remain on what constitutes metadata for APT, how to define these through a community-driven process, and how to exchange these and the data between different tools. This status quo teaches us that additional challenges exist when analyzing experiments in general and those for APT in particular. Therefore, taking strategic action is necessary and can be rewarding when using modern data science methods:

- Current experiments lead to the collection of larger datasets, thanks to a wider field of view, a higher detection efficiency as well as the increase in yield provided by laser-pulsing capabilities.
- Stronger quality demands on the analyses and increasingly more complex approaches are a reality as well in APT.
- Many individuals in the APT community see value in opening up software and file formats in an effort to improve on the documentation of the existing software. In addition, they also see value in reporting more detail about the data-acquisition, the post-processing methods, and the workflows that are used in an effort to optimize the research process. Motivated by an increasingly large part of the community it represents, the International Field Emission Society set up a Technical Committee that, in parts, oversees and helps coordinate these actions.
- With the stronger permeation of machine learning and artificial intelligence methods into a variety of fields, one may argue that missing documentation or undisclosed data reduce the speed at which new data analysis techniques can be developed, tested for their effectiveness, and broadly deployed.
- Journals and funding agencies are likely to start enforcing stricter quality demands with respect to the curation of experimental data.

One solution to cope with the above challenges is to improve the documentation and curation of experimental data and make these as comprehensive and automated tasks as possible. This aligns with the goals of the FAIR research and data stewardship principles^{6,57}. The acronym FAIR stands for research which is findable, accessible, and not only interoperable by humans and machines, but also reproducible, or ideally even repurposable, for applications in other research fields. Only concerted efforts across the community could bring research in experimental materials science closer to becoming compliant with the fair principles. This will be rewarding because methods from scientific computing can be better utilized and with this especially those manual procedures reduced, which are prone to user errors. Examples for this are the application of artificial intelligence tools, the here discussed high-throughput analyses, scientific visualization, and wizards for automated report writing.

Employing software parallelization, i.e., methods and tools from scientific computing, is another solution to improve the efficiency of APT data post-processing. Aware of the fact that not all readers are familiar with scientific computing, we recap key concepts in the supplementary methods. There are only a few examples that have started to explore the above potential of using scientific computing hardware and programming methods for APT^{58–61}. Maybe this situation has been caused by placing in the past a stronger focus on addressing scientific questions, plus getting funding for such is easier, rather than for questions on the software tools for answering scientific questions in the experimental materials science. In communities where most users are not (yet) frequently trained in software engineering or data science, this situation demands action. There are many advantages to adapting or using scientific computing in APT research, as we will see in the following.

We acknowledge that APT practitioners, like many other experimental scientists, feel comfortable with using primarily proprietary software. Yet many are open to do so in conjunction with a box of highly performing, community-led tools, which represents a versatile approach to tackle many of the challenges discussed above. Scripting options through Python and Matlab allow for assembling these tools into sophisticated workflows, as we will show^{5,6} to align better with the fair principles. These tools and workflows can in turn be interfaced with commercial software^{45,46}. In summary, we are convinced it is worthwhile to develop APT software tools, which complement rather than replace commercial software like IVAS. Furthermore, we should, as a community aim for high-throughput analyzing and automation.

In this spirit, we contribute paraprobe. The software is our first step towards open boxes of scientific computing tools for high-throughput processing of point cloud data, here exemplified for APT. Paraprobe currently does not allow to execute all the analysis tasks that are implemented in typical commercial software, but offers a platform on which more can be built. In addition, the analyses that are implemented support hybrid parallelism for computational geometry, spatial statistics, clustering, and atom probe crystallography⁶². These are functionalities that are currently not available within commercial software. Computational and experimental materials scientists can learn from each other. They face similar challenges when it comes to the development of research software bottom-up by scientists and aligning such software and descriptions of the data and metadata better with the aims of the fair data stewardship principles. Intensifying the interaction between the communities is beneficial because APT is analyzing point cloud data, which accrue similarly as atoms or material points in simulations at different length scales within many computational materials science communities. Therefore, this work can also inspire developers and users of specific point cloud processing methods within the computational materials science.

RESULTS AND DISCUSSION

Application to experimental datasets

Delivering significant and substantiated quantitative descriptors for answering materials science questions is the role of microscopy and microanalysis techniques like APT. The specific role of APT for answering materials science questions in alloy design is to deliver quantitative data at the near-atomic scale about the spatial distribution of solutes and thermodynamic phases, the number density of such phases, their composition, their spatial arrangement as well as their approximate volume and shape. Combined with other microscopy techniques, this enables to unveil specific physical mechanisms at the nanoscale and link these to the thermo-mechanical evolution and the properties of materials.

With paraprobe, we have developed a box of high-throughput tools for this. We show the benefit of these tools in typical microstructure characterization studies within alloy design. Key results of the first case study are reported in this section, while all other details are reported in the supplementary methods, including systematic benchmarks of the software. The tools are offered as open-source software that are connectable into customizable workflows using Python. The tools are supplemented by tutorials to help users with exploring the case studies in more detail or taking these as a start for setting up analyses for their own datasets.

We report the application of paraprobe in two case studies: one, which is taken from additive manufacturing (AM) research on Al-Sc-Si alloys, plus another one which is taken from alloy design of an Al-Zn-Mg-Cu aerospace alloy. There are 13 specimens in total for the first and seven specimens in total for the second case study. Several reasons motivated working with these alloy systems: Both are of interest for aerospace applications, in which case material strength is a target property. Therefore, characterizing precipitates and solutes is of key interest. Both alloy systems can also be seen as representative examples for different challenges faced when quantifying the nanochemistry: Al-Zn-Mg-Cu alloys display a mixture of nanoscale solute gradients in the matrix and different precipitate phases, which challenge tasks like clustering. Compared to classical aging treatments of aluminum alloys, additively-manufactured material experiences a different thermal treatment. Given that elements can have different melting points and diffusion rates, this results in a different thermo-chemo-mechanical evolution of precipitates in AM-processed alloys. Al-Sc-Si is one example where the size of the precipitates within even the same specimen can differ substantially inasmuch as some precipitates qualify as coarse intermetallics, while some precipitates are so small that they are practically not reliably resolvable with APT⁶³. In effect, multiple sources of inaccuracies have to be mastered when trying to understand the early stages of precipitation using APT (finite counting effects, effects of parameterization of the clustering algorithm, and ambiguities whether the cluster is better described at the atomic or the continuum scale, and which category of precipitate does it represent best terminology-wise). In this regard, the alloy systems can serve as examples how paraprobe can help practitioners master these challenges via uncertainty quantification.

Quantifying the distribution of the sizes and the number density of second-phase precipitates are the sought-after descriptors in the first case study. The link to computational materials science is that these are two frequently used descriptors to calibrate and benchmark kinetic models of precipitation thermodynamics at the mesoscopic scale. Such task is typically achieved with clustering methods. Here, we detail how such methods can be looped into high-throughput workflows, exemplified for the maximum separation method (MS), a DBScan variant, which is still one of the most frequently applied methods for studying precipitates in APT. Figure 1 summarizes the results of characterizing the early stages

of forming Al₃Sc precipitates via inspecting the spatial distribution of scandium atoms and the distribution of precipitate sizes within the specimens from the three investigated samples of the AM case study. Different core point distances (d_{\max}) were probed during the maximum separation clustering study.

The results vary systematically, both qualitatively and quantitatively, with the prior thermal history of the samples studied. For any one sample type, the results are fully consistent. For the datasets from the incipient and the intermediate samples, Fig. 1a) documents that a key assumption relevant to apply the MS method is violated: the individual spatial distribution functions of the atoms within clusters do not differ substantially from the distributions of the matrix atoms^{38,64}. This observation is particularly evident for the 1NN distributions (Fig. 1). Consequently, any interpretation of number densities at virtually all d_{\max} values for the incipient and intermediate specimens is inaccurate. This is especially visible for the global maxima of the d_{\max} curve.

An inspection of the precipitate size distributions Fig. 1b) pinpoints the shortcomings of applying the MS method to the incipient and intermediate state specimens. The distributions show that as many as 25% of the identified clusters contain only five scandium atoms, i.e., the minimum accepted count N_{\min} . Again, this is a clear argument against using the MS method for quantifying the early stages of precipitation in those AM specimens.

In contrast, the scandium 1NN distribution for the dataset of the specimen from the mature state sample is bimodal. In this case, the MS method delivers reliable precipitate number densities as Fig. 1a) confirms. It is reasonable to report the plateau value of the curve as the metallurgical relevant number density for two reasons: first, for this d_{\max} , most scandium atoms contribute to the precipitates rather than to the matrix. Second, for this d_{\max} , a potential bias in the 1NN distribution due to an accidental fusing of solute scandium atoms in the vicinity of the precipitates is lower compared to number densities read at larger d_{\max} values. In summary, these results reassure the validity of earlier findings pertaining to the application of the maximum separation clustering method^{38,64}.

Figure 1b) compares the size distributions for all clusters with the distribution for exclusively those clusters in the interior of the dataset. For the mature state, the shape of both distributions is very similar. Individual quantile values are shifted, though, in particular for the lower half of the curve. We can thus conclude that two contributions affect the distributions: the truncation of clusters by the edge of the dataset and the lower absolute number of clusters in the mature state specimen, i.e., finite counting effects. By contrast, for the incipient and intermediate state the distributions are more similar because more precipitates are included in the volume.

Given that many APT datasets may contain only a few hundred mature clusters, quantifying also such statistical effects is relevant uncertainty quantifying in addition to studying the effect of method parameterization. Our work provides additional value by delivering quasi unbiased distributions of the precipitate size (Fig. 1b)). The distributions are quasi unbiased because of the capability to detect which clusters were truncated by the dataset edge. In effect, the AM case study substantiates how paraprobe delivers additional confidence and detailed uncertainty quantification.

Practicality and relevance of the methods

The numerical costs of the above analyses are summarized in Fig. 1c). Executing the parameter sensitivity study on the three exemplar datasets for specimens of all three samples took 52 min when using 36 threads. This includes all 241 MS clustering runs per dataset, the tessellation, α -shape edge computation, and atom-to-edge distancing. Clustering was the most costly task with a total execution time fraction of 43.2%. With only 1.0% of the

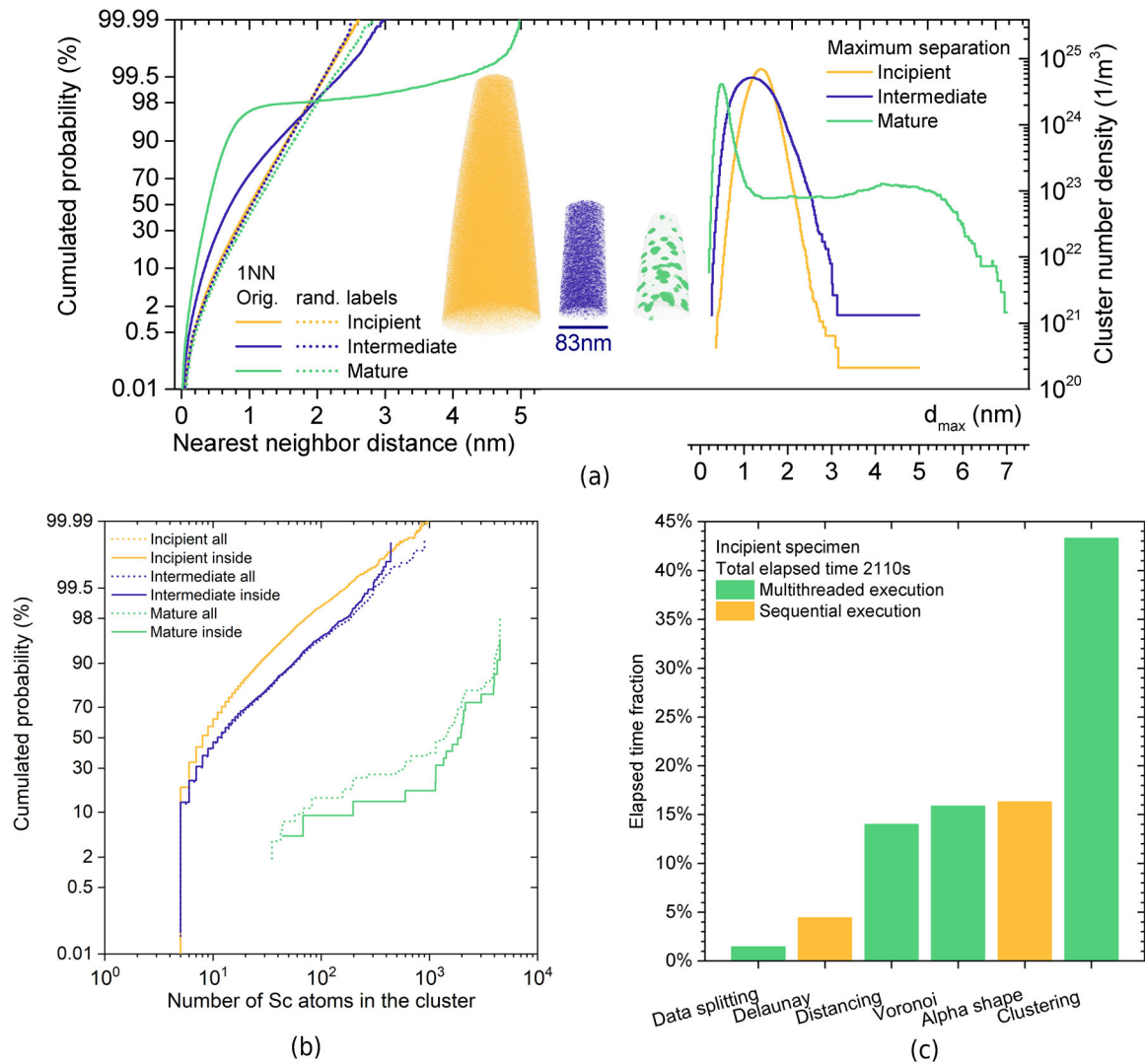


Fig. 1 We show key results of the additive manufacturing case study. Sub-figure **a** compares the reconstructed datasets of three of the 13 specimens. One representative dataset for each sample (incipient, intermediate, and mature) is shown. The color-coding distinguishes the reconstructed datasets and displays their scandium atoms. On the left side the distribution of Sc-Sc nearest neighbor distances (1NN) are compared for original (orig.) and randomized (rand.) atom type labels. On the right side the sensitivity of the cluster number density is shown as a function of d_{max} , the distance between core points while querying atom neighbors with the maximum separation (MS) method. Sub-figure **b** compares the distribution of cluster sizes for the specimens and with respect to edge effects for distributions, which consider only the interior (solid lines) or all clusters (dashed lines), respectively. Sub-figure **c** summarizes the most time consuming parts of the computations, exemplified for the incipient dataset.

total elapsed time, I/O expenditures were negligible, which is another difference compared to proprietary tools.

The supplementary methods and the supplementary references^{65,66} detail the results for the other ten datasets of the AM case study. Comparing the automatically compiled reports (in the supplementary references^{65,66}) shows that all specimens for each thermo-mechanical state yield reproducible results for all three states (incipient, intermediate, mature), respectively. The high-throughput screening with paraprobe delivers for which cases interpreting the precipitation state via the MS method is justified and for which it is not. Furthermore, the tools delivered all relevant spatial statistics, corrected for bias.

Thanks to the high-throughput approach, these analyses took few steps: First, the writing of a Python script to specify which datasets and analysis tasks should be executed via paraprobe-parmsetup. Second, the running of the paraprobe tools on the cluster. Third, the stitching together of a Python script for creating the figures and the reports via paraprobe-autoreporter.

With the combination of trivial and non-trivial parallelism, all 13 datasets were processed in a few hours. The tools could be easily extended to loop in the processing of other analysis tasks, including different clustering methods of the community for making heads-up assessments.

Immediate potential for further parallelization is available but has not been tapped in this study. We want to emphasize that the parameter runs were executed sequentially but each run of the MS method multithreaded. Alternatively, the parameter runs could be distributed trivially parallel on multiple computing nodes. This would result in hybrid-parallelized execution.

We should mention that paraprobe implements methods for processing point cloud datasets with additional mark data per point. Although, exemplified here for mark data, which are specific for APT (mass-to-charge and respective atom type label), it is possible to process also point cloud data from other sources such as molecular dynamics simulations or (material) point data of the microstructure evolution modeling community⁶⁷. This would

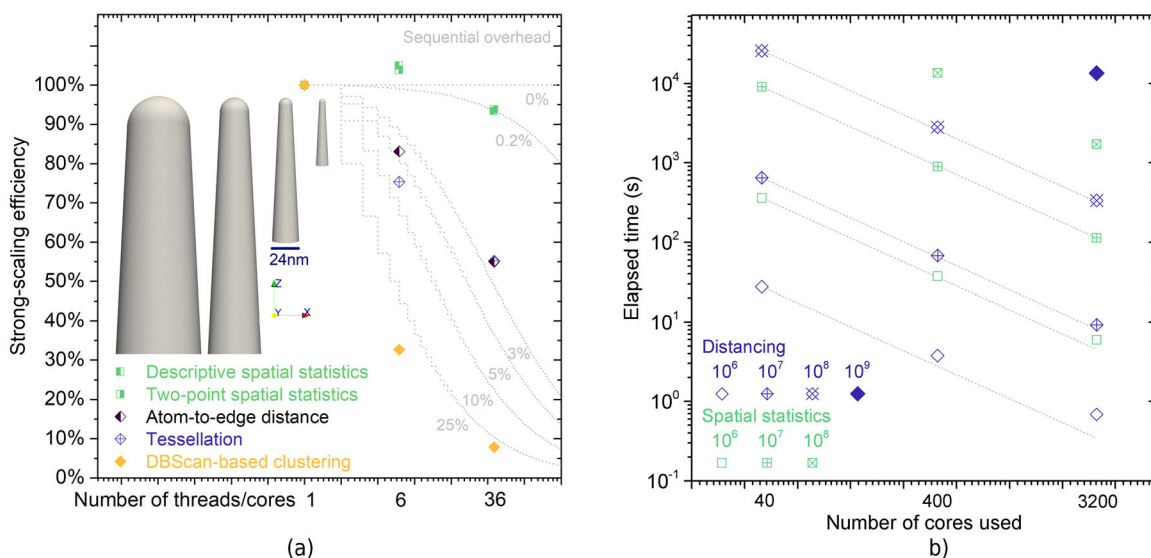


Fig. 2 We document strong-scaling efficiency for multithreaded and hybrid-parallel execution. The inset in **a** displays the synthetic datasets. These are rendered at scale (visualized via the α -shape of the edge). For the two leftmost datasets only their upper sections are shown to retain sufficient pixel resolution. The gray curves in **a** compare the strong-scaling results for different fractions of remaining sequential overhead according to Amdahl's law¹⁰⁰. Sub-figure **b** reports elapsed time. The thin dashed lines compare to the theoretical optimum of linearly scaling methods. Runtime differences were within the thickness of the data point symbols.

require, though, a modification of the paraprobe-transcoder tool to equip paraprobe with a reader for parsing x, y, z position and mark data from the file formats of the respective scientific communities. Thereafter, the same processing pipeline could be used to characterize computational geometry, spatial statistics, and clustering.

Verification and scalability

Reliable verification and software benchmarks call for ground truth data. Therefore, we created two groups of four synthetic datasets (Fig. 2a), left inset). Each dataset was built as a conical frustum with a spherical cap on top⁶⁸. Using a fixed shape, the dataset volume was scaled to contain 2×10^6 , 20×10^6 , 200×10^6 , and 2000×10^6 atoms, respectively. The first group contains four datasets for defect- and noise-free synthetic aluminum single crystals. The second group represent a copy of each dataset from the first group. For these copies $\sim 10\%$ of the atoms were replaced in total by replacing dataset volume by an ensemble of randomly dispersed spherical Al₃Sc precipitates. Details are given in the supplementary methods.

Figure 2 summarizes the key results of the performance assessment. Specifically, Fig. 2a) summarizes the strong-scaling efficiency for multithreaded execution on a workstation, or single computing node, respectively. An analysis of the individual methods as well as the memory consumption is detailed in the supplementary discussion and supplementary Fig. 3). To the best of our knowledge, this is the first such assessment of multithreaded APT tools for such a diverse set of analysis tasks. Paraprobe shows at least 55% strong-scaling efficiency when using 36 threads for all tasks but clustering. The scaling limitations for clustering are attributable to a sequential overhead, which can become as high as 25%. One contribution to this overhead is unavoidable because certain steps of the MS algorithm enforce synchronization⁶⁹.

For all other tasks also a few percent sequential overhead remains although already techniques were employed to balance dynamically the computational load. It is this overhead that results in disproportionately lower efficiency when using more threads. One key contribution to overhead is that the workload per atom, such as during tessellating, typically differs. This sets a limit with

respect to how perfectly the atoms can be distributed and processed as groups of atoms across the cores.

Figure 2b summarizes the results of combining OpenMP-multithreaded data parallelism with process data parallelism using the Message Passing Interface (MPI) library⁷⁰. Here, we exemplify an application for distancing the atoms to the α -shape and processing spatial statistics for the first group of synthetic datasets. As an example, a computing cluster with 80 nodes with 40 cores each was employed. Distances were computed for all atoms within closer than $d_{\text{srf}} = 10$ nm to the dataset edge. Al-Al spatial statistics (radial distribution function (RDF), kNN, with $k = 1, 10, 100$, and three-dimensional spatial distribution maps (SDMs)) were characterized. The radii for the regions-of-interest (ROIs) were set to $r = 10$ nm for the RDF and kNN. The ROI radii for the SDMs were set to 2 nm. SDMs were discretized in $(0.025 \text{ nm})^3$ cubic voxels.

The results confirm in all cases that, in addition to the gains from multithreading, MPI unlocked as large performance gains as additional cores were commissioned. The scalability is close to ideal, as expected for this moderate number of cores and weak coupling of the computation. The more atoms each core processes, the more effective the additional parallelization layer becomes. The reason is that the overhead for any organization of the workload gets better compensated for. The results in Fig. 2 document the strong-scaling nature of paraprobe. Additional details of the verification and benchmarking is summarized in the supplementary discussion and supplementary Figs. 2–4.

To conclude, this paper delivers a concept and box of open-source software tools for enabling high-throughput computational studies with atom probe tomography datasets and experiments. Exemplified for datasets with at most two billion ions, we deliver specific parallelized solutions for solving the following data mining tasks:

- Build α -shapes to the entire point cloud, which does not downsample close to the edge of the point cloud via an original filtering algorithm.
- Compute exact atom-to-edge distances with which edge effects for spatial statistics and precipitate size distributions can be practically eliminated.

- 6
- Fast tessellating of the entire point cloud without a need for downsampling.
 - Exemplary implementation of how to use HDF5 as an open file format for storing APT data and metadata. Thereby, our study shows how to achieve improved I/O speed, use better assistance for scientific visualization, and become prepared for studies that seek to better align with the fair data stewardship principles.

The results document at least 55% strong-scaling multithreading efficiency when using 36 OpenMP threads. With an additional layer of MPI process data parallelism, we unlock approximately three orders of magnitude faster processing compared to sequential execution when using an exemplar computing cluster with 3200 cores.

METHODS

Principle design, high-throughput workflow, and implementation

We would like to point the reader to the supplementary methods and the open-source material to explore in more detail the advanced aspects of the parallelization and the individual methods of paraprobe. In summary, Fig. 3 displays the principle setup of a workflow with paraprobe. Instead of a monolithic program, paraprobe is a collection of parallelized tools. Targeting workstations and computer clusters, paraprobe is instructed via Python scripts in the front-end, which write shell scripts for the backend tools. To assist the users with creating configuration and run files, we developed Python classes. As an alternative route to configure the individual tools, we developed a web-browser-based Python/Bokeh GUI⁷¹. Hands-on examples of the Python scripts and the raw data are offered as tutorials in the form of jupyter notebooks in the source code repository (see code availability section). These exemplify the high-throughput workflow for the two case studies.

The raw data of an APT experiment contains a collection of detector hit positions, time-of-flight, and voltage curve data. For Cameca/Ametek instruments these results are stored in proprietary container formats (RHIT until IVAS v3.6, HITS since IVAS v3.8). There is currently no generally working option to parse all content from such files without IVAS/APSuite. Therefore, all analyses with paraprobe, as far as this paper is concerned, were performed in reconstruction space. For this work, we relied on a priori existing ranging information generated for instance with IVAS. In the future one could inject alternatively the ranging tools from the community (e.g., refs. ^{12,72}) as an additional step into the workflow. Input to paraprobe is passed via POS, EPOS, APT, RNG, or RRNG files⁷ using the paraprobe-transcoder tool. Alternatively, synthetic datasets, like those used for benchmarking, can be created for arbitrary crystal structures with the paraprobe-synthetic tool.

We implemented paraprobe as a collection of C/C++ tools coordinated by Python scripts. The analyses in this work were executed on two computers: The multithreaded runs were processed with an in-house Linux workstation with 36 cores. The hybrid runs were processed on TALOS, a Linux computing cluster with 80 nodes with 40 cores per node. All machines were used exclusively, threads were pinned and placed machine-topology-aware. Details are summarized in the supplementary methods.

Before analyzing, the point cloud is split spatially into a stack of non-overlapping cuboidal point cloud regions. We split along the direction of the longest dataset axis such that each region contains a quasi equal number of N/N_{thr} atoms⁷³ with N the atom total and N_{thr} the number of threads. Each region administrates its own array of atoms. The memory management of the regions was implemented via multithreading, Open Multi-Processing (OpenMP) to be specific⁷⁴. We use strategies for advanced memory management⁷⁵ and detail these in the supplementary methods.

Quantifying the edge of the dataset

Every accurate analysis of spatial quantities for a finite dataset needs a strategy for curing edge effects. Such can arise when inspecting the long-range neighborhood of atoms at the edge of the dataset^{38,41}. One strategy could be to identify the shortest distance of an atom to the edge and use this distance as a criterion to exclude atoms from an analysis to avoid bias.

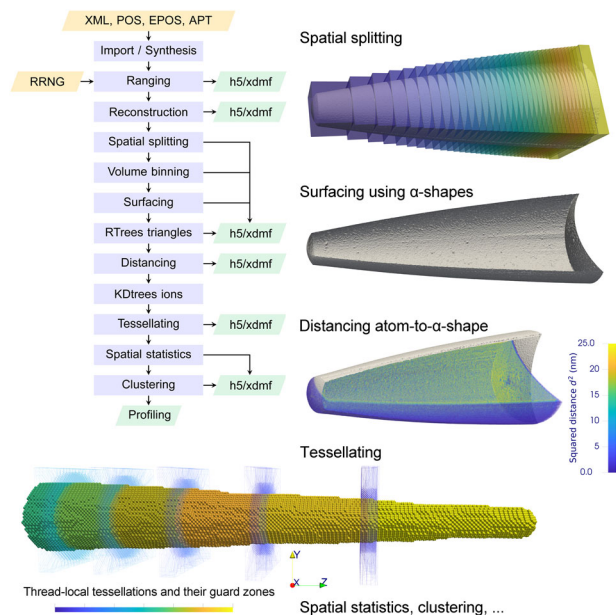


Fig. 3 Paraprobe delivers a collection of tools for implementing high-throughput workflows for automated post-processing of single or collections of APT datasets. Each tool serves a specific analysis task. Examples of tools and key processing steps are shown to the right. All results and metadata are exchanged via HDF5 files. A typical workflow has three steps: First, the desired workflow (of analysis tasks) is described via writing a Python script using the paraprobe-parmsetup classes. The script creates the necessary configuration files and a shell script for executing the workflow. Second, the analysis tasks are executed by individual paraprobe tools using a workstation or computer cluster. The flowchart to the left shows a typical combination of such analysis tasks. Third, the results are post-processed through executing a Python script using the paraprobe-autoreporter classes. These classes parse the metadata and results from their respective HDF5 files and create the desired figures alongside a PDF report. The images on the right side detail the functioning of the data splitting and show a slice of an α -shape with which the dataset edge is triangulated. The bottom image shows a tessellated dataset. Here, also the guard zones are visualized (light-blue wireframes). The guard zones are used to ensure a consistent computing of Voronoi cells at the boundary of each dataset region.

The same strategy can be applied to precipitates when they are only partially analyzed, i.e., truncated by the dataset edge.

One strategy to define the edge is to construct a triangle hull to the point cloud. A variety of methods exist and have been used for APT data: convex hulls^{41,76}, α -shapes as a generalization of convex hulls^{43,77}, or γ -shapes as a generalization of α -shapes⁷⁸. The benefit of α - and γ -shapes over convex hulls is that they can account for concave sections of the point cloud.

We offer a solution to reduce the numerical costs for processing α -shapes. Different to previous authors, who employed downsampling or accepted to work with a sub-set of the data only, we developed a method, which retains the accuracy of the point cloud close to the edge.

The key observation is that most atoms in the interior of the point cloud do not contribute a supporting vertex of a triangle to an α -shape. Consequently, a filtering algorithm is proposed, which filters out these interior atoms. Thereby, only the relevant atoms have to be computed. For multi-million atom datasets and larger, there are typically at least two orders of magnitude fewer of these relevant atoms than there are interior atoms. This enables the processing of even the multi-hundred million atom datasets. The details of the filtering algorithm are described in the supplementary methods. The subsequent α -shape construction has two steps: first, the computation of a Delaunay triangulation of the filtered atom point cloud^{76,79}. Second, the triangulation of α -shapes for specific α

values. Both steps were executed sequentially. In the future parallel implementations of these steps could be added.

Atom-to-edge distancing

An α -shape offers a triangulated representation of the edge of the dataset. This enables the computation of atom-to-edge distances. Approximate and exact analytical methods can be used. Paraprobe computes distances d analytically. These distances can then be evaluated against a threshold distance d_{thr} . Such a threshold can define the thickness of a skin to quantify the dataset edge. The skin starts at the dataset edge and extends inwards to identify which atoms are counted as neighbors but should not be visited with own regions-of-interest to eliminate bias.

The key challenge when computing exact distances is that potentially a large number of atom-to-triangle tests have to be evaluated. Therefore, paraprobe implements a multi-step filtering algorithm, which reduces the number of atom-to-triangle tests per atom: First, a coarse distance is evaluated. Second, this value is used to identify a smaller set of candidate triangles via an R-tree^{80,81} of the α -shape. Details are given in the supplementary methods. All distancing works multithreaded. Atoms are machined off region after region. For each region all threads process atoms via dynamically-scheduled multithreading. Based on this general and fast strategy to compute distances between atoms and triangles, we are currently extending paraprobe to enable also the computation of distances to iso-surfaces and geometrical primitives.

Descriptive and two-point spatial statistics

Spatial statistics⁸² characterize the spatial environment of points. Different probability density functions and their distributions are commonly used for APT data. Examples are kNN, i.e., the distribution of distances between atoms of a certain type and their individual k^{th} -nearest neighbor (of a certain type); or the RDF^{83,84}. RDFs link to methods for small-angle X-ray scattering^{35,63,85}. RDF and kNN represent annularly integrated representatives of the more general, so-called two-point (spatial) statistics⁸⁶.

These are functions that quantify three-dimensional probability mass values, which describe how many neighboring atoms of a particular type the atoms have in a particular direction \mathbf{r} and radial distance R . Serial sections of these functions in the central atom's plane of location are better known by atom probers as SDMs⁸⁷. Paraprobe implements all the above-mentioned spatial statistics with a customizable binning with rectangular transfer functions and hybrid parallelization.

We refer to a single combination of central atoms and (their) neighbors as a spatial statistics query task. Paraprobe enables users to a priori formulate a list of combinations of multiple statistics, multiple query task combinations, and multiple atom types. At runtime, this task list is machined off with an internal batch queue processor, whose details are explained in the supplementary methods.

Detection of clusters or precipitates

Clustering algorithms applied to APT data^{38,63,88} enable the quantification of the number density and the distribution of sizes for phase regions or clusters, precipitates respectively, if these phase regions are crystalline. For reasons of practicality, the terms cluster and precipitate are used interchangeably in this work. A variety of clustering methods has been reported^{7,38–40,89}. Especially variants and generalizations of the DBScan⁹⁰ clustering algorithm, such as the maximum separation (MS) method^{37,64}, the core-linkage³⁸, or the hierarchical DBScan⁴⁰ method are employed for APT data. Given the importance of DBScan variants, we decided to focus in this work on their potential for parallelization. Paraprobe thus executes the OpenMP-parallelized DBScan implementation of Götz et al.⁶⁹

We modified their code to enable batch execution of individually multithreaded DBScan runs. As a specific variant of DBScan, the MS method requires the calibration of parameters: the core point distance d_{max} (equivalent to ϵ in the original DBScan reference⁹⁰), the number of neighboring points within d_{max} distance to call the point a core point (here $k=1$), and the minimum number of atoms to consider a cluster as a significant one N_{min} . All spatial queries of atoms or triangles use efficient spatial indices to reduce unnecessary queries and reduce the costs of individual queries. The details are explained in the supplementary methods.

Parallelized volume tessellations

A tessellation is an overlap-free distributing of a space⁷⁶ for a given set of points and a mathematical space distributing rule. The rule that defines a Voronoi tessellation, i.e., which assigns each position in space to the individually closest member of the point cloud, yields several useful results for APT data: a defined volume, and thus concentration value per atom^{91,92}, three-dimensional Voronoi cells with a topology, which are useful for cluster identification⁹², and cell facets, with which microstructural features^{67,92,93} can be reconstructed.

Despite these benefits, the construction of tessellations for APT datasets with more than a few million atoms faced so far unsolved challenges because existent computational geometry libraries were used sequentially and out-of-the-box. Paraprobe breaks with this strategy. Instead, we build on a solution from the cosmology community: the key idea is to split the tessellation task first into multiple smaller tessellations. Second, these are fused at the edges. Following this idea of Peterka and coworkers⁹⁴, paraprobe splits the tessellation of the entire point cloud into as many tessellations as there are regions. Now these regions are independent and thus processable via multithreading. For this purpose, we implemented a multithreaded wrapper around the Voro++ library. Each thread processes one region.

Guard zones were attached on either side of a region and exact partial copies of the point clouds from the adjoining regions copied to ensure that also the cells at the region edge are computed with correct individual shapes. Figure 3 shows an example of these guard zones (light-blue wireframes) and the resulting tessellation for six threads. Additional implementation details are described in the supplementary methods.

Efficient storing and sharing of data and metadata via HDF5

File formats with open specifications offer a transparent way for storing APT data and metadata. We are convinced that examples like the Hierarchical Data Format (HDF5)⁹⁵ offer a more performant tool than the traditional formats and I/O strategies, which the APT community applied in the past. In concert with open metadata standards plus a to be developed ontology, this enables the APT community to store and align their computational workflows better with the aims of the fair data stewardship principles^{6,57}. Interestingly, these practical advantages remained so far virtually unexplored for APT.

Therefore, we implemented a proof-of-concept how metadata, via HDF5, could be used for managing post-processed data and metadata within APT, also to support activities of the International Field Emission Society Technical Committee. HDF5 has advantages over traditional file formats: the source code is open, the library offers in-place compression functionality, and is optimized for both sequential and distributed-memory parallel I/O. Interfaces for many programming languages exists for accessing HDF5 to help organizing data. For these reasons, paraprobe stores all output in HDF5 files.

Experiments and datasets of the case studies

We analyzed specimens from two alloy design case studies, which cover typical examples of the context in which APT measurements are embedded when answering materials science questions. It is common in such studies to prepare multiple specimens for a given thermo-mechanical state because on the one hand specimens can fracture early during an APT experiment, and thus create insufficient atom count, and on the other hand it is good practice to assure reproducibility between the repeat specimens (at least to the level it is practically possible for a method that probes the nanoscale). The first case study contains specimens from additively-manufactured samples of a research project on characterizing the effects of intrinsic reheating on the precipitate population during AM operations. The specimens contained clusters and precipitates in different growth states. These states are referred to as the incipient, the intermediate, and the mature state, respectively (Fig. 1). We chose these terms to distinguish the specimens because they reflect that the precipitation process is in different stages with respect to how much solute concentration remains left in the matrix in front of the precipitates.

The samples in the incipient and intermediate state were produced via directed energy deposition (DED)⁹⁶ from an Al–0.49Sc–0.45Si (wt.%) alloy in the as-produced state. Clusters in the DED sample formed in response to the intrinsic reheating of the deposited layers during AM. Multiple specimens were taken from the bottom and the top parts of the sample, yielding the incipient and the intermediate states, respectively. The sample in the mature state originates from an Al–0.44Sc–0.02Si (wt.

%) alloy, which was processed via laser powder bed fusion (L-PBF)⁹⁶. After building by AM, the mature state sample was heat-treated at 350 °C for 10 h. In response to this aging treatment, the precipitates grew to an average diameter of 20 nm approximately. The specimens were prepared with a lift-out procedure after xenon ion milling with an FEI Helios PFIB dual-beam focused ion beam scanning electron microscope⁹⁷.

Characterizing the number density of scandium-bearing precipitates and their sizes, was the specific aim of the AM case study. Such pieces of information about the nanochemistry serve typically as input for formulating mean-field kinetic models⁹⁸. The presence of cluster of different diameter and non-negligible scandium and silicon content in solid solution creates the need to quantify the uncertainties in the number density and the size distributions. This should be best practice for every such alloy design study with APT, regardless which particular method for an analysis task, here clustering, one employs. We exemplify such high-throughput uncertainty quantification for the maximum separation method^{37,64}.

First, an α -shape was computed for each dataset via filtering with cubic bins $d_{\text{bin}} = 0.5$ nm. Thereafter, nearest neighbor (1NN) spatial statistics were characterized for Sc-Sc using spherical regions-of-interest (ROIs) with $r = 5$ nm radius. ROIs were placed at scandium atoms with at least $d_{\text{srf}} \geq 2.0$ nm distance to the dataset edge. Statistics were computed for original (orig.) and randomized (rand.) atom type labels. Uncertainties due to parameter sensitivities of the maximum separation method were quantified by probing 241 linearly spaced individual runs. The core point distances d_{max} ranged between 0.2 nm and 5.0 nm with 0.02 nm step. Clusters with less than $N_{\text{min}} < 5$ were not considered. Tessellations were built for all atoms. The Voronoi cells of atoms within $d_{\text{ero}} < 1.0$ nm distance to the dataset edge were eroded. The number of scandium atoms ranged from 8.64×10^4 to 6.74×10^5 . Further details are documented in the settings files of the supplementary reference⁶⁵.

As the second case study we analyzed Al-Zn-Mg-Cu specimens from Zhao et al.³⁵, seven specimens in total. The experimental methods and reconstruction protocols are detailed in the original paper. The case study quantifies the matrix and precipitate concentrations for zinc, magnesium, and copper solutes in an effort to understand better the artificial aging response in such 7XXX series alloys with aerospace applications. The authors of the original study³⁵ discussed therein a method, which post-processes several spatial statistics (RDF, and 1NN) for each type of solute. The motivation behind this is to explore links between spatial correlation functions from small-angle scattering and those used in APT^{35,63,85}. The main practical challenge is that large ROI radii (2.5 nm to 10 nm and beyond) are probed. With IVAS this is known to be particularly costly. Also, due to the lack of a documentation how the above spatial statistics in IVAS are in detail computed, it is essential to employ rigorous strategies for the removal of edge effects. Paraprobe solves the above analysis tasks not only faster but also with rigorous uncertainty quantification. All detailed settings are given in the supplementary reference⁶⁶.

DATA AVAILABILITY

All datasets, code, and results of the additive manufacturing⁶⁵ and the Al-Zn-Mg-Cu⁶⁶ case studies are offered open-source. This includes the reconstructed APT datasets as POS/EPOS files, the processed results, and the Python source code for generating the associated workflows. Python scripts were developed (paraprobe-parmsetup) to generate the input for running the paraprobe tools. Python scripts were also developed (paraprobe-autoreporter) for composing figures and creating automatic PDF reports from the HDF5 files of the tools. To help potential users with setting up such scripts, we wrote Jupyter-notebook-based tutorials in the online documentation. We have split the repository, for practical reasons, into the configuration, settings, input POS/EPOS, and essential results files⁹⁹. The entire repository of compressed data from all benchmarks occupies several terabytes because they include the geometry of every cell of the multi-hundred million atom synthetic datasets. These data are available from the authors upon serious request.

CODE AVAILABILITY

The source code of paraprobe and a documentation is maintained online: • <http://gitlab.mpcdf.mpg.de/mpie-aptfm-toolbox/paraprobe> • <http://paraprobe-toolbox.readthedocs.io> The repository contains also CPU- and GPU-parallelized tools for atom probe crystallography, which will be assessed in a future study⁶².

Received: 11 April 2020; Accepted: 12 December 2020;

Published online: 29 January 2021

REFERENCES

- Curtarolo, S. et al. The high-throughput highway to computational materials design. *Nat. Mater.* **12**, 191–201 (2013).
- Pizzi, G., Cepellotti, A., Sabatini, R., Marzari, N. & Kozinsky, B. AiiDA: automated interactive infrastructure and database for computational science. *Comput. Mater. Sc.* **111**, 218–230 (2016).
- Montoya, J. H. & Persson, K. A. A high-throughput framework for determining adsorption energies on solid surfaces. *Npj Comput. Mat.* **3**, 14 (2017).
- Himanen, L., Geurts, A., Foster, A. S. & Rinke, P. Data-driven materials science: status, challenges, and perspectives. *Adv. Sci.* **6**, 1900808, <https://doi.org/10.1002/advs.201900808> (2019).
- Janßen, J. et al. pyiron: An integrated development environment for computational materials science. *Comput. Mater. Sc.* **163**, 24–36 (2019).
- Draxl, C. & Scheffler, M. in *Handbook of Materials Modeling* (eds Yip, S. & Androni, W.) (Springer, Cham, 2020).
- Gault, B., Moody, M. P., Cairney, J. M. & Ringer, S. P. *Atom Probe Microscopy*, 1 edn (Springer, New York, 2012).
- Larson, D. J., Prosa, T. J., Ulfög, R. M., Geiser, B. P. & Kelly, T. F. *Local Electrode Atom Probe Tomography*, 1 edn (Springer Science, New York, 2013).
- Lefebvre, W., Vurpillot, F. & Sauvage, X. *Atom Probe Tomography: Put Theory Into Practice*, 2 edn (Academic Press, Amsterdam, 2016).
- Miller, M. K., Cerezo, A., Hetherington, M. G. & Smith, G. D. W. *Atom Probe Field Ion Microscopy*, 1 edn (Clarendon Press, Oxford, UK, 1996).
- Hudson, D., Smith, G. D. W. & Gault, B. Optimisation of mass ranging for atom probe microanalysis and application to the corrosion processes in Zr alloys. *Ultramicroscopy* **111**, 480–486 (2011).
- Haley, D., Choi, P. & Raabe, D. Guided mass spectrum labelling in atom probe tomography. *Ultramicroscopy* **159**, 338–345 (2017).
- Gault, B. et al. Advances in the reconstruction of atom probe tomography data. *Ultramicroscopy* **111**, 448–457 (2011).
- Kirchmayer, A. et al. Combining experiments and atom probe tomography-informed simulations on equation 1 precipitation strengthening in the polycrystalline Ni-base superalloy A718Plus. *Adv. Eng. Mater.* **22**, 2000149 (2020).
- Herbig, M. Spatially correlated electron microscopy and atom probe tomography: current possibilities and future perspectives. *Scr. Mater.* **148**, 98–105 (2018).
- Hono, K. Atom probe microanalysis and nanoscale microstructures in metallic materials. *Acta Mater.* **47**, 3127–3145 (1999).
- Kuzmina, M., Herbig, M., Ponge, D., Sandlöbes, S. & Raabe, D. Linear complexions: confined chemical and structural states at dislocations. *Science* **349**, 1080–1083 (2015).
- Valley, J. W. et al. Hadean age for a post-magma-ocean zircon confirmed by atom-probe tomography. *Nat. Geosci.* **7**, 219–223 (2014).
- Piazolo, S. et al. Deformation-induced trace element redistribution in zircon revealed using atom probe tomography. *Nat. Commun.* **7**, 10490 (2016).
- White, L. F. et al. Atomic-scale age resolution of planetary events. *Nat. Commun.* **8**, 15594 (2017).
- Saxey, D. W., Moser, D. E., Piazolo, S., Reddy, S. M. & Valley, J. W. Atomic worlds: current state and future of atom probe tomography in geoscience. *Scr. Mater.* **148**, 115–121 (2018).
- Cojocaru-Mirédin, O., Schwarz, T. & Abou-Ras, D. Assessment of elemental distributions at line and planar defects in Cu(In,Ga)Se₂ thin films by atom probe tomography. *Scr. Mater.* **148**, 106–114 (2018).
- Perea, D. E. et al. Atom probe tomographic mapping directly reveals the atomic distribution of phosphorus in resin embedded ferritin. *Sci. Rep.* **6**, 22321 (2016).
- Rusitzka, K. A. K. et al. A near atomic-scale view at the composition of amyloid-beta fibrils by atom probe tomography. *Sci. Rep.* **8**, 17615 (2018).
- Voyles, P. M., Muller, D. A., Grazul, J. L., Citrin, P. H. & Gossmann, H.-J. L. Atomic-scale imaging of individual dopant atoms and clusters in highly n-type bulk Si. *Nature* **416**, 826–829 (2002).
- Barnes, J. P. et al. Atom probe tomography for advanced nanoelectronic devices: Current status and perspectives. *Scr. Mater.* **148**, 91–97 (2018).
- Giddings, A. D. et al. Industrial application of atom probe tomography to semiconductor devices. *Scr. Mater.* **148**, 82–90 (2018).
- Kontis, P. et al. The effect of chromium and cobalt segregation at dislocations on nickel-based superalloys. *Scr. Mater.* **145**, 76–80 (2018).
- Li, T. et al. Atomic-scale insights into surface species of electrocatalysts in three dimensions. *Nat. Catal.* **1**, 300–305 (2018).
- Gin, S. et al. Atom-probe tomography, TEM and ToF-SIMS study of borosilicate glass alteration rim: a multiscale approach to investigating rate-limiting mechanisms. *Geochim. Cosmochim. Acta* **202**, 57–76 (2017).

31. Sepehri-Amin, H. et al. Correlation of microchemistry of cell boundary phase and interface structure to the coercivity of $\text{Sm}(\text{Co}_{0.784}\text{Fe}_{0.100}\text{Cu}_{0.088}\text{Zr}_{0.028})_{1.719}$ sintered magnets. *Acta Mater.* **126**, 1–10 (2017).
32. Schreiber, D. K., Perea, D. E., Ryan, J. V., Evans, J. E. & Vienna, J. D. A method for site-specific and cryogenic specimen fabrication of liquid/solid interfaces for atom probe tomography. *Ultramicroscopy* **194**, 89–99 (2018).
33. Chang, Y. et al. Ti and its alloys as examples of cryogenic focused ion beam milling of environmentally-sensitive materials. *Nat. Commun.* **10**, 942 (2019).
34. McCarroll, I. E., Bagot, P. A. J., Devaraj, A., Perea, D. E. & Cairney, J. M. New frontiers in atom probe tomography: a review of research enabled by cryo and/or vacuum transfer systems. *Mater. Today Adv.* **7**, 100090 (2020).
35. Zhao, H., Gault, B., Ponge, D., Raabe, D. & de Geuser, F. Parameter free quantitative analysis of atom probe data by correlation functions: Application to the precipitation in Al-Zn-Mg-Cu. *Scr. Mater.* **154**, 106–110 (2018).
36. Hellman, O. C., Vandenbroucke, J. A., Rüsing, J., Isheim, D. & Seidman, D. N. Analysis of three-dimensional atom-probe data by the proximity histogram. *Microsc. Microanal.* **6**, 437–444 (2000).
37. Hyde, J. M. & English, C. A. An analysis of the structure of irradiation induced Cu-enriched clusters in low and high nickel welds. In *Proc. MRS Fall Meeting 2000: Symposium R-Microstructural Processes in Irradiated Materials* (eds Lucas, G. E., Snead, L. L., Kirk, M. A., and Elliman, R. G.) 650, 6–12 (Cambridge University Press, Cambridge, 2000).
38. Stephenson, L. T., Moody, M. P., Liddicoat, P. V. & Ringer, S. P. New techniques for the analysis of fine-scaled clustering phenomena within atom probe tomography (APT) data. *Microsc. Microanal.* **13**, 448–463 (2007).
39. Zelenty, J., Dahl, A., Hyde, J., Smith, G. D. W. & Moody, M. P. Detecting clusters in atom probe data with gaussian mixture models. *Microsc. Microanal.* **23**, 269–278 (2017).
40. Ghamarian, I. & Marquis, E. A. Hierarchical density-based cluster analysis framework for atom probe tomography data. *Ultramicroscopy* **200**, 28–38 (2019).
41. Haley, D., Petersen, T., Barton, G. & Ringer, S. P. Influence of field evaporation on radial distribution functions in atom probe tomography. *Philos. Mag.* **89**, 925–943 (2009).
42. Felfer, P., Ceguerra, A., Ringer, S. & Cairney, J. Applying computational geometry techniques for advanced feature analysis in atom probe data. *Ultramicroscopy* **132**, 100–106 (2013).
43. Felfer, P. & Cairney, J. A computational geometry framework for the optimisation of atom probe reconstructions. *Ultramicroscopy* **169**, 62–68 (2016).
44. Ulfing, R. M. et al. Hardware and software advances in commercially available atom probe tomography systems. *Microsc. Microanal.* **23**, 40–41 (2017).
45. Reinhard, D. A. et al. Improved Data Analysis with IVAS 4 and AP Suite. *Microsc. Microanal.* **25**, 302–303 (2019).
46. Day, A. C. et al. Recent developments in APT analysis automation and support for user-defined custom analysis procedures in IVAS 4. *Microsc. Microanal.* **25**, 338–339 (2019).
47. Boll, T., Al-Kassaba, T., Yuan, Y. & Liu, Z. Investigation of the site occupation of atoms in pure and doped $\text{TiAl/Ti}_3\text{Al}$ intermetallic. *Ultramicroscopy* **107**, 796–801 (2007).
48. Moody, M. P., Stephenson, L. T., Ceguerra, A. V. & Ringer, S. P. Quantitative binomial distribution analyses of nanoscale like-solute atom clustering and segregation in atom probe tomography data. *Microsc. Res. Tech.* **71**, 542–550 (2008).
49. Moody, M. P., Gault, B., Stephenson, L. T., Haley, D. & Ringer, S. P. Qualification of the tomographic reconstruction in atom probe by advanced spatial distribution map techniques. *Ultramicroscopy* **109**, 815–824 (2009).
50. Yao, L., Gault, B., Cairney, J. M. & Ringer, S. P. On the multiplicity of field evaporation events in atom probe: a new dimension to the analysis of mass spectra. *Philos. Mag. Lett.* **90**, 121–129 (2010).
51. Saxey, D. Correlated ion analysis and the interpretation of atom probe mass spectra. *Ultramicroscopy* **111**, 473–479 (2011).
52. Ceguerra, A. V. et al. The rise of computational techniques in atom probe microscopy. *Curr. Opin. Solid. State. Mater. Sci.* **17**, 224–235 (2013).
53. Haley, D. & London, A. *APTtools*. <http://apptools.sourceforge.net> (2020).
54. Ringer, S. P. *Atom Probe Workbench*. <http://www.massive.org.au/cvl/cvl-workbenches/atom-probe-workbench> (2020).
55. Haley, D. & Ceguerra, A. *3Depict-Visualisation & Analysis for Atom Probe*. <http://threedepict.sourceforge.net> (2020).
56. Keutgen, J., London, A. & Cojocaru-Mirédin, O. Solving peak overlaps for proximity histogram analysis of complex interfaces for atom probe tomography data. *Microsc. Microanal.* 1–9 (2020).
57. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).
58. Seal, S. et al. Tracking nanostructural evolution in alloys: Large-scale analysis of atom probe tomography data on blue gene/L. In *Proc. 37th International Conference on Parallel Processing* (ed. O’Conner, L.) 338–345 (The Institute of Electrical and Electronics Engineers, Inc., Los Alamitos, 2008).
59. Seal, S. K., Yoganath, S. B. & Miller, M. K. Nanoscale cluster detection in massive atom probe tomography data. In *Proc. IEEE International Parallel and Distributed Processing Symposium Workshops*, (ed. O’Conner, L.) 1180–1189 (The Institute of Electrical and Electronics Engineers, Inc., Los Alamitos, 2014).
60. Lu, H., Seal, S. K., Muzyn, G., Guo, W. & Poplawsky, J. D. Efficient, parallel at-scale correlation analysis for atom probe tomography on hybrid architectures. In *Proc. IEEE International Parallel and Distributed Processing Symposium Workshops*, (ed. O’Conner, L.) 54–63 (The Institute of Electrical and Electronics Engineers, Inc., Los Alamitos, 2018).
61. Katnagallu, S. et al. Advanced data mining in field ion microscopy. *Mater. Charact.* **146**, 307–318 (2018).
62. Kühbach, M., Kasemer, M., Gault, B. & Breen, A. On open and strong-scaling tools for atom probe crystallography: high-throughput methods for indexing crystal structure and orientation. Preprint at <http://arxiv.org/abs/2009.00735v1> (2020).
63. de Geuser, F. & Gault, B. Metrology of small particles and solute clusters by atom probe tomography. *Acta Mater.* **188**, 406–415 (2020).
64. Jäggle, E. A., Choi, P.-P. & Raabe, D. The maximum separation cluster analysis algorithm for atom-probe tomography: Parameter determination and accuracy. *Microsc. Microanal.* **20**, 1662–1671 (2014).
65. Kühbach, M. et al. Supplementary material and data to “On strong-scaling and open-source tools for analyzing atom probe tomography data” on the additive manufacturing case study on Zenodo. <http://zenodo.org/record/3906906> (2020).
66. Kühbach, M. et al. Supplementary material and data to “On strong-scaling and open-source tools for analyzing atom probe tomography data” on the Al-Zn-Mg-Cu case study on Zenodo. <http://zenodo.org/record/3904304> (2020).
67. Kühbach, M. & Roters, F. Quantification of 3D spatial correlations between state variables and distances to the grain boundary network in full-field crystal plasticity spectral method simulations. *Model. Simul. Mat. Sci. Eng.* **28**, 055005 (2020).
68. Kühbach, M., Breen, A. J., Herbig, M. & Gault, B. Building a library of simulated atom probe data for different crystal structures and tip orientations using tapsim. *Microsc. Microanal.* **25**, 320–330 (2019).
69. Götz, M., Bodenstern, C. & Riedel, M. HPDBSCAN: highly parallel DBSCAN. In: *Proc. Workshop on Machine Learning in High-Performance Computing Environments* (ed. Kern, J.) 1–10 (The Association for Computing Machinery, New York, 2015).
70. Snir, M., Otto, S., Huss-Lederman, S., Walker, D. & Dongarra, J. *MPI-The Complete Reference, Volume 1, The MPI Core*, 2 edn (MIT Press, Cambridge, 1998).
71. Bokeh Development Team. Bokeh: Python library for interactive visualization. <http://bokeh.org> (2020).
72. Wei, Y. et al. Machine-learning-enhanced time-of-flight mass spectrometry analysis. Preprint at <http://arxiv.org/abs/2010.01030v1> (2020).
73. Patwary, M. A. et al. PANDA: Extreme scale parallel K-nearest neighbor on distributed architectures. In *Proc. IEEE International Parallel and Distributed Processing Symposium* (ed. O’Connor, L.) 494–503 (The Institute of Electrical and Electronics Engineers, Inc., Los Alamitos, 2016).
74. Chandra, R. et al. *Parallel Programming in OpenMP*, 1 edn. (Morgan Kaufmann, San Francisco, 2001).
75. Hennessy, J. L. & Patterson, D. A. *Computer architectures: a quantitative approach*, 5 edn (Morgan Kaufmann, Amsterdam, 2012).
76. Okabe, A., Boots, B., Sugihara, K. & Chiu, S. N. *Spatial Tessellations: Concepts and Applications of Voronoi Diagrams*, 2 edn (John Wiley & Sons, Chichester, 2000).
77. Edelsbrunner, H. & Mücke, E. P. Three-dimensional alpha shapes. *ACM Trans. Graph.* **13**, 43–72 (1994).
78. Cameron, M. E., Sloan, K. R. & Sun, Y. in *Geometric Modeling for Scientific Visualization*. (eds Brunnett, G., Hamman, B., Müller, H. & Linsen, L.) (Springer, Berlin, 2004).
79. The CGAL Project. *CGAL User and Reference Manual* (CGAL Editorial Board, 2018), 4.12 edn. <http://doc.cgal.org/4.12/Manual/packages.html>. (2018).
80. Brinkhoff, T., Kriegel, H. P. & Seeger, B. Parallel processing of spatial joins using R-trees. In *Proc. Twelfth International Conference on Data Engineering*, (ed. Su, S. Y. W.) 258–265 (IEEE Computer Society, Washington D.C., 1996).
81. Balasubramanian, L. & Sugumaran, M. A state-of-art in R-tree variants for spatial indexing. *Int. J. Comput. Appl.* **42**, 35–41 (2012).
82. Cressie, N. A. C. *Statistics for Spatial Data*, 1 edn (John Wiley & Sons, Chichester, 1991).
83. Sudbrack, C. K., Noebe, R. D. & Seidman, D. N. Direct observations of nucleation in a nondilute multicomponent alloy. *Phys. Rev. B* **73**, 212101 (2006).
84. Philippe, T., Duguay, S. & Blavette, D. Clustering and pair correlation function in atom probe tomography. *Ultramicroscopy* **110**, 862–865 (2010).
85. de Geuser, F. & Lefebvre, W. Determination of matrix composition based on solute-solute nearest-neighbor distances in atom probe tomography. *Microsc. Res. Tech.* **74**, 257–263 (2011).
86. Cecen, A., Yabansu, Y. C. & Kalidindi, S. R. A new framework for rotationally invariant two-point spatial correlations in microstructure datasets. *Acta Mater.* **158**, 53–64 (2018).

87. Geiser, B. P., Kelly, T. F., Larson, D. J., Schneir, J. & Roberts, J. P. Spatial distribution maps for atom probe tomography. *Microsc. Microanal.* **13**, 437–447 (2007).
88. Marquis, E. A. *Microstructural Evolution and Strengthening Mechanisms in Al-Sc*, 1 edn (Materials Science and Engineering, Northwestern University, 2002).
89. Gwalani, H., Gwalani, B., O'Neill, M., Mikler, A. R. & Banerjee, R. Simulation of solute clusters in metallic systems. *Model. Simul. Mat. Sci. Eng.* **27**, 085014 (2019).
90. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proc. Second International Conference on Knowledge Discovery and Data Mining*, (eds Simoudis, E., Han, J. & Fayyad, U.) 226–231 (AAAI Press, Menlo Park, 1996).
91. Breen, A. et al. Spatial decomposition of molecular ions within 3D atom probe reconstructions. *Ultramicroscopy* **132**, 92–99 (2013).
92. Felfer, P., Ceguerra, A. V., Ringer, S. P. & Cairney, J. M. Detecting and extracting clusters in atom probe data: a simple, automated method using Voronoi cells. *Ultramicroscopy* **150**, 30–36 (2015).
93. Felfer, P., Scherrer, B., Demeulmeester, J., Vandervoort, W. & Cairney, J. M. Mapping interfacial excess in atom probe data. *Ultramicroscopy* **159**, 438–444 (2015).
94. Morozov, D. & Peterka, T. Efficient delaunay tessellation through K-D tree decomposition. In *Proc. International Conference for High Performance Computing, Networking, Storage and Analysis*, 728–738 (IEEE Press, 2016).
95. Prabhat & Koziol, Q. (eds.) *High Performance Parallel I/O*, 1 edn (Chapman & Hall, CRC Computational Science, 2014).
96. ASTM International, *ISO/ASTM52900-15 Standard Terminology for Additive Manufacturing—General Principles Terminology*. <https://www.astm.org/> (2015).
97. Zhao, H. et al. Segregation assisted grain boundary precipitation in a model Al-Zn-Mg-Cu alloy. *Acta Mater.* **156**, 318–329 (2018).
98. Robson, J. D. Modelling the overlap of nucleation, growth and coarsening during precipitation. *Acta Mater.* **52**, 4669–4676 (2004).
99. Kühbach, M., Bajaj, P., Çelik, M. H., Jäggle, E. A. & Gault, B. Supplementary material and data to “on strong-scaling and open-source tools for analyzing atom probe tomography data” on zenodo. <http://zenodo.org/record/2540529> (2020).
100. Amdahl, G. M. Validity of the single processor approach to achieving large-scale computer capabilities. In *American Federation of Information Processing Societies (ed.) AFIPS Conference Proceedings*, 483–485 (Thompson Book Co., Washington D.C., 1967).

ACKNOWLEDGEMENTS

M.K. gratefully acknowledges the funding and computing time grants through BiGmax, the Max-Planck-Society's Research Network on Big-Data-Driven Materials Science and the funding from the German Research Foundation through project RO 2342/8-1. The authors appreciate computer administration advice from Berthold Beckschäfer and Achim Kuhl. The work catalyzed from scientific discussions with Andrew Breen, Baptiste Gault, Leigh Stephenson, Jan Janßen, and Franz Roters on how to professionalize tools for APT.

AUTHOR CONTRIBUTIONS

M.K. leads the paraprobe project. He designed the tools, implemented the code, performed the analyses, and wrote the manuscript. P.B. contributed the experiments and corresponding manuscript section on additive manufacturing. H.Z. contributed the Al-Zn-Mg-Cu datasets and helped testing early versions of the tools. M.H.C. implemented a proof-of-concept Bokeh GUI for the tools, which M.K. developed further into the current GUI. E.J. and B.G. contributed through continuous scientific advice, manuscript writing suggestions, and proofreading.

FUNDING

Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

The online version contains supplementary material available at <https://doi.org/10.1038/s41524-020-00486-1>.

Correspondence and requests for materials should be addressed to M.K.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2021