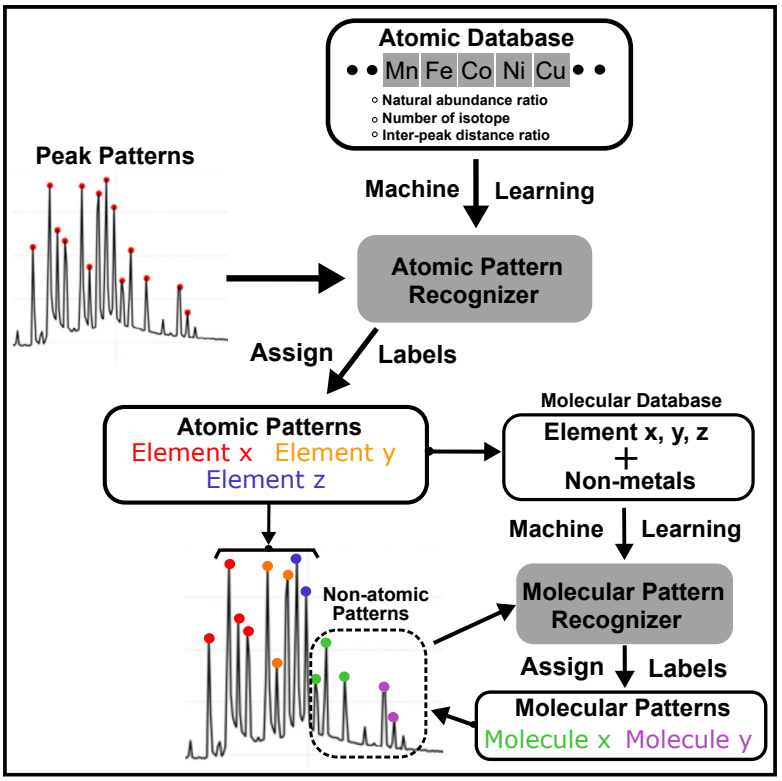
Article

Patterns

Machine-learning-enhanced time-of-flight mass spectrometry analysis

Graphical Abstract



Highlights

- A machine-learning method provides reliable atomic/molecular labels for ToF-MS
- No human labeling or prior information required
- The training dataset is artificially generated based on isotopic abundances
- Method validated on a variety of materials and two ToF-MS-based techniques

Authors

Ye Wei, Rama Srinivas Varanasi, Torsten Schwarz, ..., Hao Chen, Dierk Raabe, Baptiste Gault

Correspondence

y.wei@mpie.de (Y.W.),
b.gault@mpie.de (B.G.)

In Brief

A machine-learning application for the accelerated data processing and interpretation of time-of-flight mass spectrometry is presented. The machine learns patterns in a human-label-free manner, making the process easy to implement and the result highly reproducible.

Wei et al., 2021, Patterns 2, 100192
February 12, 2021
https://doi.org/10.1016/j.patter.2020.100192

CellPress

Patterns

CellPress
OPEN ACCESS

Article

Machine-learning-enhanced time-of-flight mass spectrometry analysis

Ye Wei,[1,5,*] Rama Srinivas Varanasi,[1] Torsten Schwarz,[1] Leonie Gomell,[1] Huan Zhao,[1] David J. Larson,[4] Binhan Sun,[1] Geng Liu,[3] Hao Chen,[3] Dierk Raabe,[1] and Baptiste Gault[1,2,*]

[1]Max-Planck-Institut für Eisenforschung, Max-Planck-Strasse 1, 40237 Düsseldorf, Germany
[2]Department of Materials, Royal School of Mines, Imperial College, London SW7 2AZ, UK
[3]Key Laboratory for Advanced Materials of Ministry of Education, School of Materials Science and Engineering, Tsinghua University, Beijing 100084, China
[4]CAMECA Instruments, 5470 Nobel Drive, Madison, WI 53711, USA
[5]Lead contact
*Correspondence: y.wei@mpie.de (Y.W.), b.gault@mpie.de (B.G.)
https://doi.org/10.1016/j.patter.2020.100192

THE BIGGER PICTURE Time-of-flight mass spectrometry (ToF-MS) is a mainstream analytical technique widely used in biology, chemistry, and materials science. ToF-MS provides quantitative compositional analysis with high sensitivity across a wide dynamic range of mass-to-charge ratios. A critical step in ToF-MS is to infer the identity of the detected ions. Here, we introduce a machine-learning-enhanced algorithm to provide a user-independent approach to performing this identification using patterns from the natural isotopic abundances of individual atomic and molecular ions, without human labeling or prior knowledge of composition. Results from several materials and techniques are compared with those obtained by field experts. Our open-source, easy-to-implement, reliable analytic method accelerates this identification process. A wide range of ToF-MS-based applications can benefit from our approach, e.g., hunting for patterns of biomarkers or for contamination on solid surfaces in high-throughput data.

Development/Pre-production: Data science output has been rolled out/validated across multiple domains/problems

SUMMARY

Mass spectrometry is a widespread approach used to work out what the constituents of a material are. Atoms and molecules are removed from the material and collected, and subsequently, a critical step is to infer their correct identities based on patterns formed in their mass-to-charge ratios and relative isotopic abundances. However, this identification step still mainly relies on individual users' expertise, making its standardization challenging, and hindering efficient data processing. Here, we introduce an approach that leverages modern machine learning technique to identify peak patterns in time-of-flight mass spectra within microseconds, outperforming human users without loss of accuracy. Our approach is cross-validated on mass spectra generated from different time-of-flight mass spectrometry (ToF-MS) techniques, offering the ToF-MS community an open-source, intelligent mass spectra analysis.

INTRODUCTION

Mass spectrometry is a widespread approach for revealing what constitutes a solution or a material. An array of techniques are used in the life sciences, geology, and materials science. Among this arsenal, time-of-flight mass spectrometry (ToF-MS) is one of the mainstream techniques in which an ion's mass-to-charge (m/z) ratio is determined via a ToF measurement.[1] It can provide a quantitative analysis of the composition of the sampled material with high precision and for a wide range of atomic and molecular masses.[2] The principles of ToF-MS are common to techniques such as matrix-assisted laser desorption/ionization (MALDI), secondary ion mass spectrometry (SIMS), or atom probe tomography (APT). Each of these techniques relies on a different concept to emit the ions from the sample, and this versatility means that their common underlying analysis approach viz.

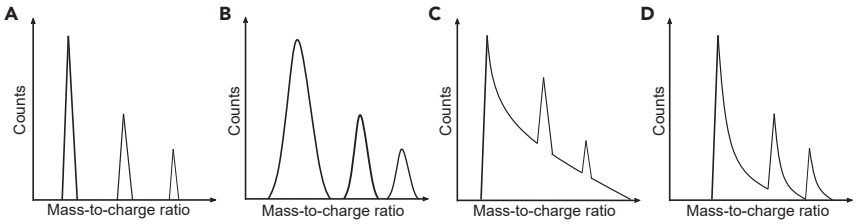


Figure 1. Examples of peak patterns under various experimental conditions

(A) Perfect peak pattern.

(B) Peak pattern with broadened peak width due to primary spatial distribution of ions.

(C) Peak patterns with long thermal tails.

(D) Peak patterns with short thermal tails.

ToF-MS has found use in chemical reaction studies, large-molecule characterization, and the quantification of dopants in semiconductors or the atomic-scale distribution of impurities at grain boundaries in metallic alloys, for instance.[3–8]

The ToF-MS data are essentially a plot of the counts as a function of the m/z ratio—typically a peak appears for each isotope of each element present—and the amplitude is proportional to the relative amount of each species within the sampled volume. Fast and accurate identification and interpretation of the rich patterns and correlations in the spectral data are of great importance and can lead to discoveries.[9] Yet the interpretation and identification rely on the user's expertise, making it slow and prone to error and hindering reproducibility.

Challenges in the development of automatic ToF-MS data analysis are two-fold. First, in ToF-MS, ions of the same species typically show distribution in their velocity or distribution in their instant of departure from the specimen. These lead to distribution in flight times. As a result, depending on the experimental conditions, ToF-MS peak patterns can take various shapes and are not always simple to recognize (Figure 1).[10] Second, molecular patterns are commonly encountered in ToF-MS spectra, i.e., not only signals from atomic ions are detected.[11–15] Combining individual atoms into a molecular ion usually leads to a new pattern comprising the distribution of the combination of isotopes from each individual element. Building a database for all possible molecular formula is practically impossible.

Machine learning (ML) is well known for its powerful ability to recognize patterns and signals.[16] Recently, the mass spectrometry community has embraced ML techniques for large-scale data analysis. The data-analyzing speed of ion-trap-based mass spectrometry has been dramatically accelerated,[17,18] whereas ToF-MS data analysis still largely relies on database searching.[19,20]

Some pioneering works demonstrated the potential of applying statistical/ML techniques to ToF-MS spectra analysis. For example, unsupervised ML has been used in exploratory data analysis for ToF-SIMS and ToF-MALDI.[21–24]

Lately, a Bayesian approach has been adopted for peak identification in APT.[25,26] The Bayesian approach implemented by A. Mikhalychev et al.[26] is able to identify and deconvolute many different types of ToF-APT mass spectra simultaneously. With reasonable prior information, this method can lead to robust results. However, prior knowledge is often provided by users. If a bad prior is assumed, the computation can become very expensive.

Here, we introduce a ML-based approach that automates the process of assigning elemental and molecular identities to peaks and series of peaks within ToF-MS spectra. Moreover, uncertainties are attached to these identities indicating to what extent the peak patterns are affected by the noise level and shape features. We name this approach "ML-ToF." It is shown that ML-ToF can handle various ToF-MS spectra without prior knowledge of composition information and from the analysis of a variety of materials systems and techniques. Indeed, we cross-validate ML-ToF on ToF-APT and ToF-SIMS spectra. The materials investigated include a high-strength Al alloy

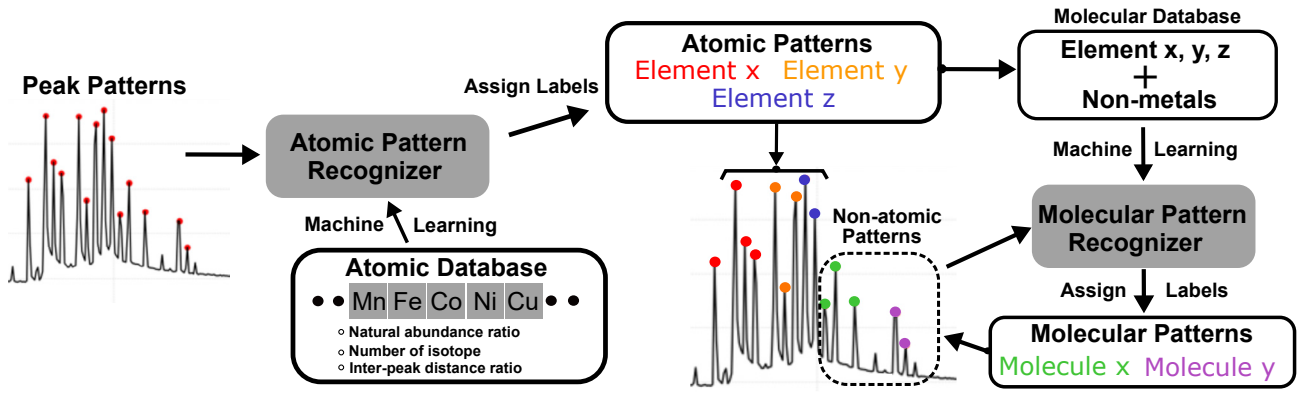


Figure 2. Flowchart of ML-assisted time-of-flight mass spectrum identification (ML-ToF)

An atomic pattern recognizer takes a mass spectrum as input and identifies all atomic patterns (mainly pure metal elements). A molecular database is then constructed by combining atomic patterns from elements with non-metal elements (e.g., hydrogen, oxygen, nitrogen). Trained in such an on-the-fly database, a machine-learning-based molecular pattern recognizer assigns molecular identities to non-atomic patterns. In such a way, ML-ToF recognizes both the elemental and the molecular fingerprints in mass spectra.

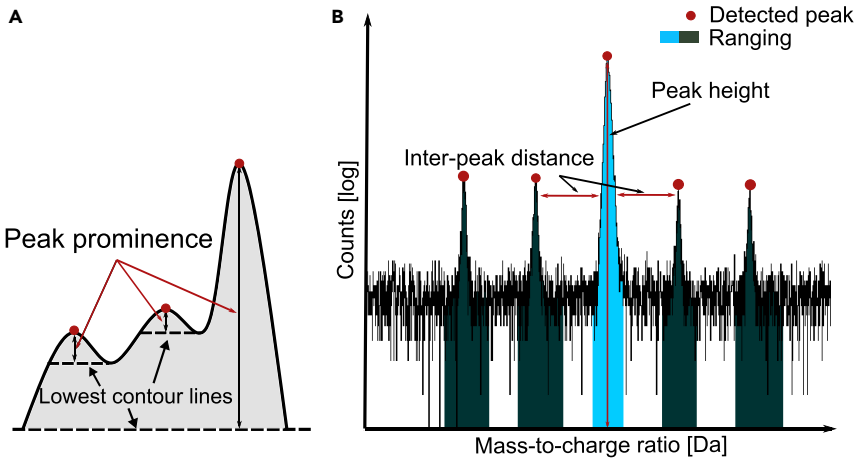


Figure 3. Examples of peak detection parameters in the SciPy Python package

(A) Schematic diagram showing the definition of peak prominence. Peak prominence is defined as the vertical distance between the peak and the lowest contour line (the dashed lines).

(B) Peak detection example from the ToF-APT dataset, showing the interpeak distance between detected peaks and peak height. Blue and dark green regions represent the range of peaks assigned by human users.

developed for aerospace applications, medium-Mn steel found in automotive applications, Cu-In-based materials used in solar cell absorbers, and SmCo-based permanent magnets. Furthermore, we benchmark the results by comparing ML-ToF-assigned labels with those yielded by field experts. ML-ToF drastically reduces the duration of the peak recognition process. In general, it takes ML-ToF microseconds to obtain a labeled spectrum, whereas human users could take minutes or even hours. An overview of our approach is shown in Figure 2.

RESULTS

Peak pattern detection

Mass spectra can be regarded as a one-dimensional array whose values are always positive. We focus on patterns with sufficient signal-to-background level to demonstrate that our approach can work properly with discernible patterns. We import the peak detection algorithm from a Python library (SciPy package, *de facto* standard package for signal processing in Python) that finds the peak positions and the corresponding intensity values.[27] The peak detection algorithm takes the mass spectra as input and searches for local maxima by a simple comparison of intensity. A subset of these peaks can be further chosen by specifying conditions of peak properties. There are three major peak properties: peak height, interpeak distance, and peak prominence. The prominence is defined as the intensity difference between the peak's height and its adjacent local minima, as indicated by Figure 3A. In Figure 3B, one can find the definition of peak height (the absolute count value in log scale). Throughout ToF-APT examples we used the same parameters for the detection (see Figure 3): peak height = 4 (log count); interpeak distance = 0.25 Da; prominence = 0.5 (log count). By visual inspection, the peak detection algorithm with this set of parameters can capture the vast majority of peaks.

In the manual procedure, users need to select a start and end position for each peak, as shown in Figure 3B. This procedure is often referred to as "ranging,"[28] and this process can lead to errors due to the different peak shapes, which depend in part on the instrument used and the experimental conditions. For instance, the laser pulse energy or the base temperature was shown to have an influence.[29–31] Here, we confine the task of ML-ToF to the identification of elemental or molecular patterns and assume the intensity represents the peak intensity at the detected position instead of the entire peak range. This assumption works well in practice: ML-ToF can recognize the peaks even when they exhibit long tails. Tails originate from either energy deficits or uncertainty on the instant at which the ion left the specimen's surface[32–35] (see Discussion). The detected m/z ratios and the corresponding intensity serve as the input of ML-ToF.

ToF-MS pattern recognition

In general terms, patterns existing in the mass spectra can be categorized into two types: (1) atomic pattern, exhibiting the natural abundance ratio of one particular element, and (2) molecular pattern, formed by two or more elements with mixed abundance ratio distribution. In this section, we introduce a systematic approach that identifies both types simultaneously. Two main

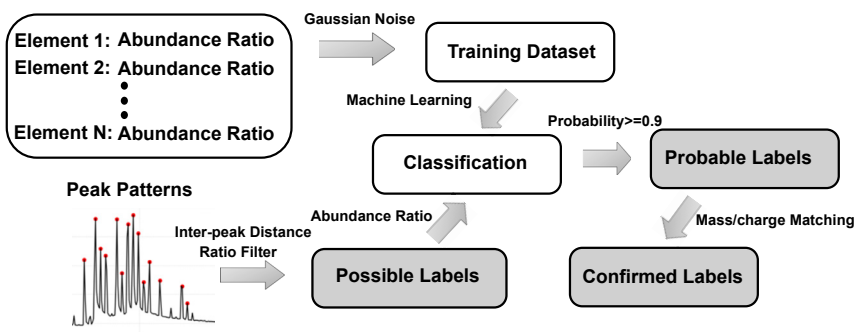


Figure 4. Protocol of atomic pattern recognizer

Patterns to be recognized are peaks with *interpeak distance ratio* and their respective abundance ratio. After the probable labels are obtained, a database search based on mass to charge is performed to identify the exact composition.

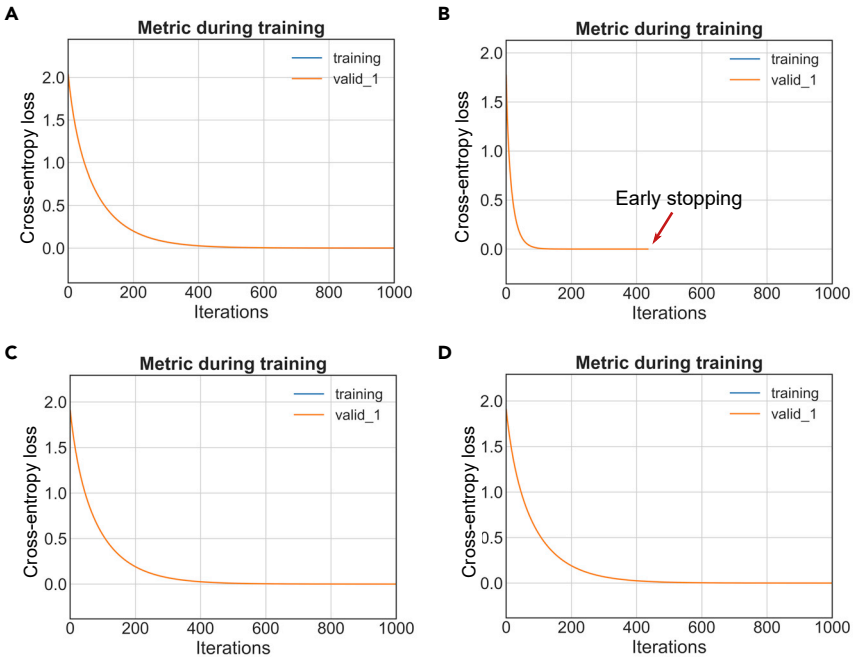


Figure 5. Training history

(A–D) Training histories of the LightGBM model for three- (A), four- (B), five- (C), and seven- (D) peak patterns are shown. In the training histories of objective function *L*, we have training and validation curves (indicated by training and valid_1, correspondingly). In all four cases, training and validating loss histories are almost the same. Hence, the two curves overlap completely.

aspects are addressed, i.e., the strategy to construct a reasonable database and the search and identification of the most probable patterns.

Atomic pattern recognizer

First, we introduce the atomic pattern recognizer designed to identify all the atomic patterns. The general protocol is demonstrated in Figure 4.

Database. ML can produce optimal results only if it is trained in a good database. In our case, the atomic pattern database consists of three parts: the number of isotope peaks, the natural abundance ratio, and the interpeak distance ratio (IDR). The IDR is defined as the distance between two neighboring peaks divided by the smallest neighboring distance within a group of peaks. For example, Fe^{+} has four peaks at 54, 56,

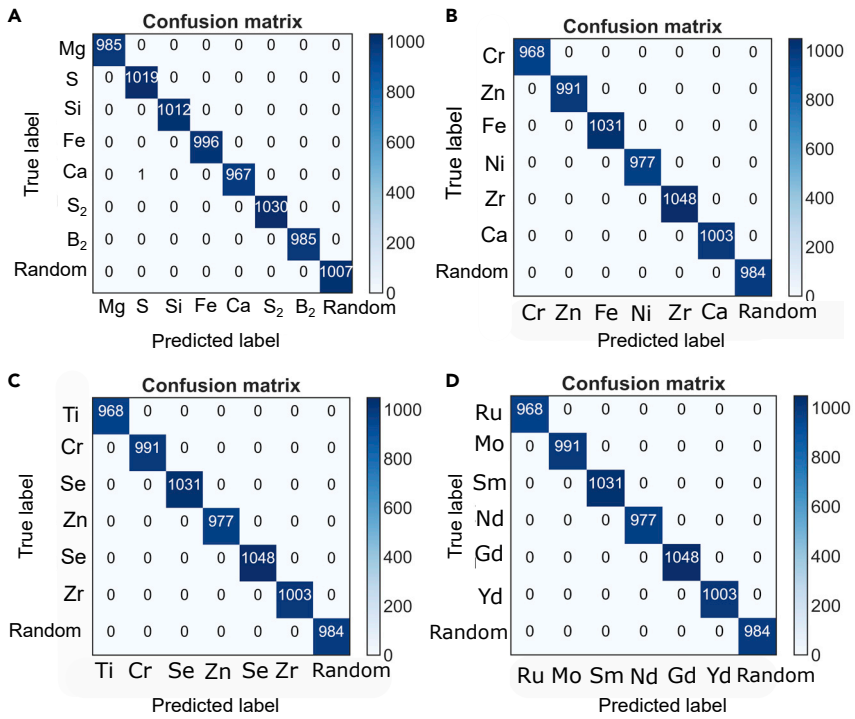


Figure 6. Confusion matrix

(A–D) Confusion matrices for three- (A), four- (B), five- (C), and seven- (D) peak patterns. The confusion matrix indicates that the models achieve 100% accuracy on the abundance ratio classification task. Small randomness is introduced in the training/testing splitting. Therefore, the size of the test dataset is not always 1,000 but quite close to it.

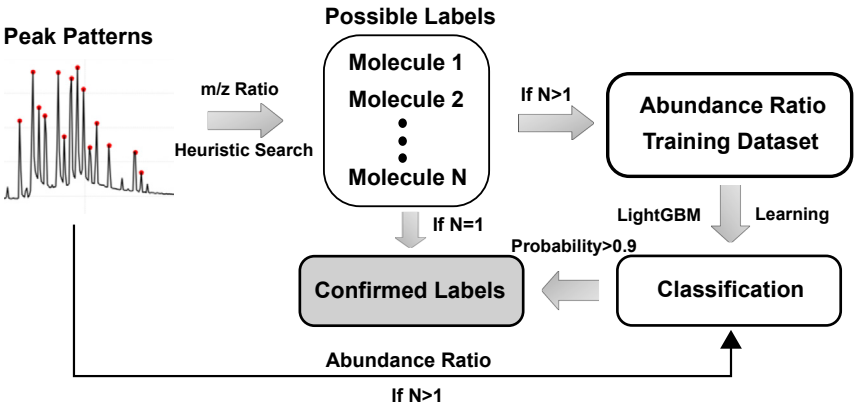


Figure 7. Molecular pattern recognizer

57, and 58 Da. So the distance ratio is (56 − 54)/(58 − 57):(57 − 56)/(58 − 57):(58 − 57)/(58 − 57) = 2:1:1. As such, even if Fe is in the form of charge state 2 with four peaks at 27, 28, 28.5, and 29 Da, the IDR is still 2:1:1. We do not have to impose any constraints on the specific charge state of the elements. This is important, as the charge-to-state ratio can vary significantly (i.e., element Fe can have 1+, 2+, or 3+ charge state) based on the experimental parameters and even within a single dataset.[36]

The database contains the most commonly encountered elements (excluding the inert gases) and some lanthanides. Currently, it contains 37 elements and 3 compounds, such as S_2 and C_2. These compounds are included because some elements have a strong tendency to form molecular ions, as frequently observed experimentally. Further information regarding the database can be found under Supplemental experimental procedure 1.1.

Interpeak distance ratio filter. As can be seen in Figure 4, matching the IDR is the first step toward a full pattern recognition. For a given peak pattern, the IDR filter searches for all possible candidates with matched IDR. Subsequently, the algorithm will examine the abundance ratio of these candidates. In practice, ToF mass spectra often contain calibration errors. Therefore, ML-ToF rounds the m/z ratio's digits up to 0, 1/4, 1/3, 2/3, 3/4, and 1 Da so that the IDR can be correctly calculated.

Learning the abundance ratio. The next step is concerned with pattern recognition of the isotopic abundance ratio. Classification of the abundance ratio is not a trivial task. Different patterns sometimes aggregate at similar m/z ratios. It is often very difficult to deconvolute them. The ML technique is naturally suited to data-driven classification tasks, thanks to its ability to learn and improve from experience without human intervention[16] automatically. Unlike the conventional yes/no answer, ML algorithms produce a list of possible answers with corresponding likelihoods. Even if an exact match from the given input to the theoretical database cannot be found, the ML-based algorithm can still provide a ranking of likely labels. In other words, ML looks for partially retained patterns and thus assigns a higher matching probability.

For elements with two isotopes, ML-ToF calculates the measured intensity ratio between the peaks ($r_m = P_1/P_2$) and compares with the expected ones from the natural abundances (r_t). If the absolute value of the deviation $(r_m - r_t)/(r_t)$ exceeded a certain threshold (here we chose empirically 0.3), then we classified this as unidentified peaks. For example, the pattern for Cu has a natural abundance ratio of 69.17:30.83, therefore the theoretical ratio r_t = 69.17/30.83 = 2.24. ML-ToF will not assign element Cu to this pattern if its abundance ratio goes outside the range [1.56, 2.91]. For monoisotopic elements (e.g., Al, As, Co), since there is no abundance ratio, ML-ToF searches for their different charge states and assigns the element if two or more of its corresponding charge states are found (e.g., Al^+ at 27 Da and Al^{2+} at 13.5 Da).

In the present study, we selected Light Gradient Boosting Machine (LightGBM) as our learning model. LightGBM belongs to the framework of Gradient Boosting Decision Tree (GBDT).[37] GBDT is an ensemble model of weaker learners that are trained in sequence. In each training iteration, a decision tree learns from the errors up to the current iteration. Via a gradient descent approach, every subsequent tree minimizes the loss function between the actual output and the weighted sum of predictions from previous iterations. The final model is the weighted average of all weak learners. GBDT has achieved state-of-the-art performance in many ML tasks, such as multiclass classification[38] and ranking tasks.[39]

Our label-predicting task is essentially a multilabel classification task. In such a setting, the algorithm tries to minimize the objective function L:

$$L = -\frac{1}{N}\left(\sum_{i=1}^{N} y_i \cdot \log(s_i)\right). \qquad \text{(Equation 1)}$$

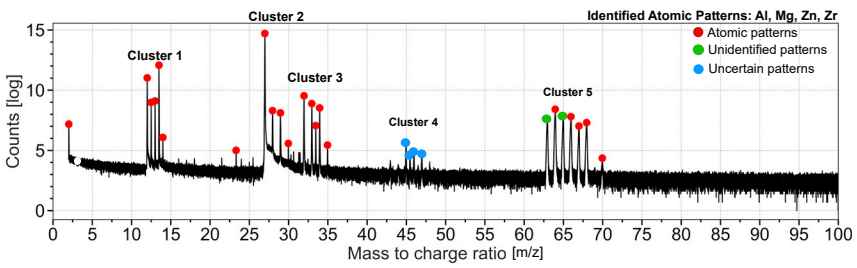


Figure 8. ML-ToF identification of a simple alloy system

Ion mass spectrum of a simple alloy system. The color of the circle markers indicates the state of the peaks. Red, green, and blue markers indicate atomic (identified), unidentified, and uncertain peaks, respectively; the majority of the ML-ToF assigned labels are consistent with APT operators.

Table 1. Peak pattern identity analysis for Al-Zn-Mg-Cu-(Zr) alloys

Cluster number	1	1	2	2	2	3	4	5	5
m/z	12 12.5 13	13.5	27	28	29	32 33 33.5 34 35	45 45.5 46 47	64 66 67 68 70	63 65
Expert	Mg^{2+}	Al^{2+}	Al^{+}	AlH^{+}	AlH_2^{+}	An^{2+}	Zr^{2+}	Zn^{+}	Cu^{+}
ML-ToF	Mg^{2+} (100%)	Al^{2+}	Al^{+}	AlH^{+}	AlH_2^{+}	Zn^{2+} (100%)	Random (51%) Zr^{2+} (45%)	Zn^{+} (100%)	None
Theory	78.99: 10.00:11.01	None	None	None	None	48.27: 27.98:4.10: 19.20:0.63	51.45: 11.22:17.15: 17.38:2.80	45.85:25.02: 12.26:16.17: 0.71	69.15:30.85
Measure	78.00: 10.13:11.87	None	None	None	None	48.63:27.73: 4.26:18.50:0.87	45.42:15.18: 21.37:18.03	48.63:27.73: 4.26:18.50:0.87	42.41:57.59

Five rows can be found for each individual cluster: mass-to-charge ratio, expert-assigned element, ML-ToF-assigned element, theoretical normalized intensity (theory), and measured normalized intensity (measurement).

L represents the cross-entropy. Here, this ML-specific entropy formulation serves as a measure for the difference between two probability distributions and is used as a loss function for classification models; N represents the number of labels, y_i is the ground truth, and s_i denotes predictions of the ML model. This objective function measures how off the machine's prediction is from the truth. The smaller the loss of objective function is, the closer the prediction of the machine is to the ground truth. Zero loss would imply that the model has achieved 100% accuracy. In general, using the cross-entropy function instead of the sum of mean square errors for a classification problem leads to a faster training as well as improved generalization.[40] In contrast to other black-box ML models like a neural network, the decision tree enjoys a unique advantage; namely, it is an explainable ML model, which provides not only the predictions but also methods to interpret them. A specific example can be found in Figure S1. Other parameters of the current LightGBM model and the corresponding explanations can be found in Supplemental experimental procedure 1.2.

We generate 5,000 data points for each element. During the training, the total dataset is further split into a first one used for the training (around 4,000 data points) and a second (around 1,000 data points) to validate the trained model. More details of database construction can be found in Supplemental experimental procedure 1.1. Figures 5A–5D illustrate the training histories of the LightGBM model for three-, four-, five-, and seven-peak patterns. The model for three-peak classification achieves near-zero loss after about 200 iterations and then plateaus at zero. Loss histories of four-peak, five-peak, and seven-peak patterns show similar trends. Notably, the four-peak pattern model converges to zero at a much faster rate, reaching near-zero loss at 100 iterations. Thus this model stops early at 500 iterations. Training and validating losses are almost identical in all four cases, resulting in two completely overlapping curves.

Confusion matrix is a useful tool for visualizing the performance of a model. It enables a direct comparison between the ML prediction and the ground truth on the test dataset. These confusion matrices (shown in Figures 6A–6D) indicate that the LightGBM models can perfectly predict the element given its abundance ratio. In addition, the training dataset introduced "redundancy" to deal with the partial pattern or overlapped pattern. For instance, three patterns are assigned to Fe: (1) atomic mass 54, 56, 57, 58 Da; abundance ratio 5.8:91.8:2.1:0.3; (2) atomic mass 54, 56, 57 Da; abundance ratio 5.8:91.8:2.1; and (3) atomic mass 56, 57, 58 Da; abundance ratio 91.8:2.1:0.3. This is because sometimes the signal-to-noise ratio of some peaks is too weak to be detected. Or a strong Ni presence (major peaks at 58 Da) destroys the first pattern of Fe. In these cases, ML-ToF is still able to recognize the presence of Fe. Such a redundancy scheme guarantees that ML-ToF has a certain degree of robustness against various noise sources.

Matching the mass-to-charge ratio. A "probable label" is defined as a peak pattern with more than 90% certainty (assigned by the LightGBM model). However, the probable label is not yet the final identified label. For example, if a pattern

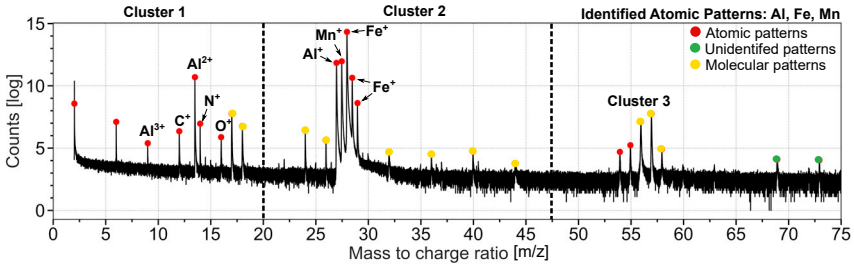


Figure 9. Identification of Fe-Mn-C-Al alloy system

Ion mass spectrum of Fe-Mn-C-Al alloy. Markers are colored based on the indicated state of the peaks (red for identified and green unidentified, yellow suggests molecular ions); dashed lines are used to separate clusters; peaks identified by the atomic pattern recognizer are indicated. ML-ToF identifies the majority of the peaks, among which atomic patterns constitute 98% intensity of the detected peaks, and about 1% are of possible molecule origins.

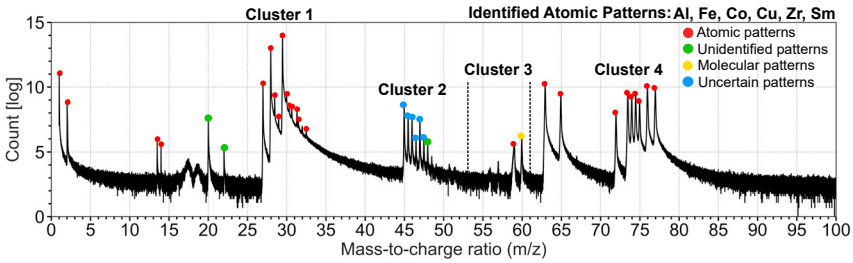


Figure 10. Identification of Sm-Co-based hard magnet

Ion mass spectrum of an Sm-Co-based hard magnet. The color of the circle markers indicates the state of the peaks.

satisfies both the IDR and the abundance ratio of element Fe, it is still possible that this pattern can be another element. Therefore, as the last step, the probable label is confirmed if its m/z ratio can be matched to a m/z ratio database, i.e., a pattern with the same IDR and abundance ratio of an element. In the case of Fe, for instance, if its m/z ratio were 54, 56, 57, 58 Da, then ML-ToF would predict Fe^+, but if its m/z ratio were 60, 72, 73, 74 Da, ML-ToF would indicate FeO^+.

Molecular pattern recognizer

When two or more elements with a different natural abundance ratio combine, the resulting molecule forms a new fingerprint. As we mentioned in the introduction, the new fingerprint differs not only in the atomic number but also in the abundance ratio. This type of combination is often found between the non-metal elements (e.g., carbon, oxygen, nitrogen, sulfur) and sometimes metallic elements too.[41] This poses a significant challenge to the database's construction, since it is impossible to search for all combinations by brute force. To identify the molecular fingerprint, we introduce a molecular pattern recognizer, which adopts a different workflow compared with the atomic pattern recognizer, as outlined in Figure 7.

For any undetermined patterns, a molecular pattern recognizer first performs a heuristic search (Figure 7) by matching their m/z ratios to an on-the-fly molecular label database and assign a molecular label to this pattern if a match is found. This on-the-fly database contains all possible recombinations between the identified atomic patterns and the non-metal elements. The range of this new molecular database depends on the maximum detected m/z ratio. If there are multiple possible candidates, an abundance-ratio-based LightGBM will be trained and will find the most probable labels. This part is similar to the atomic pattern recognizer.

DISCUSSION

Atom probe tomography

APT is a microscopy and microanalysis technique that provides the three-dimensional compositional mapping of materials at the near-atomic scale.[13,42,43] Accurate analysis of atom probe data typically involves assigning an elemental nature to each ion based on its m/z-ratio in the ToF-APT mass spectrum. In this section, we evaluate the performance of our approach on ToF-APT spectra from different alloy systems.

Aerospace high-strength Al alloy

Al-Zn-Mg-Cu-(Zr) alloys are widely employed in aerospace and automobile applications due to their low mass density and high strength.[44,45] These alloys are strengthened by a high-volume fraction of nanoscale precipitates.[46,47] ToF-APT of this alloy system generally has clear peak patterns and involves only a few molecular ions (demonstrated in Figure 8). This first example shows three possible categories for these detected peaks: identified peaks, unidentified peaks, and uncertain peaks. Overall, the patterns identified by ML-ToF are consistent with the expert's indexing, and the ML-ToF-identified peaks account for 99.9% of the total intensity of detected patterns.

The peaks are grouped into five clusters to facilitate visualization, and they are separately described in Table 1. We provide a list of tables that compare expert-assigned elements to those assigned by ML-ToF. For clusters 1, 3, 4, and 5, theoretical and measured normalized intensity (all involved normalized intensities sum up to 100) are also present. More specifically, one can observe that for clusters 1, 3, and 5, ML-ToF and expert are in complete agreement; ML-ToF assigns 100% certainty to its selected candidates (shown in parentheses after the assigned element). However, in cluster 4 (m/z ratio: 45, 45.5, 46, 47 Da), the ML algorithm is confused between a random (51%) and a Zr pattern (45%).

Table 2. Peak pattern identity analysis for Fe-Mn-C-Al alloy

Cluster number	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	3	3
m/z	2	6	9	12	13.5	14	16	17	18	24	26	27.5	27 28 28.5 29	32	36	40	44	54 56 57 58	55
Expert	H_2^+	C^{2+}	Al^{3+}	C^+	Al^{2+}	N^+	O^+	HO^+	C_3^{2+}	C_2^+	CN^+	Mn^+	Fe^{2+}	O_2^+	Fe O^{2+}	FeC_2^{2+}	None	Fe^+ and FeH^+	Mn^+
ML-ToF	H_2^+	C^{2+}	Al^{3+}	C^+	Al^{2+}	N^+	O^+	HO^+	C_3^{2+}	C_2^+	CN^+	Mn^+	Al^+ and Fe^{2+}	O_2^+	Fe O^{2+}	FeC_2^{2+} and C_2O^+ and CN_2^+	$AlOH^+$	Fe^+ and FeH^+	Mn^+

Table 3. Molecular pattern database

Molecular ion	$Fe_xH_aC_bN_cO_d$	$Al_xH_aC_bN_cO_d$	$Mn_xH_aC_bN_cO_d$	$H_aC_bN_cO_d$
Database size	329	455	222	750

x = 1, 2; a = 0, 1, 2; b = 0, 1, 2, 3, 4; c = 0, 1, 2, 3, 4; d = 0, 1, 2, 3, 4; charge state = 1, 2 and mass-to-charge ratio is restricted to below 75 Da, since no peaks are detected beyond such. The search of molecular pattern is performed within this dataset.

Two main reasons lead to this result. The first relates to the detection criteria: the fifth peak intensity is too low, such that a peak at 48 Da is not detected. The second relates to the abundance ratio: the measured abundance ratio significantly differs from Zr's natural abundance ratio. The normalized intensity of the second peak (in theory, the percentile is 11.22% but measured to be 15.18%) deviates 36% from theory. This deviation likely originates from the detection of Zr-H peaks.[48] Despite the uncertainty, ML-ToF still ranks Zr as the second most likely candidate, with 45% certainty.

Moreover, in the case of the green-colored peaks within cluster 5, ML-ToF is not able to assign any identity to peak patterns with m/z ratio values of 63 and 65 Da, while the expert would assign them as Cu^+. This is owing to the fact that ML-ToF makes predictions of two peak patterns based on a simple threshold method. In this case, the measured intensity ratio between the two peaks is 0.73. Meanwhile, if it is stand-alone element Cu, this ratio would be 2.24. Hence ML-ToF observes a remarkable deviation (67.1%) and rejects candidate Cu, contrary to an expert assignment. Cu in its 1+ charge state is also prone to be detected as CuH_2^{1+}, which will then lead to CuH_2 to overlap with the Zn peak at 67 Da, which, in part, explains the discrepancy between the measured and the theoretical ratios for Zn, which did not affect ML-ToF's capacity to identify Zn correctly. In general, when ML-ToF associates a peak with two or more atomic/molecular labels, one can apply the element deconvolution technique[49] to differentiate different labels in the same peak in terms of their spatial distribution.

Medium-Mn steel

Medium-manganese steels are promising candidates for the automotive industry owing to their excellent mechanical properties.[50] Atom probe studies help us understand the local chemistry, particularly the crystal defects, such as dislocations and grain boundaries,[51–54] thereby providing insights into the atomic-scale mechanisms at play in this class of steels. Figure 9 illustrates a mass spectrum for the more complex Fe-Mn-C-Al alloy system. More than 99% of the ions are within detected peaks assigned an identity that is consistent with that given by the field expert.

ML-ToF successfully identified the existence of the elements Fe, Mn, and Al. Non-metal elemental patterns of O, N, and C are identified too. Therefore, a new database is proposed, which contains four different types of molecular patterns: $Fe_xH_aC_bN_cO_d$, $Al_xH_aC_bN_cO_d$, $Mn_xH_aC_bN_cO_d$, and $H_aC_bN_cO_d$. The number of metals (x) is set to 1, 2, 3, 4; H (a) to 0, 1, 2; and C (b), N (c), and O (d) to 0, 1, 2, 3, 4 and charge state to 1, 2. These ranges include almost all the common types of molecular patterns. In addition, the search for molecular patterns is restricted to values below 70 Da since no peaks occur beyond this value. Combining all the above-mentioned conditions, we construct a molecular pattern database shown in Table 3.

Table 2 shows both the expert's and ML-ToF's assignment of peaks. In cluster 2, both Al^+ and Fe^+ were assigned to the peak at 27 Da, a known overlap that makes the quantification by APT of Al in Fe or Fe in Al challenging. Even in the presence of Al, the atomic pattern recognizer is still able to recognize the Fe isotope pattern with 100% certainty. At 40 Da, the algorithm offers some multiple candidates (FeC_2^{2+}, CN_2^+, C_2O^+, with the same number of atoms) compared with the expert's choice of FeC_2^{2+}. In such a case, the algorithm would also choose FeC_2^{2+} since Fe is the most abundant element (80% of intensity is assigned to element Fe).

Sm-Co-based hard magnet

Sm-Co-based materials are known for their outstanding magnetic properties related to their complex microstructure.[55,56] By changing the pinning mechanisms and pinning strength, the coercivity of the alloy $Sm_2(Co, Fe, Cu, Zr)_{17}$ can be controlled by substituting Fe for Co.[57] In this example (Figure 10), ML-ToF shows its robustness against broadened peaks due to the relatively high laser power used for this analysis.

As shown in Table 4, in cluster 1, ML-ToF identified aluminum due to the detection of peaks at Al^+ (peak at 13.5 Da) and Al^{2+} (peak at 27 Da). Also, ML-ToF identifies Zr^{3+}, albeit with reduced certainty (85%). This is likely due to the

Table 4. Peak pattern identity analysis for Sm-Co-based hard magnet

Cluster number	1	1	1	1	2	2	3	3	4
m/z	27 28 28.5 29	29.5	31.5 32.5	30 30.7 31 31.3 32	45 45.5 46 46.5 47 47.5 48	48.3 49.7 50 50.751.3	56 57 58	59 60	72 73 74 74.5 75 76 77
Expert	Fe^{2+}	Co^{2+}	Cu^{2+}	Zr^{3+}	Zr^{2+} and ZrH^{2+}	Mn^+	Fe^+	Co^+ and CoH^+	Sm^{2+}
ML-ToF	Al^+ and Fe^{2+}	CO^{2+}	Cu^{2+}	Zr^{3+} (85%)	Zr^{2+} (48.3%)	Peak not detected	Peak not detected	Co^+ and CoH^+	Sm^{2+}

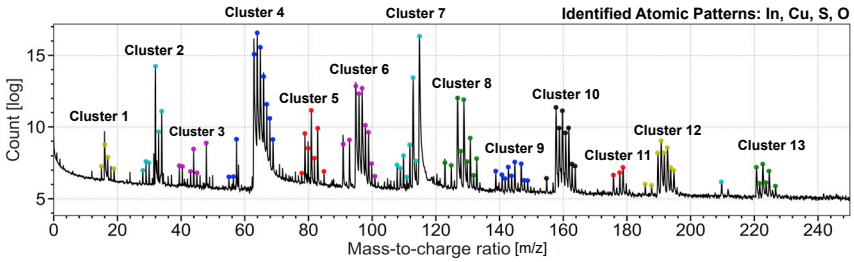


Figure 11. Identification of Cu-In-S system

Ion mass spectrum of a solar cell absorber system. Because most of the peaks are molecular pattern, for better visualization, circular markers with different colors are used to separate different clusters. The atomic pattern recognizer has identified In, Cu, S, and O as the atomic elements.

long thermal tails of the peaks. In cluster 2, ML-ToF identified Zr^{2+} with 48.3% certainty at 45, 45.5, 46, 47, and 48 Da. This relatively low probability (still considerably higher than the second-highest pattern: random [30%]) indicates the existence of other types of ions, which is pointed out by an expert as ZrH^{2+}. ML-ToF fails to assign any labels to peaks at 48.3, 49.6, 50, 50.6, and 51.3 Da. This is largely due to their relatively low signal-to-background ratio, which does not meet our detection criteria. In cluster 3, peaks at 56, 57, and 58 Da are not detected due to their low signal-to-noise ratio but still labeled by experts as Fe^{+}. Finally, at cluster 72, 73, 74, 74.5, 75, 76, and 77 Da, the element Sm is identified.

Elemental signatures like N^{+} (peak at 14 Da), As^{+} (peak at 75 Da), Sc^{+} (peak at 45 Da), and Ca^{2+} are identified too. But since we did not detect other charge states from these one/two peak elements, ML-ToF rejects these possible candidates. This can be considered as an inherent limit of the instrument itself rather than ML-ToF.

Solar cell absorber

Here, we showcase ML-ToF's application to a much more complex mass spectrum. $Cu(In,Ga)S_2$ is a compound semiconductor with a direct band gap, which can be tuned between 1.55 and 2.4 eV for pure $CuInS_2$ and $CuGaS_2$, respectively.[58] It is, therefore, suitable as an absorber material in solar cells, especially as a top junction in tandem solar cells, to overcome the Shockley-Queisser limit.[59] However, the microstructure, especially the composition-structure relationships of grain boundaries, for this material is not well known.[60,61] Here, we present for clarity only the mass spectrum of the Cu-In-S system (without Ga).

Indexing the complex mass spectrum, shown in Figure 11, is more difficult than the previous two cases. ML-ToF identifies atomic fingerprints: In, Cu, S, and O. As they tend to recombine with one another, the newly formed molecular pattern will change in terms of not only the atomic number but also the abundance ratio. Such an example is shown in Table 6. Cu and S form a compound (CuS) with atomic numbers of 95, 97, and 99, and a new abundance ratio of 63.7:32.2:1.3. Table 7 shows that the size of the new molecular database is also considerably larger than in the case of medium-Mn steel. Nevertheless, as we can see in the peak identity analysis in Table 5, ML-ToF provides a result almost identical to that of the field expert without any prior knowledge.

As can be seen from Table 5, for clusters 1, 2, 4, 7, 8, 10, and 11, ML-ToF's choice of element identity is identical to the expert's. For cluster 3, ML-ToF fails to assign any labels to peak 48 Da, whereas the expert assigns Ti^{+}. This is because the background signal is relatively higher compared with the side peaks of Ti. Therefore only one peak is detected, whereas, in theory, element Ti should show five peaks. Regarding cluster 5 (81–83 Da), the expert chose $CuOH_2^{+}$, while ML-ToF chose $CuOH^{+}$ and $CuOH_2^{+}$.

Two other interesting cases are worth mentioning. The first case is CuN_2^{+}, which is identified at (91–93 Da, cluster 6) but not confirmed by ML-ToF. A closer look reveals that this ambiguity is due to the fact that ML-ToF did not identify the pattern associated with nitrogen at 7 or 14 Da, i.e., N_2+ and N^{+}. Therefore no N-containing compounds in the new molecular pattern database involve nitrogen. In the second case, ML-ToF can predict the identity ($Cu_4S_2^{+}$ and InS^{+}) at 142–145 Da (cluster 9), while the user did not assign any identity to them.

Overall, ML-ToF has shown high fidelity in handling complicated cases, even identifying some peaks for which humans did not assign any label. More importantly, it takes ML-ToF only a half-second to complete the task. Experts would have taken 15 min on average, sometimes even longer, when scientists had no prior experience with the material system.

Secondary ion mass spectrometry

ToF-SIMS is another analytical imaging mass spectrometry technique, which provides unique insights into surface chemistry.[62–64] The large-scale and high-dimensional data generated by contemporary ToF-SIMS instruments consists of x-y-z spatial information and mass spectrum associated with each pixel. In comparison to APT, the strength of SIMS is its sensitivity associated with the larger probe volumes. The associated drawback is lower spatial resolution. A single ToF-SIMS dataset contains hundreds to thousands of mass spectra. In comparison to APT mass spectra, peak patterns of ToF-SIMS generally have a high signal-to-noise ratio. Although many peak patterns have very low intensity, these peaks are still of great importance and need to be identified. Hence the detection criteria are also different from those for ToF-APT: peak height = 0.0001 (log count); interpeak distance = 0.25 Da; prominence = 0.0001 (log count). In the following examples, we demonstrate the efficiency of ML-ToF on ToF-SIMS mass spectra of different complexities. Here we omit the tabular peak analysis and directly insert ML-ToF-assigned-labels as the expert-assigned labels are available for only a few peaks.

Table 5. Peak pattern identity analysis for Cu-In-S alloy system

Cluster number	1	2	3	4	5	6	7	8	9	10	12	13
m/z	16	32 33 34	31.5 32.5	48	57.5	63 65	64 65 66 67 68	79	80 81 82 83	91 93	95 96 97 98 99	113 115	127 128 129 130 131	142 143 144 145	158 159 160 161 162 163	190 191 192 193 194 195	221 222 223 224 225 226
Expert	O^+	S^+	Cu^{2+}	Ti^{2+}	In^{2+}	Cu^+	S^{2+}	CuO^+	$CuOH_2^+$	CuN_2^+	CuS^+ and S_3^+	In^+	CuS_2^+ and InO^+	None	CuS_3^+ and Cu_2S^+	$Cu_2S_2^+$ and CuS_4^+	Cu_3S^{2+}
ML-ToF	O^+	S^+	Cu^{2+}	None	In^{2+}	Cu^+	S^{2+}	CuO^+	$CuOH^+$ and $CuOH_2^+$		CuS^+ and S_3^+	In^+	CuS_2^+ and InO^+	Cu_4S^{2+}	CuS_3^+ and Cu_2S^+	$Cu_2S_2^+$ and CuS_4^+	Cu_3S^+ and $Cu_2S_3^+$

Table 6. An example of a new molecule pattern formation (Cu and S form CuS)

m/z ratio	63 65 (Cu)	32 33 34 (S)	95 97 99 (CuS)
Abundance ratio	69.1:30.9	95:0.8:4.2	65.7:32.2:1.3

Molecule CuS shows a new pattern.

Corrosion and wear Co-based alloy

The chemical composition (wt%) of this alloy characterized by nanoscopic SIMS is Ni 0.32, Cr 0.20, Al 0.08, and Y 0.4, balanced with Co, which is designed as a corrosion- and wear-resistant alloy employed in turbine blades.[65] The mass spectrum shown in Figure 12 was constructed by TOF-SIMS Explorer 1.3.1.0 software from the total ions information of the scanned region. In this spectrum, ToF-ML identifies Al^+, Cr^+, Co+, Ni+, Ca^+, Ti^+. This composition is relatively simple. However, abundant complex molecular fingerprints are identified by ML-ToF, as evidenced in Figure 12.

Unknown alloy from mine dump

Finally, ML-ToF was tested on an unknown alloy sample from a mine dump in Erzgebirge, Germany (Figure 13). There is no specification for nominal composition. The spectrum is produced by dynamic SIMS, showing complex peak patterns. ToF-ML identifies a variety of elements and compounds: Na^+, Al^+, Fe^+, Co^+, Cu^+, Ni^+, As^+, Mo^+, Bi^+, NaO^+, MnO^+, and CuO^+. ML-ToF is able to extract rich information even with no prior knowledge on the material.

Conclusions

We have developed a gradient-boosting-decision-tree-based approach that converts raw ToF mass spectra to its elemental or molecular identified form. The training dataset is generated based on natural abundance ratios, which does not require any human labeling. The workflow is validated on experimental datasets from APT and SIMS. Its outputs are compared with identification provided by different operators.

The main bottleneck of our approach mainly lies at the detection limits. Higher signal-to-noise ratio of the spectrum will lead to more identified patterns. Maximum peak intensity can be very sensitive to various noise sources (e.g., shot noise). To further increase the robustness of ML-ToF, one can use integral intensity as the input. A suggested criterion for such integration is to start from the maximum peak position and continue to the position whose peak intensity is 3 σ above the surrounding noise level. Sigma is the standard

Table 7. New database

Molecular ion	$Cu_xS_yO_aH_b$	$In_xS_yO_aH_b$	$S_yO_aH_b$
Database size	2,059	1,602	450

x = 1, 2, 3, 4; y = 1, 2, 3; a = 0, 1, 2, 3; b = 0, 1, 2; charge state = 1, 2, and mass-to-charge ratio is restricted to below 300 Da, since no peaks are detected beyond that. The search of molecular patterns is performed within this dataset.

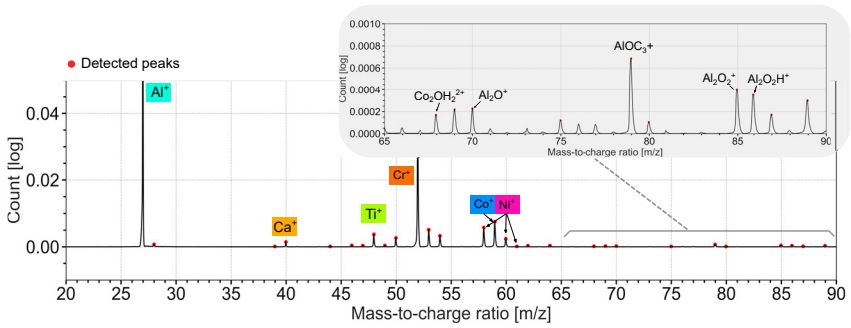


Figure 12. Identification of spectral patterns from secondary ion mass spectrometry using ML-ToF

The region of interest of mass-to-charge ratio ranges from 20 to 90 Da. ML-ToF also identifies complex molecular patterns. This can be seen in the zoom-in region (65–90 Da); note that the count [log] value is very small, because the spectrum was already normalized once by the nanoscopic SIMS software.

deviation of the surrounding noise level, assuming the noise is Gaussian distributed; 3 σ above means there is a 95% chance that the signal at this level is not noise.[66] Another limitation is that the atomic dataset does not include all elements in the periodic table, because sufficient testing and validation must be performed when new elements are added to the training data. Mass spectra containing these new elements were not typically available at the time the method was being developed. The next step is to collect more data and extend ML-ToF to more element types, thus making ML-ToF a universal technique for ToF spectral data analysis. Currently, ML-ToF still relies on brute-force search of molecular ion combinations. To accelerate this search process, one could envision a heuristic search algorithm to be integrated into the ML-ToF (e.g., beam search[67]), which rules out impossible combinations of ions.

The identification of monoisotopic species is another bottleneck of ML-TOF. Current ML-ToF could be improved using the mass defects (i.e., actual atomic masses) as an indicator for the existence of monoisotopic species, if ToF mass spectra were accurately calibrated. Finally, the implementation of real-time ML-ToF for mass spectra pattern recognition during the atom probe experiment has the potential of avoiding peak overlapping problems, thus further boosting the accuracy of APT. Finally, our method is open source, easy to implement, and capable of making instant, accurate, and consistent predictions. A wide range of ToF-based techniques can benefit from this approach, e.g., hunting for patterns of biomarkers in high-throughput ToF-MALDI data or for contamination on the solid surface in SIMS data, etc. ML-ToF enables significant acceleration of the identification process and paves the way for more reliable and more reproducible data analysis.

EXPERIMENTAL PROCEDURES

Resource availability

Lead contact

The lead contact for this article is Ye Wei, y.wei@mpie.de.

Materials availability

This study did not generate any physical material.

Data and code availability

To ensure transparency and reusability, the entire program is written in Python and is available at Github: https://github.com/DeepHeisenberg/Time-of-flight-Mass-spectra-analysis.

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.patter.2020.100192.

ACKNOWLEDGMENTS

B.G. acknowledges support from the SFB TR 270 - Hommage, project Z01. Y.W. appreciates funding by BiGmax, the Max Planck Society's Research Network on Big-Data-Driven Materials Science. The authors are grateful to Professors Leopoldo Molina-Luna and Oliver Gutfleisch for providing the Sm-Co hard magnet sample. B.G. and Y.W. acknowledge financial support from the ERC-CoG-SHINE-771602. L.G. is grateful for funding from the Stiftung des Deutschen Volkes. R.S.V. is grateful for funding from IMPRS-SURMAT. H.Z. is grateful for funding from the CSC. T.S. is grateful for funding from the German Research Foundation (DFG) (contract GA 2450/1-1) and the Luxembourgish Fonds National de la Recherche (CORRKEST). The authors are grateful to Uwe Tezins, Andreas Sturm, and Christian Broß for their support of the APT and FIB facilities at the Max-Planck-Institut für Eisenforschung GmbH.

AUTHOR CONTRIBUTIONS

Y.W. and B.G. conceived the idea and wrote the original draft; Y.W. developed the code; R.S.V., T.S., H.Z., and L.G. provided the APT dataset; D.J.L., B.S., G.L., and H.C. provided the SIMS data. All authors provided suggestions and were involved in the manuscript revision.

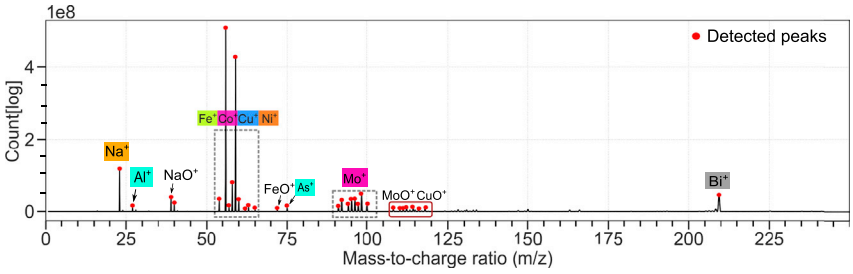


Figure 13. ML-ToF successfully assigns labels to the vast majority of peaks from mass spectra of an unknown alloy sample from a mine dump in Erzgebirge, Germany

DECLARATION OF INTERESTS

The authors declare that there is no conflict of interest.

Received: September 29, 2020
Revised: November 13, 2020
Accepted: December 17, 2020
Published: January 21, 2021

REFERENCES

1. Wolff, M.M., and Stephens, W.E. (1953). A pulsed mass spectrometer with time dispersion. Rev. Sci. Instr. *24*, 616–617, https://doi.org/10.1063/1.1770801.
2. Maher, S., Jjunju, F., and Taylor, S. (2015). Colloquium : 100 years of mass spectrometry: perspectives and future trends. Rev. Mod. Phys. *87*, 113–135, https://doi.org/10.1103/RevModPhys.87.113.
3. Sulzer, P., Petersson, F., Agarwal, B., Becker, K.H., Jürschik, S., Märk, T.D., Perry, D., Watts, P., and Mayhew, C.A. (2012). Proton transfer reaction mass spectrometry and the unambiguous real-time detection of 2,4,6 trinitrotoluene. Anal. Chem. *84*, 4161–4166, https://doi.org/10.1021/ac3004456.
4. Pedersen, S., Herek, J.L., and Zewail, A.H. (1994). The validity of the "diradical" hypothesis: direct femtoscond studies of the transition-state structures. Science *266*, 1359–1364, https://doi.org/10.1126/science.266.5189.1359. eprint. https://science.sciencemag.org/content/266/5189/1359.full.pdf.
5. Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., Yoshida, T., and Matsuo, T. (1988). Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry. Rapid Commun. Mass Spectrom. *2*, 151–153, https://doi.org/10.1002/rcm.1290020802.
6. Kissel, J., and Krueger, F.R. (1987). The organic component in dust from comet Halley as measured by the PUMA mass spectrometer on board Vega 1. Nature *326*, 755–760, https://doi.org/10.1038/326755a0.
7. Karas, M., and Hillenkamp, F. (1988). Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. Anal. Chem. *60*, 2299–2301, https://doi.org/10.1021/ac00171a028.
8. Liebscher, C.H., Stoffers, A., Alam, M., Lymperakis, L., Cojocaru-Mirédin, O., Gault, B., Neugebauer, J., Dehm, G., Scheu, C., and Raabe, D. (2018). Strain-induced asymmetric line segregation at faceted Si grain boundaries. Phys. Rev. Lett. *121*, 1, https://doi.org/10.1103/PhysRevLett.121.015702.
9. Aebersold, R., and Mann, M. (2016). Mass-spectrometric exploration of proteome structure and function. Nature *537*, 347–355, https://doi.org/10.1038/nature19949.
10. Boesl, U. (2017). Time-of-flight mass spectrometry: introduction to the basics. Mass Spectrom. Rev. *36*, 86–109. arXiv: NIHMS150003. https://doi.org/10.1002/mas.21520.
11. Tsong, T. (1984). Pulsed-laser-stimulated field ion emission from metal and semiconductor surfaces: a time-of-flight study of the formation of atomic, molecular, and cluster ions. Phys. Rev. B *30*, 4946–4961, https://doi.org/10.1103/PhysRevB.30.4946.
12. Sha, W., Chang, L., Smith, G.D.W.D.W., Mittemeijer, E.J.J., Liu, C., and Mittemeijer, E.J.J. (1992). Some aspects of atom-probe analysis of Fe-C and Fe-N systems. Surf. Sci. *266*, 416–423, https://doi.org/10.1016/0039-6028(92)91055-G.
13. Müller, M., Saxey, D., Smith, G., and Gault, B. (2011). Some aspects of the field evaporation behaviour of GaSb. Ultramicroscopy *111*, 487–492, https://doi.org/10.1016/j .ultramic. 2010. 11.019.
14. Gordon, L.M., Tran, L., and Joester, D. (2012). Atom probe tomography of apatites and bone-type mineralized tissues. ACS Nano *6*, 10667–10675, https://doi.org/10.1021/nn3049957.
15. Rusitzka, K.A.K., Stephenson, L.T., Szczepaniak, A., Gremer, L., Raabe, D., Willbold, D., and Gault, B. (2018). An atomic-scale view at the composition of amyloid-beta fibrils by atom probe tomography. Sci. Rep. *8*, 1–10, https://doi.org/10.1038/s41598- 018-36110-y.
16. Jordan, M.I., and Mitchell, T.M. (2015). Machine learning: trends, perspectives, and prospects. Science *349*, 255–260. arXiv: arXiv:1011.1669v3. https://doi.org/10.1126/science.aaa8415.
17. Elias, J.E., Gibbons, F.D., King, O.D., Roth, F.P., and Gygi, S.P. (2004). Intensity-based protein identification by machine learning from a library of tandem mass spectra. Nat. Biotechnol. *22*, 214–219, https://doi.org/10.1038/nbt930.
18. Gessulat, S., Schmidt, T., Zolg, D.P., Samaras, P., Schnatbaum, K., Zerweck, J., Knaute, T., Rechenberger, J., Delanghe, B., Huhmer, A., et al. (2019). Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. Nat. Methods *16*, 509–518, https://doi.org/10.1038/s41592-019-0426-7.
19. Sadygov, R.G., Cociorva, D., and Yates, J.R. (2004). Large-scale database searching using tandem mass spectra: looking up the answer in the back of the book. Nat. Methods *1*, 195–202, https://doi.org/10.1038/nmeth725.
20. Sinitcyn, P., Rudolph, J.D., and Cox, J. (2018). Computational methods for understanding mass spectrometry-based shotgun proteomics data. Annu. Rev. Biomed. Data Sci. *1*, 207–234, https://doi.org/10.1146/annurev-biodatasci-080917-013516.
21. Biesinger, M.C., Paepegaey, P.-Y., McIntyre, N.S., Harbottle, R.R., and Petersen, N.O. (2002). Principal component analysis of TOF-SIMS images of organic monolayers. Anal. Chem. *74*, 5711–5716, https://doi.org/10.1021/ac020311n.
22. McCombie, G., Staab, D., Stoeckli, M., and Knochenmuss, R. (2005). Spatial and spectral correlations in MALDI mass spectrometry images by clustering and multivariate analysis. Anal. Chem. *77*, 6118–6124, https://doi.org/10.1021/ac051081q.
23. Bluestein, B.M., Morrish, F., Graham, D.J., Guenthoer, J., Hockenbery, D., Porter, P.L., and Gamble, L.J. (2016). An unsupervised MVA method to compare specific regions in human breast tumor tissue samples using ToF-SIMS. Analyst *141*, 1947–1957, https://doi.org/10.1039/c5an02406d.
24. Verbeeck, N., Caprioli, R.M., and Van de Plas, R. (2020). Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. Mass Spectrom. Rev. *39*, 245–291, https://doi.org/10.1002/mas.21602.
25. Vurpillot, F., Hatzoglou, C., Radiguet, B., Da Costa, G., Delaroche, F., and Danoix, F. (2019). Enhancing element identification by expectation-maximization method in atom probe tomography. Microsc. Microanal. *25*, 367–377, https://doi.org/10.1017/S1431927619000138.
26. Mikhalychev, A., Vlasenko, S., Payne, T., Reinhard, D., and Ulyanenkov, A. (2020). Bayesian approach to automatic mass-spectrum peak identification in atom probe tomography. Ultramicroscopy *215*, 113014, https://doi.org/10.1016/j.ultramic.2020.113014.
27. Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods *17*, 261–272, https://doi.org/10.1038/s41592-019- 0686-2.
28. Hudson, D., Smith, G.D.W., and Gault, B. (2011). Optimisation of mass ranging for atom probe microanalysis and application to the corrosion processes in Zr alloys. Ultramicroscopy *111*, 480–486, https://doi.org/10.1016/j.ultramic.2010.11.007.
29. Yao, L., Cairney, J.M., Zhu, C., and Ringer, S.P. (2011). Optimisation of specimen temperature and pulse fraction in atom probe microscopy experiments on a microalloyed steel. Ultramicroscopy *111*, 648–651.
30. Tang, F., Gault, B., Ringer, S.P., and Cairney, J.M. (2010). Optimization of pulsed laser atom probe (PLAP) for the analysis of nanocomposite Ti-Si-N films. Ultramicroscopy *110*, 836–843, https://doi.org/10.1016/j.ultramic.2010.03.003.
31. La Fontaine, A., Gault, B., Breen, A., Stephenson, L., Ceguerra, A.V., Yang, L., Dinh Nguyen, T., Zhang, J., Young, D.J., and Cairney, J.M. (2015). Interpreting atom probe data from chromium oxide scales.

Ultramicroscopy *159*, 354–359, https://doi.org/10.1016/j.ultramic.2015.02.005.

32. Müller, E.W., and Krishnaswamy, S.V. (1974). Energy deficits in pulsed field evaporation and deficit compensated atom-probe designs. Rev. Sci. Instr. *45*, 10531059.
33. Vurpillot, F., Gault, B., Vella, A., Bouet, M., and Deconihout, B. (2006). Estimation of the cooling times for a metallic tip under laser illumination. Appl. Phys. Lett. *88*, 94105, https://doi.org/10.1063/1.2181654.
34. Vurpillot, F., Houard, J., Vella, A., and Deconihout, B. (2009). Thermal response of a field emitter subjected to ultra-fast laser illumination. J. Phys. D Appl. Phys. *42*, 125502, https://doi.org/10.1088/0022-3727/42/12/125502.
35. Gault, B., Moody, M.P.M., Cairney, J.M., Ringer, S.S.P., Cariney, J., and Ringer, S.S.P. (2012). Atom probe microscopy. In Springer Series in Materials Science, *Vol. 160* (Springer S). https://doi.org/10.1007/978-1-4614-3436-8.
36. Kingham, D.R. (1982). The post-ionization of field evaporated ions: a theoretical explanation of multiple charge states. Surf. Sci. *116*, 273–301, https://doi.org/10.1016/0039-6028(82) 90434-4.
37. Friedman, J.H. (2001). Greedy function approximation: a gradient boosting machine. Ann. Stat. *29*, 1189–1232, https://doi.org/10.2307/2699986.
38. Li, P. (2012). Robust LogitBoost and adaptive base class (ABC) LogitBoost, CoRR abs/1203.3491. arXiv, 1203.3491.
39. Burges, C.J.C. (2010). From rankNet to LambdaRank to lambdaMART: an overview. Learning *11*, 23–581, https://doi.org/10.1111/j.1467-8535.2010.01085.x.
40. Bishop, C.M. (2006). Pattern Recognition and Machine Learning (Information Science and Statistics) (Springer-Verlag).
41. Tsong, T.T. (1986). Observation of doubly charged diatomic cluster ions of a metal. J. Chem. Phys. *85*, 639–640.
42. Miller, M.K., Kelly, T.F., Rajan, K., and Ringer, S.P. (2012). The future of atom probe tomography. Mater. Today *15*, 158–165, https://doi.org/10.1016/S1369- 7021(12)70069-X.
43. Larson, D.J., Prosa, T.J., Ulfig, R.M., Geiser, B.P., and Kelly, T.F. (2013). Local Electrode Atom Probe Tomography (Springer Science), p. 318.
44. Starke, E., and Staley, J. (1996). Application of modern aluminum alloys to aircraft. Prog. Aerospace Sci. *32*, 131–172, https://doi.org/10.1016/0376-042K95) 00004-6.
45. Mondolfo, L. (2013). Aluminum Alloys: Structure and Properties (Elsevier).
46. Dumont, M., Lefebvre, W., Doisneau-Cottignies, B., and Deschamps, A. (2005). Characterisation of the composition and volume fraction of $\eta'$and pη'recipitates in an Al-Zn-Mg alloy by a combination of atom probe, small-angle X-ray scattering and transmission electron microscopy. Acta Mater. *53*, 2881–2892, https://doi.org/10.1016/j.actamat.2005.03.004.
47. Zhao, H., De Geuser, F., Kwiatkowski da Silva, A., Szczepaniak, A., Gault, B., Ponge, D., and Raabe, D. (2018). Segregation assisted grain boundary precipitation in a model Al-Zn-Mg-Cu alloy. Acta Mater. *156*, 318–329, https://doi.org/10.1016/j.actamat.2018.07.003.
48. Mouton, I., Breen, A.J., Wang, S., Chang, Y., Szczepaniak, A., Kontis, P., Stephenson, L.T., Raabe, D., Herbig, M., Britton, T.B., et al. (2019). Quantification challenges for atom probe tomography of hydrogen and deuterium in Zircaloy-4. Microsc. Microanal. *25*, 481–488, https://doi.org/10.1017/S143192761801615X.
49. Johnson, L., Thuvander, M., Stiller, K., Odén, M., and Hultman, L. (2013). Blind deconvolu- tion of time-of-flight mass spectra from atom probe tomography. Ultramicroscopy *132*, 60–64. IFES 2012. https://doi.org/10.1016/j.ultramic.2013.03.015.
50. Lee, Y.-K., and Han, J. (2015). Current opinion in medium manganese steel. Mater. Sci. Technol. *31*, 843–856, https://doi.org/10.1179/1743284714y.0000000722.
51. Kuzmina, M., Herbig, M., Ponge, D., Sandlobes, S., and Raabe, D. (2015). Linear complexions: confined chemical and structural states at dislocations. Science *349*, 1080–1083, https://doi.org/10.1126/science.aab2633.
52. Kuzmina, M., Ponge, D., and Raabe, D. (2015). Grain boundary segregation engineering and austenite reversion turn embrittlement into toughness: example of a 9 wt.% medium Mn steel. Acta Mater. *86*, 182–192, https://doi.org/10.1016/j.actamat.2014.12.021.
53. da Silva, A.K., Ponge, D., Peng, Z., Inden, G., Lu, Y., Breen, A., Gault, B., and Raabe, D. (2018). Phase nucleation through confined spinodal fluctuations at crystal defects evidenced in Fe-Mn alloys. Nat. Commun. *9*, https://doi.org/10.1038/s41467-018-03591-4.
54. da Silva, A.K., Kamachali, R.D., Ponge, D., Gault, B., Neugebauer, J., and Raabe, D. (2019). Thermodynamics of grain boundary segregation, interfacial spinodal and their relevance for nucleation during solid-solid phase transitions. Acta Mater. *168*, 109–120, https://doi.org/10.1016/j.actamat.2019.02.005.
55. Maury, C., Rabenberg, L., and Allibert, C.H. (1993). Genesis of the cell microstructure in the Sm(Co, Fe, Cu, Zr) permanent magnets with 2:17 type. Physica Status Solidi (a) *140*, 57–72. https://onlinelibrary.wiley.com/doi/pdf/10.1002/pssa.2211400104.
56. Gutfleisch, O., Willard, M.A., Brück, E., Chen, C.H., Sankar, S.G., and Liu, J.P. (2011). Magnetic materials and devices for the 21st century: stronger, lighter, and more energy efficient. Adv. Mater. *23*, 821–842, https://doi.org/10.1002/adma.201002180.
57. Duerrschnabel, M., Yi, M., Uestuener, K., Liesegang, M., Katter, M., Kleebe, H.J., Xu, B., Gutfleisch, O., and Molina-Luna, L. (2017). Atomic structure and domain wall pinning in samarium-cobalt-based permanent magnets. Nat. Commun. *8*, 1–7.
58. Scheer, R. (2011). Chalcogenide Photovoltaics: Physics, Technologies, and Thin Film Devices, H.S.., ed. (Wiley-VCH Verlag GmbH).
59. De Vos, A. (1980). Detailed balance limit of the efficiency of tandem solar cells. J. Phys. D Appl. Phys. *13*, 839–846, https://doi.org/10.1088/0022-3727/13/5/018.
60. Lomuscio, A., Rödel, T., Schwarz, T., Gault, B., Melchiorre, M., Raabe, D., and Siebentritt, S. (2019). Quasifermi-level splitting of Cu-poor and Cu-rich CuInS2 absorber layers. Phys. Rev. Appl. *11*, https://doi.org/10.1103/PhysRevApplied.11.054052.
61. Schwarz, T., Lomuscio, A., Siebentritt, S., and Gault, B. (2020). On the chemistry of grain boundaries in CuInS2 films. Nano Energy *76*, 105081, https://doi.org/10.1016/j.nanoen.2020.105081.
62. Liebl, H. (1967). Ion microprobe mass analyzer. J. Appl. Phys. *38*, 52775283, https://doi.org/10.1063/1.1709314.
63. Wittmaack, K. (1975). Pre-equilibrium variation of the secondary ion yield. Int. J. Mass Spectrom. Ion Phys. *17*, 39–50, https://doi.org/10.1016/0020-7381(75)80005-2.
64. Magee, C.W., Harrington, W.L., and Honig, R.E. (1978). Secondary ion quadrupole mass spectrometer for depth profiling - design and performance evaluation. Rev. Sci. Instr. *49*, 477–485, https://doi.org/10.1063/1.1135438.
65. Yang, L., Choi, R., Zheng, Y., Bidabadi, M.H.S., Rehman, A., Zhang, C., Chen, H., and Yang, Z.-G. (2020). Spalling resistance of thermally grown oxide based on NiCoCrAlY(Ti) with different oxide peg sizes. Rare Met. https://doi.org/10.1007/s12598-019-01339-7.
66. Grafarend, E. (2006). Linear and Nonlinear Models: Fixed Effects, Random Effects, and Mixed Models (de Gruyter).
67. Reddy, D.R. (1977). Speech Understanding Systems: A Summary of Results of the Five-Year Research Effort. https://apps.dtic.mil/dtic/tr/fulltext/u2/a049288.pdf.