Unsupervised Machine Learning for the Geographic Origin Determination of Cu-bearing Tourmaline

Hao A. O. Wang, Michael S. Krzemnicki, Markus Wälle and Rainer A. Schultz-Guttler

ABSTRACT: Determining the geographic origin of Cu-bearing tourmaline poses a significant challenge in gemmology, particularly when traditional microscopic methods yield inconclusive results. This study applies a combined analytical and computational approach using 469 gem-quality samples from Brazil, Mozambique and Nigeria. A total of 57 elements (from Li to U) were quantified using full-mass-spectrum LA-ICP-TOF-MS. The high-dimensional elemental dataset was reduced to interpretable 2D maps using non-linear unsupervised machine-learning algorithms, including t-distributed stochastic neighbour embedding (t-SNE) and uniform manifold approximation and projection (UMAP). These methods successfully identified complex patterns and distinct subgroups, revealing compositional similarities not captured by traditional linear approaches. The resulting clusters provided a clear framework for geographic origin determination of unknown samples. Elemental signatures of key elements (i.e. Na, Ca, Li, Ti, Fe, Mn, Cu, Ga, Sr, La and Pb) highlighted their influence on clustering and related geochemical variations to colour and geographic origin. Unsupervised machine-learning algorithms do not rely on predefined origin labels. This reduces errors caused by uncertain origin information and helps reveal statistical outliers that may point to new or undocumented sources. By integrating colour information with compositional clustering, the method also provides a possible framework for identifying heat treatment in high-clarity stones.

The Journal of Gemmology, **39**(8), 2025, pp. 772–787, https://doi.org/10.15506/JoG.2025.39.8.772

Copper-bearing tourmalines (i.e. fluor-elbaite and fluor-liddicoatite) are beautiful and fascinating members of the tourmaline family. First recognised in the late 1980s in Paraíba State, Brazil (Koivula & Kammerling 1989; Fritsch *et al*. 1990; Henn *et al*. 1990), these gems display colours distinct from those of other tourmalines, with hues that range from vivid blue—known in the trade as 'neon' or 'electric' blue—to green and purple. Major, minor and trace amounts of Cu and Mn are responsible for their intense colouration. Structurally, these tourmalines accommodate Cu substitution at the distorted octahedral Y-site, which generates characteristic Cu-related absorption bands in the 700–900 nm spectral region (Henn *et al*. 1990).

Following the earlier discoveries in Brazil, additional Cu-bearing tourmaline deposits were identified in 2001 in the Edeko area of Nigeria (Smith *et al*. 2001; Zang *et al*. 2001). Later, in 2004, discoveries in the Mavuco region of north-eastern Mozambique further expanded the known geographic and geochemical range of these rare tourmalines (Wentzell 2004; Abduriyim & Kitawaki 2005; Laurs *et al*. 2008). More recently, another alluvial deposit of Cu-bearing tourmaline was reported in the Maraca region of Mozambique, about 20 km from Mavuco (Karampelas & Klemm 2010; Milisenda & Müller 2017). As is common practice in the trade, Cu-bearing

Figure 1: These Cu-bearing tourmalines (2–20 ct) are representative of the samples analysed in this study. They are all from Brazil except for the purple specimen on the right, which originates from Mavuco, Mozambique. The individual images are scaled to similar visual size for comparison, and to illustrate the typical colour range of Cu-bearing tourmaline. Composite photo by H. A. O. Wang and Julien Xaysongkham, SSEF.

tourmalines from all these geographic origins may be called 'Paraíba tourmaline', which refers to the Brazilian locality of Paraíba where this tourmaline variety was first mined (LMHC 2023). However, specimens of 'neon' blue colour from Brazil remain the benchmark in the trade, in part for their beauty and rarity, but mostly because they link to the historical discovery of this gem.

Although the geographic origin of Cu-bearing tourmaline is a critical factor influencing its market value, this determination remains a challenge for gemmological laboratories. Due to their high clarity and diverse colours (Figure 1), traditional methods (such as spectroscopic analysis and microscopic examination of inclusions) are increasingly being complemented by advanced techniques such as laser ablation inductively coupled plasma mass spectrometry (LA-ICP-MS), which produces detailed elemental data that can reveal subtle geochemical differences helpful for distinguishing geographic origin (Abduriyim *et al*. 2006; Katsurada *et al*. 2019).

Recent advancements in machine learning (ML), a subdomain of artificial intelligence (AI; see Box A), have introduced new strategies for analysing complex datasets, improving the consistency and accuracy of gemstone origin determination. To our knowledge, the first scientific application of ML for identifying the geographic origin of gemstones was done by Dereppe *et al*. (2000), who explored artificial neural networks to classify emeralds from various origins. Since then, most studies of ML applied to gem materials have focused on techniques that employ models such as deep learning (Chow & Reyes-Aldasoro 2021; Bendinelli *et al*. 2024), artificial neural networks, random forests and support vector machines (Chow & Reyes-Aldasoro 2021; Hardman *et al*. 2024; Seneewong-Na-Ayutthaya *et al*. 2025), and partial least squares regression (Dutrow *et al*. 2024). All these *supervised* ML methods (see Box A) require large datasets confidently labelled with origin information. In gemmology, assembling such datasets is challenging because gem production is dynamic. New deposits are continuously discovered, and artisanal or small-scale mining operations typically lack the rigorous documentation required to establish a definitive chain of custody. Even in well-studied deposits, geochemical variability can blur the boundaries of origin, leaving some identifications to be based more on expert judgment than on unequivocal data (Giuliani & Groat 2019).

To address these limitations, we previously introduced and explored an *unsupervised* ML workflow for gemstone origin determination (Wang & Krzemnicki 2021; Krzemnicki *et al*. 2024). Unlike supervised methods, unsupervised ML does not require labelling of data with geographic origin before calculation (again, see Box A); instead, it identifies similarities in chemical composition. Most gems analysed in a lab were not collected *in situ*, so origin information is typically inferred from trusted sources, which may be subjective. In contrast, elemental data is obtained through objective analytical methods, such as LA-ICP-MS.

LA-ICP-MS can quantify more than 50 elements, but the resulting high-dimensional dataset (that is, with a large number of variables) is difficult for humans to interpret directly. Dimensionality-reduction techniques, such as t-distributed stochastic neighbour embedding (t-SNE; van der Maaten & Hinton 2008), allow effective projection of high-dimensional data into a lower-dimensional space without *a priori* labelled data (Wang & Krzemnicki 2021). In the present study, we expand our previous approach by incorporating an additional unsupervised ML technique—uniform manifold approximation and projection (UMAP; McInnes *et al*. 2018; Healy & McInnes 2024)—to further investigate the inherent geochemical signatures of Cu-bearing tourmaline. The goals of this approach were to detect subtle clustering patterns and to refine the classification of geographic origin.

BOX A: ARTIFICIAL INTELLIGENCE VS MACHINE LEARNING (SUPERVISED AND UNSUPERVISED APPROACHES)

Artificial intelligence (AI) refers to a broad field focused on designing and developing computer systems that can perform certain tasks which have typically required human intelligence. The foundations of AI research were laid by English mathematician Alan Turing (Turing 1950). A few years later, the term was coined by John McCarthy in a proposal for a workshop at Dartmouth College (McCarthy *et al*. 1955). AI tasks include reasoning through complex situations, understanding natural language (i.e. human language in the context of AI) and making decisions under uncertainty.

Machine learning (ML) is a specialised subfield of AI. Rather than relying on explicitly programmed instructions, ML enables computers to learn patterns and relationships directly from data. By repeatedly being exposed to examples, such as elemental compositions of gems with or without their geographic origin labels, ML models can identify statistical patterns that allow them to classify, predict or group data based on input data. For instance, given trace-element data, an ML model might learn to distinguish between gems from different geographic origins (e.g. Dereppe *et al*. 2000; Bendinelli *et al*. 2024; Seneewong-Na-Ayutthaya *et al*. 2025).

ML algorithms typically fall into three types:

1. **Supervised learning.** A model is trained on labelled examples, connecting input (e.g. elemental composition) to known output (e.g. geographic origin). The input is normally split into training, validation and testing datasets.

2. **Unsupervised learning.** A model examines unlabelled data to uncover inherent patterns, clusters or anomalies. It can group stones by elemental similarity or identify unusual specimens as outliers.

3. **Reinforcement learning.** An agent—an autonomous decision-making entity typically implemented using one or more models—learns optimal strategies by interacting with an environment and receiving rewards. Although less common in gemmology today, it holds potential for applications such as automated grading or robotic sample handling.

When choosing between supervised and unsupervised ML approaches, it is important to consider the strengths and limitations of each method. Supervised ML relies on labelled datasets, such as known geographic origins, and is highly effective when accurate, comprehensive ground-truth information is available. However, in gemmology, origin labels are often difficult to verify and may be based on subjective or incomplete information. Mislabelled data can introduce errors during model training, reducing the reliability of geographic origin determinations.

Unsupervised ML, by contrast, does not require pre-assigned labels. Instead, it identifies clusters and patterns based solely on the intrinsic structure of the data. This makes unsupervised methods preferable when such ground truth is either unavailable or unreliable. Unsupervised approaches also can serve as an exploratory step before applying supervised models. By revealing the underlying structure of the data and minimising bias from potentially unreliable labels, unsupervised methods help refine datasets and improve the output quality of subsequent ML methods.

In gemmology, AI and ML are transforming traditional workflows by augmenting, rather than replacing, the role of expert gemmologists (Wang & Krzemnicki 2021; Seneewong-Na-Ayutthaya *et al*. 2025). These tools serve as powerful assistants: rapidly processing large datasets, highlighting outliers or anomalies, and uncovering hidden patterns that may not be immediately evident. This human-machine collaborative approach enhances both efficiency and accuracy, allowing gemmologists to focus on higher-level interpretation and decision making, effectively acting as supervisors of the data-driven process.

Understanding the distinction between AI's broad, goal-oriented applications and ML's more focused, data-driven methodologies will allow gemmologists and researchers to choose the most suitable tools for specific tasks such as origin determination, quality grading and treatment detection. It will also help prevent the misuse or overstatement of terminology, such as the tendency to label basic statistical analyses as 'AI'.

MATERIALS AND METHODS

Samples

The study comprised analyses of 469 gem-quality Cu-bearing tourmalines (Table I). The samples were obtained from several reliable sources, including the SSEF reference collection, the collection of Prof. Dr Henry A. Hänni, reputable clients and mining companies, and samples with origin labelling confidently determined by gemmologists. Most of the specimens were fluor-elbaite (hereafter, simply *elbaite*), a Na-rich tourmaline with the general formula $Na(Li_{1.5}Al_{1.5})Al_6Si_6O_{18}(BO_3)_3(OH)_3(F)$, from deposits in the Paraíba and Rio Grande do Norte states of Brazil, the Mavuco region of Mozambique and the Edeko region of Nigeria. Also included were Ca-rich tourmaline samples of the fluor-liddicoatite species (hereafter, simply *liddicoatite*), with the general formula $Ca(Li_2Al)Al_6Si_6O_{18}(BO_3)_3(OH)_3(F)$, from the Maraca region of Mozambique. All samples, either faceted or rough, had at least one polished surface to minimise contamination during laser ablation inductively coupled plasma time-of-flight mass spectrometry (LA-ICP-TOF-MS) analysis.

The samples spanned a wide range of colours, mainly blue to green and purple, with various saturation levels. For simplicity, colours were categorised by main hue—blue, green and purple—using standardised illumination at 4500 K. Some of the samples used in this study had undergone heat treatment (disclosed and undisclosed), a common practice used to reduce the purple hue and enhance blue colouration in Cu-bearing tourmaline (Abduriyim *et al*. 2006). Although heat-treatment detection was not the primary focus of this study, the methodology presented also offers a promising new direction for identifying heat treatment in this type of tourmaline.

Table I: Summary of the 469 Cu-bearing tourmaline samples analysed in this study.

Geographic origin	No. samples	Weight range (ct)	Dominant hues
Brazil	253*	Melee–102.8	Blue, green, purple
Mozambique (Mavuco)	129	0.6–73.0	Blue, green, purple
Mozambique (Maraca)	72	0.9–37.8	Blue, green
Nigeria	15	0.6–61.3	Blue, green, purple

* Includes 71 samples collected during a field trip in 2017 to the São José da Batalha, Mulungu and Alto dos Quintos mines (Klumb 2018).

LA-ICP-TOF-MS Analysis

As outlined in Figure 2, analyses were conducted using a 193 nm ArF excimer laser-ablation system (NWR193UC, ESI, USA) coupled to an ICP-TOF-MS

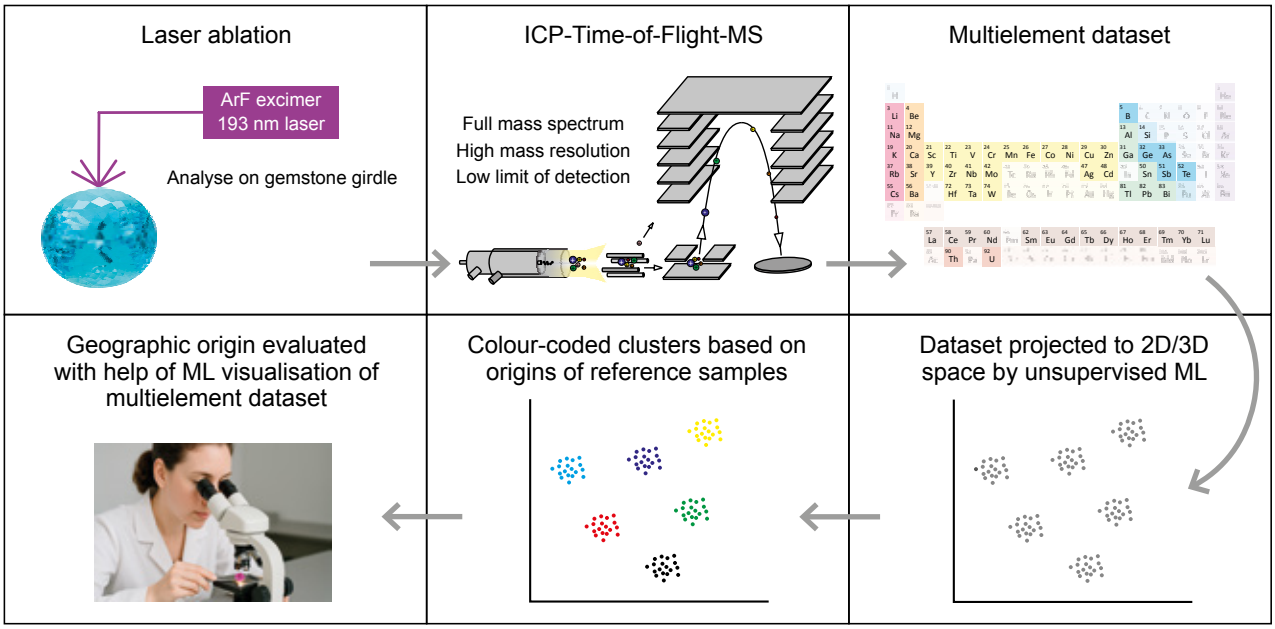

Figure 2: A schematic workflow illustrates the procedure using unsupervised ML to determine the geographic origin of an unknown sample. LA-ICP-TOF-MS measurements generate multielement data, which are projected in 2D/3D space using algorithms such as t-SNE and UMAP. The resulting low-dimensional plots are then coloured according to geographic origin or other attributes, such as elemental concentration or sample colour, to reveal compositional relationships. Finally, a gemmologist compares the composition of an unknown sample to the reference data to give an opinion on its geographic origin.

(icpTOF R, Tofwerk AG, Switzerland). The instrument was tuned daily to simultaneously measure from $^{7}Li^{+}$ to $^{238}U^{+}$, achieving a mass resolving power of approximately 3,000 at $^{238}U^{+}$. Each sample was ablated in hole-drilling mode on three to four inclusion-free spots. The laser spot was 100 µm in diameter for the samples and 75 µm for the standard reference materials (SRMs). Each position received 600 laser shots at 20 Hz and about 5.6 J/cm² fluence. Helium was used as the carrier gas to transport the aerosol into the plasma. Five pre-cleaning shots were applied at each position to remove surface contaminants.

Because no suitable matrix-matched tourmaline standard exists for the full suite of trace elements of interest, we employed a dual-SRM calibration using NIST 610 and NIST 612 glasses for external calibration. Each SRM was measured before and after the unknowns to monitor instrument drift. Quantification was performed using $^{29}Si^{+}$ as the internal standard, followed by total mass normalisation to account for matrix effects (Guillong *et al*. 2005). If the concentration of a given element fell below its detection limit, the value was replaced with a random number drawn from the log-normal distribution of detection limits for that element across all analyses. Further details of this approach are provided in Wang and Krzemnicki (2021). The dataset comprised Cu-bearing tourmaline analyses collected over a five-year period.

Unsupervised Machine Learning

Following quantification of 57 elements, to explore the complex multielement dataset and identify natural clusters and compositional similarities among samples, we used two unsupervised ML algorithms: t-SNE and UMAP. All computational analyses were conducted using Python software (version 3.11.9) on a laptop with an Intel i7 central processing unit. We then generated two-dimensional (2D) scatterplots, and geographic-origin information was colour coded and overlaid on the plots. The same approach was used to examine correlations between elemental concentrations and the dominant hues of the samples.

Because it was not possible to obtain confident provenance information for every specimen, the study included both reference samples of known origin and samples of originally unknown origin. The former—those collected in the field or provided by verified sources—served as reference anchors in the unsupervised ML embedding. Once clusters were defined based on their multielement similarities, unknown samples could be positioned relative to these reference clusters. If an unknown sample consistently plotted within a cluster defined by reference material from a specific locality, its geographic origin was then inferred.

t-SNE. A detailed explanation of the t-SNE algorithm is provided in Box B. In this study, t-SNE was implemented using the scikit-learn package, version 1.5.1 (Pedregosa *et al*. 2011), to reduce the dimensionality of the multielement dataset to two dimensions. Before analysis, the elemental concentration dataset was log-transformed to reduce skew and compress the dynamic range. The t-SNE algorithm was configured with an 'seuclidean' distance metric, a learning rate of 100, a maximum of 5,000 iterations and operation in 'exact' mode for enhanced accuracy. An exaggeration factor of 30 was applied at the beginning of the optimisation process to enhance cluster separation. Each calculation round took about three minutes.

We tested a range of perplexity values (10–200; see Box B) to optimise the balance between local and global structure preservation (Figure DD-1 in *The Journal*'s online data depository). *Local structure* refers to relationships among nearby data points—samples with similar compositions that remain close together statistically. *Global structure* describes the overall arrangement of all clusters in a dataset. Based on the selection of a perplexity of 30, the resulting 2D coordinates were visualised as scatterplots. Although 3D projections offered better subgroup separation when interactively viewed, 2D plots were chosen for publication due to their simplicity and clarity.

UMAP. As an alternative method to t-SNE, we employed UMAP (see Box C) for dimension reduction. Using the umap-learn package (version 0.5.7; McInnes *et al*. 2018), we projected the high-dimensional elemental dataset into 2D space to better visualise complex inter-sample relationships. As with t-SNE, the dataset was first log-transformed. The UMAP analysis was then configured with a Euclidean distance metric. Each calculation round took less than one minute. We explored a minimum distance (DIST) of 0.1–0.99 and a variable number (10–200) of nearest neighbours (NN) to find a suitable clustering pattern that balanced the global and local structures (see Figure DD-2 in the data depository for comparison). We chose DIST = 0.5 and NN = 30 to efficiently capture both local and global data structures, forming a robust method for subsequent cluster analyses and comparisons with t-SNE outcomes.

BOX B: t-DISTRIBUTED STOCHASTIC NEIGHBOUR EMBEDDING (t-SNE)

t-SNE is an unsupervised ML technique used to visualise complex high-dimensional datasets in two or three dimensions. In gemmology, it is particularly useful for interpreting multielement chemical data, such as the elemental fingerprints of gem materials. Each measured element in a gem represents one dimension in the dataset. With dozens of elements per sample, the resulting data exist in a high-dimensional space that can be difficult to interpret visually. The t-SNE method helps by projecting this high-dimensional structure into a lower-dimensional space, typically two dimensions, while preserving the relationships between samples as faithfully as possible (van der Maaten & Hinton 2008). Additional resources for explaining the t-SNE algorithm include Wattenberg *et al*. (2016) and Kemal (2020). Numerous dimensionality-reduction techniques exist in addition to t-SNE. Readers interested in a broader overview are encouraged to consult comparative review papers on these methods (van der Maaten *et al*. 2009; Wani 2025). The authors evaluated several methods and found that t-SNE and UMAP performed best for the Cu-bearing tourmaline dataset used in this study.

The process starts by measuring how similar each analysis (or data point in the plot) is to every other analysis based on their elemental compositions. These similarities are converted into probabilities. Next, t-SNE maps these points into two dimensions, again using probabilities, but now derived from a special statistical function called the *t-distribution*. This distribution is particularly useful because it has heavier tails compared to the standard Gaussian distribution, meaning it better handles points that are far apart, thus clearly separating different clusters. Note that the axes in t-SNE plots have no physical or compositional meaning; only the relative distances and spatial relationships between data points are interpretable.

The critical step for t-SNE is to ensure these probabilities from the simplified two-dimensional map closely reflect those from the original high-dimensional dataset. To accomplish this, t-SNE uses a mathematical measure called *Kullback-Leibler* (KL) *divergence* (Kullback & Leibler 1951). By minimising KL divergence, the algorithm ensures that points close together in the original high-dimensional space remain close in the simpler 2D map, while dissimilar points are clearly separated.

A key parameter in t-SNE is *perplexity* (Figure DD-1), which represents how many neighbours surround each data point (i.e. its statistically similar neighbours). A lower perplexity tends to focus on local structure, potentially leading to the breakup of global clusters and the appearance of many small, tightly packed groups. A higher perplexity considers more neighbours, which can provide a better representation of the global structure of the data, potentially resulting in more cohesive and broader clusters. However, very high perplexity values might blur local details.

While t-SNE excels at revealing hidden clusters and patterns, it is a non-deterministic method, meaning repeated runs may produce slightly different layouts. For more details, refer to the original description of t-SNE (van der Maaten & Hinton 2008) or visit https://scikit-learn.org/stable/modules/generated/sklearn.manifold.TSNE.html.

RESULTS AND DISCUSSION

Comparison of Different Methods of Data Analysis

To assess the effectiveness of various data-separation and dimensionality-reduction techniques for determining the geographic origin of Cu-bearing tourmaline, we applied four approaches to the same 57-element dataset: bivariate scatterplots, principal component analysis (PCA), t-SNE and UMAP. In each case, the resulting 2D scatterplots were colour coded after calculation using known geographic origins—Brazil, Mozambique (Mavuco and Maraca) and Nigeria—to independently evaluate the clustering performance of pre-assigned labels.

Bivariate Scatterplots. We began with traditional two-element scatterplots, such as Cu vs Ga (Figure 3a). Brazilian samples showed higher Cu and lower Ga concentrations, while Maraca and Mavuco samples exhibited higher Ga. Nigerian tourmalines displayed broad variability in Cu but were confined to a narrow Ga range. Despite these trends, overlap among regions was significant. Thus, such plots offer limited discrimination. Evaluating additional element pairs becomes unwieldy as the number of elements increases.

BOX C: UNIFORM MANIFOLD APPROXIMATION AND PROJECTION (UMAP)

UMAP is another unsupervised ML technique that can simplify high-dimensional data. Like t-SNE, it is designed to visualise complex multi-dimensional datasets and uncover hidden patterns. UMAP assumes that the data lie on a continuous, non-linear surface (called a manifold) and aims to reconstruct the structure of this manifold in fewer dimensions without losing important relationships. The process starts by building a graph (or network) of the high-dimensional dataset by identifying the nearest neighbours for each data point and quantifying how strongly they are connected. The resulting 'fuzzy' graph captures the local structure of the data. In the second step, UMAP optimises a low-dimensional layout that preserves these relationships as closely as possible (McInnes *et al*. 2018).

UMAP typically runs faster and handles large datasets more efficiently than t-SNE, making it better suited for extensive data analysis (McInnes *et al*. 2018). However, t-SNE generally excels at capturing fine local details, while UMAP maintains a balance between preserving local and global relationships, providing a more comprehensive view of the data's organisation. UMAP tends to be less sensitive to parameter choices than t-SNE, yielding consistent results with minimal fine tuning. For gemmological applications, a good practice is to use both methods: first use UMAP as a fast screening to quickly search for broader global layouts, and then use t-SNE to enhance the local data structures.

UMAP also includes user-defined parameters such as: *number of nearest neighbours* to control how local the analysis is (smaller values emphasise fine-grained clustering and larger values capture broader patterns); *minimum distance* to affect how tightly points are packed together in the reduced space (lower values lead to denser clusters); and *distance metric* to define how distances are measured in the high-dimensional (e.g. Euclidean) space (Figure DD-2).

In gem research, UMAP can be especially useful for revealing patterns among chemically similar samples, identifying outliers in the data and distinguishing subtle differences in multi-element fingerprints. Because it does not require origin labels, it is ideal for exploring unlabelled or partially labelled datasets, and can serve as a valuable first step before applying supervised learning or manual classification. UMAP offers a complementary approach to t-SNE to cross-validate data visualisation. For more details about UMAP, visit https://umap-learn.readthedocs.io/en.

Component Analysis. We applied PCA using MATLAB 2018b software to explore multivariate relationships within the dataset. In this approach, the Z-score standardised dataset was mathematically transformed into a new set of orthogonal axes, known as principal components. Each principal component represents a linear combination of the original variables and is ordered according to the amount of variance it explains. This transformation reduced the dimensionality of the dataset while retaining as much of the original variability as possible, allowing the dominant geochemical trends to be visualised and interpreted more effectively than the bivariate scatterplots.

As shown in Figure 3b, PCA effectively separated the Maraca (Mozambique) samples, which are liddicoatite, from the others (elbaite) due to their distinctly greater Ca, Li and rare-earth element (REE) contents. However, among the elbaite tourmalines, PCA showed substantial overlap, limiting its ability to resolve subtle geochemical differences. This behaviour reflects a common limitation of PCA when datasets are dominated by compositional gradients. The first two principal components (PC1 and PC2) are mainly driven by high-variance elements that separate liddicoatite from elbaite. Consequently, trace elements such as Cu and Ga contribute little to these components, limiting their discriminating power. To uncover these more subtle geochemical relationships, we performed a separate PCA restricted to elbaite samples (Figure DD-3 in the data depository), but overlaps between different geographic origins were still significant.

Machine-learning Methods. The unsupervised ML algorithms t-SNE and UMAP are non-linear and produced more nuanced and informative plots (Figures 3c–e). Critically, neither algorithm used geographic-origin labels during computation (Figure 3c). They were only added afterwards, ensuring an unbiased clustering result. Compared to PCA and bivariate plots, t-SNE (Figure 3d) and UMAP (Figure 3e) not only offered separation between elbaite and liddicoatite samples, but also significantly improved

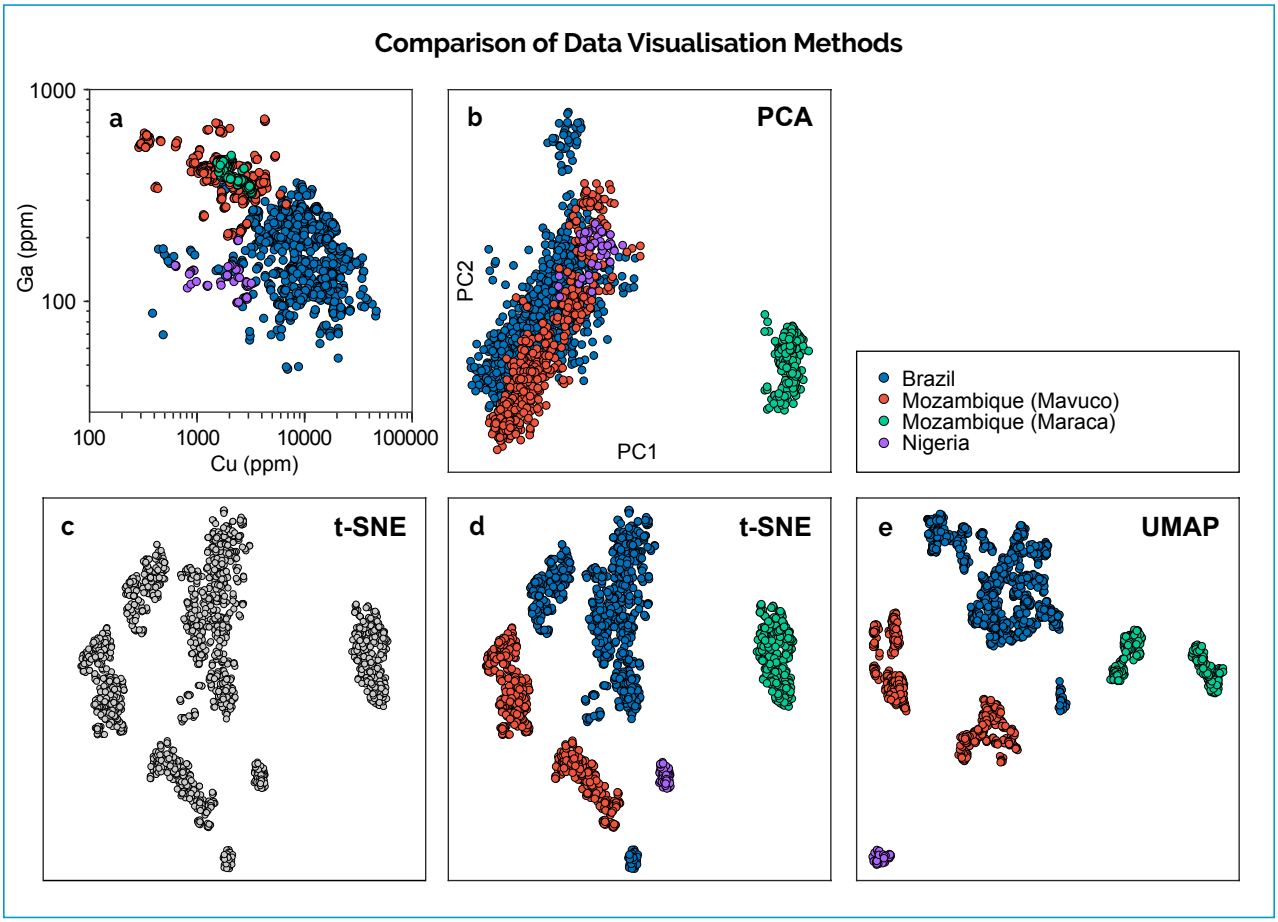

Figure 3: Four data-reduction methods were applied to the 57-element dataset of Cu-bearing tourmalines from Brazil, Mozambique (Mavuco and Maraca) and Nigeria, for comparison: (**a**) Cu vs Ga bivariate scatterplot; (**b**) plot of PCA scores; (**c**) unlabelled t-SNE plot; (**d**) t-SNE plot colour-coded by origin; and (**e**) UMAP projection colour-coded by origin. Both of the non-linear methods (d, e) provide clearer separation of geographic groups and intra-origin subclusters than the linear approaches (a, b). Note that the axes in the t-SNE and UMAP plots have no inherent meaning; only the relative distances and groupings of points carry interpretive value.

clustering among samples from Brazil, Mozambique and Nigeria.

Notably, both methods also revealed subclustering of data within Brazil and Mozambique origins (Figures 3d, e). Brazilian samples split into more than two distinct subgroups, and Mavuco and Maraca samples divided into two subgroups. While more subgroups may be revealed by tuning algorithm parameters, over-segmentation might reflect experimental variation rather than true chemical differences of geological significance.

Although UMAP appears to produce tighter and more distinct subgroups than t-SNE, this may be due to specific parameter settings of the algorithms. Therefore, we recommend a combined workflow in which both t-SNE and UMAP are applied to the multielement dataset, showing different capabilities in global and local clustering to (1) cross-validate emergent clusters, (2) identify geographic origin and intra-origin substructures, and (3) detect anomalous points that may need further investigation.

In addition, new reference samples with well-documented geographic origins should be regularly analysed to validate existing cluster labels. If newly analysed samples form separate clusters, this may suggest the presence of previously unrecognised clusters or, potentially, new deposits, either due to delayed reporting of new finds or, in rare cases, intentional concealment of source information. This dual-method approach can increase confidence in origin assignments and support the continuous development of a robust, automated, supervised ML classification system for gemmological laboratories.

Elemental Signature Visualisation

In order to highlight how specific elements contribute to the cluster separation of specific geographic origins, the relative concentrations of selected elements are represented as 'heat maps' on the t-SNE-generated scatterplot (Figure 4). (While this is shown here only for t-SNE, the same can be done with UMAP.) Each data point was coloured according to its relative

concentration for a given element, from deep blue (low concentration) to deep red (high concentration). These colour gradients were normalised independently for each element, so they reflect relative variations within an element across the dataset, but they do not allow direct comparison between different elements. The corresponding concentration ranges for these elements are summarised in Table DD-I in the data depository.

Sodium dominates the elbaite tourmaline clusters (i.e. Brazil, Nigeria and Mavuco [Mozambique]; Figure 4b), as expected from elbaite's composition (Na-Li tourmaline). The Ca and Li concentrations peak sharply in the Maraca (Mozambique) cluster, consistent with the presence of liddicoatite (Ca-Li tourmaline; Figure 4c, d). Within the Maraca cluster, Ca exhibits a clear gradient, where samples at the lower end of the plot have more Ca than at the upper end. This effect is difficult to interpret because the stones from Maraca are from an alluvial deposit, so no reliable geological context information is available (Katsurada & Sun 2017). By contrast, Na-rich elbaite samples show less origin-related differentiation based on Na content alone, indicating that other elements are needed to resolve subgroups in the elbaite population.

Figures 4e–4l show additional major, minor and trace elements (Ti, Fe, Mn, Cu, Ga, Sr, La and Pb) on a logarithmic scale, reflecting their wide dynamic range. The elements Ti (Figure 4e), Fe (Figure 4f) and Mn (Figure 4g) show pronounced concentration differences within each geographic-origin group, indicating their contribution mainly to local cluster structures (elemental variations within a geographic origin), rather than global clustering (elemental variations between different countries of origin). Copper concentrations (Figure 4h) are highest in Brazilian samples, while Ga concentrations (Figure 4i) are more elevated in Mozambique samples regardless of species, consistent with previous observations (Katsurada *et al*. 2019). Samples from Nigeria show

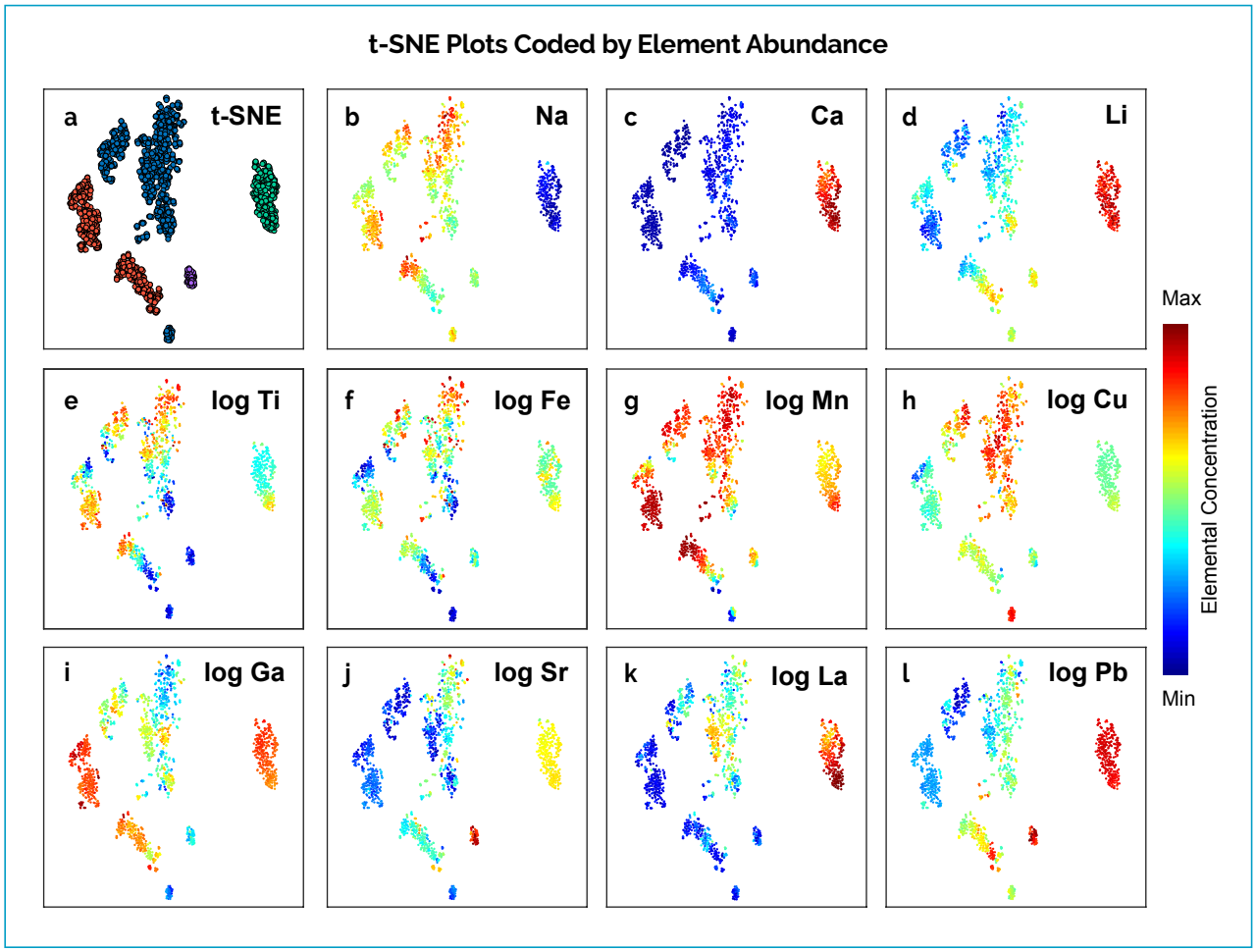

Figure 4: t-SNE plots of multielement data for Cu-bearing tourmaline are colour-coded by (**a**) geographic origin (from Figure 3d, for reference), (**b–d**) major elements (Na, Ca and Li; on a linear scale) and (**e–l**) other major, minor and trace elements (Ti, Fe, Mn, Cu, Ga, Sr, La and Pb; on a logarithmic scale). The colour gradients in plots b–l range from low (blue) to high (red) for the relative concentrations of each element. These plots illustrate how specific elemental enrichments define geographic clusters and intra-origin subclusters.

a distinctively higher Sr concentration (Figure 4j), making it a useful discriminator for origin determination. As a representative rare-earth element, La (Figure 4k) shows a higher concentration (about 10 ppm) in a small Brazilian subgroup, although it does not seem to link to any specific deposit. A higher La concentration occurs in the Maraca subgroup, which also shows a similar trend as Ca and Mn (i.e. higher concentration at the lower end of the cluster). Pb concentrations are generally higher in samples from Maraca and Nigeria (Figure 4l), while lower levels occur in samples from Mavuco and Brazil. Notably, a small subgroup within the Mavuco group shows elevated Pb, but it remains clustered with other Mavuco samples. This suggests that a dimensionality-reduction method, such as t-SNE, effectively captures overall geochemical similarity by considering the entire elemental profile, rather than individual elements in isolation, thus demonstrating the robustness of the unsupervised ML approach.

Fine Classification of Cu-Bearing Tourmaline

In the t-SNE plot (Figure 5a), the Nigerian samples form a tight, isolated cluster in the lower centre, marked by elevated Sr and Ta concentrations. The Maraca tourmaline group is defined by higher Li, Ca, Sn and REE, along with moderate Ta concentrations. Within the Maraca group, a smaller subgroup enriched in Sc and V with moderate Fe occurs towards the lower part of the cluster. Within the Brazilian domain, several subclusters emerged. One subcluster at the top of the t-SNE plot is characterised by higher Ti, Mn and Fe concentrations. Adjacent to it is a small Brazilian subgroup with moderate REE (La–Tb) concentrations, while another distinct subgroup contains lower Be and Bi (see also Figure DD-4 in the data depository). Interestingly, a similar low Be and Bi signature appears in a Mozambican subgroup, but the two subgroups remain clearly separated in the t-SNE plot due to differences in other trace elements. A small Brazilian subgroup with relatively high Cd, averaging 43 ppm (compared to an average of 2.7 ppm in the rest of the dataset), forms a discrete cluster in the bottom centre; the exact deposit of origin for these samples remains unknown. A few scattered data points near the main Brazilian subcluster may be the result of laser shots hitting inclusions, but they were labelled as Brazilian samples since they originated from reference material collected during the field trip.

The UMAP plot (Figure 5b) shows similar substructures but presents a slightly different spatial organisation due to the differing methodologies. Isolated subclusters—such as the Cd-bearing Brazilian samples, as well as those from Nigeria and Maraca—remain clearly defined. The high Ti-Mn-Fe Brazilian cluster also appears, albeit as part of a more connected substructure. Interestingly, UMAP separates the Maraca subgroup into two distinct clusters, whereas t-SNE presents them as a more continuous distribution. This difference likely arises from the different parameter sensitivities of the two algorithms, but it also suggests the potential for even finer classification within the Maraca samples (see Figure DD-5 in the data depository).

Mapping of Sample Colours

To explore the relationship between sample colour and elemental composition, we mapped the dominant colours onto the same t-SNE and UMAP plots used for elemental fingerprinting (Figure 6). Both plots reveal that samples with similar colours tend to cluster together, suggesting some correlation between colour and elemental composition. For example, the Brazilian cluster in the upper region of the t-SNE plot is predominantly green and coincides with higher concentrations of Ti, Mn and Fe. By contrast, two green-dominant subgroups in the Mavuco (Mozambique) region show only moderate Fe enrichment, but higher Mn and Ti. In addition, a small subgroup of green Maraca samples (bottom of the cluster) also coincides with elevated Ti, Mn and Fe concentrations. These observations confirm a previous finding that the green colour of Cu-bearing tourmaline is related to the presence of Ti, Mn and Fe in addition to Cu (Laurs *et al*. 2008). However, elemental analysis must be combined with spectroscopic studies to determine whether one or a combination of elements is causing the observed colour.

The purple hue of Cu-bearing tourmaline has been attributed primarily to Mn^{3+}. Various experiments have indicated that heat treatment at about 500°C or higher can reduce the purple hue, commonly resulting in a more desirable 'neon' blue colour dominated by Cu^{2+}. As such, a purple tint generally indicates lack of heat treatment. Interestingly, a comparison of Figures 4g and 6a reveals that the highest Mn concentrations occur in green, rather than purple, samples. This apparent discrepancy may be explained by differences in Mn oxidation states. Specifically, the presence of Mn^{2+} and/or Mn^{2+}–Ti^{4+} intervalence charge transfer (IVCT; Rossman & Mattson 1986), combined with Cu^{2+}, could contribute to the green colouration of these tourmalines. However, because elemental analysis by LA-ICP-TOF-MS does not distinguish oxidation

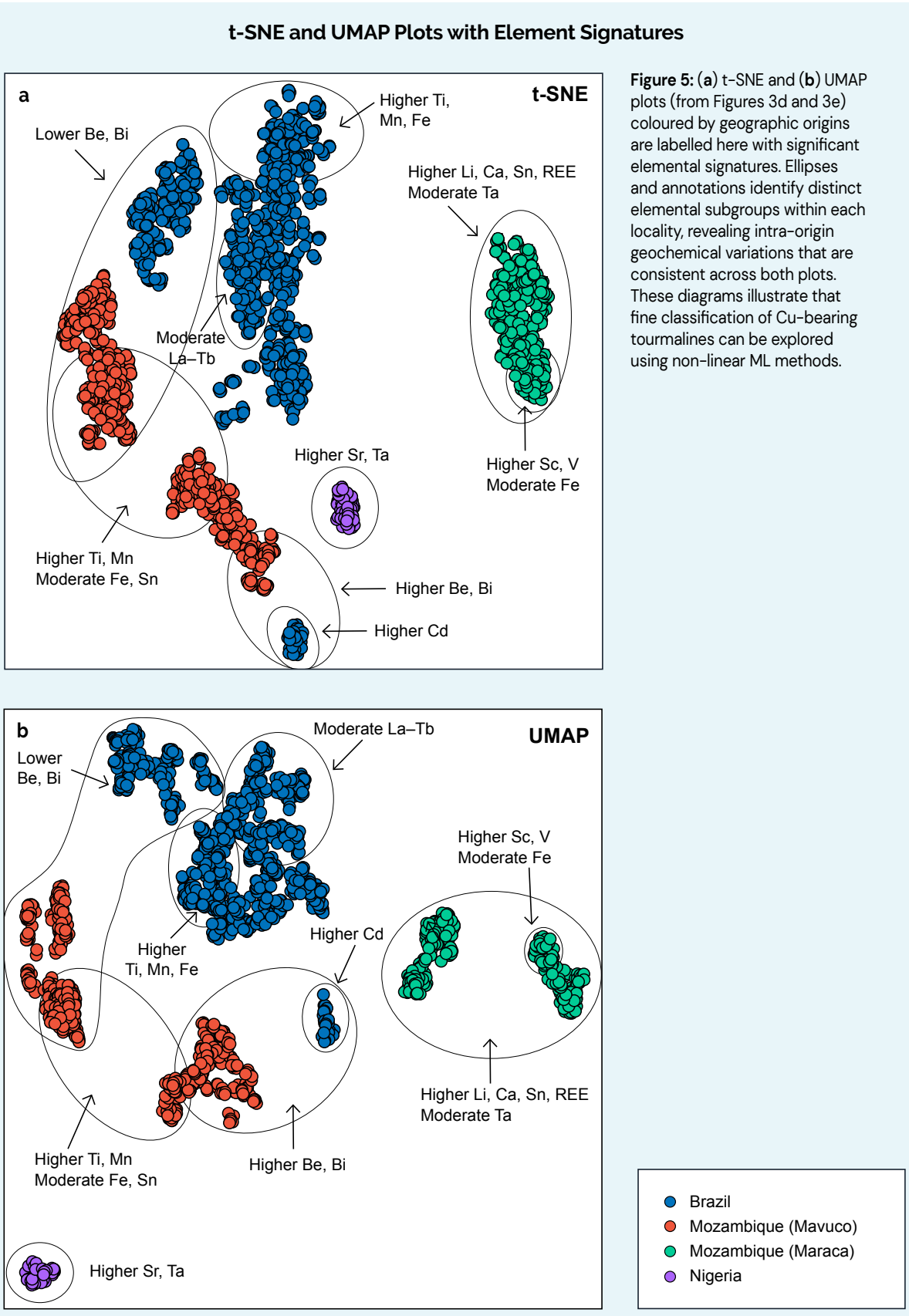

Figure 5: (**a**) t-SNE and (**b**) UMAP plots (from Figures 3d and 3e) coloured by geographic origins are labelled here with significant elemental signatures. Ellipses and annotations identify distinct elemental subgroups within each locality, revealing intra-origin geochemical variations that are consistent across both plots. These diagrams illustrate that fine classification of Cu-bearing tourmalines can be explored using non-linear ML methods.

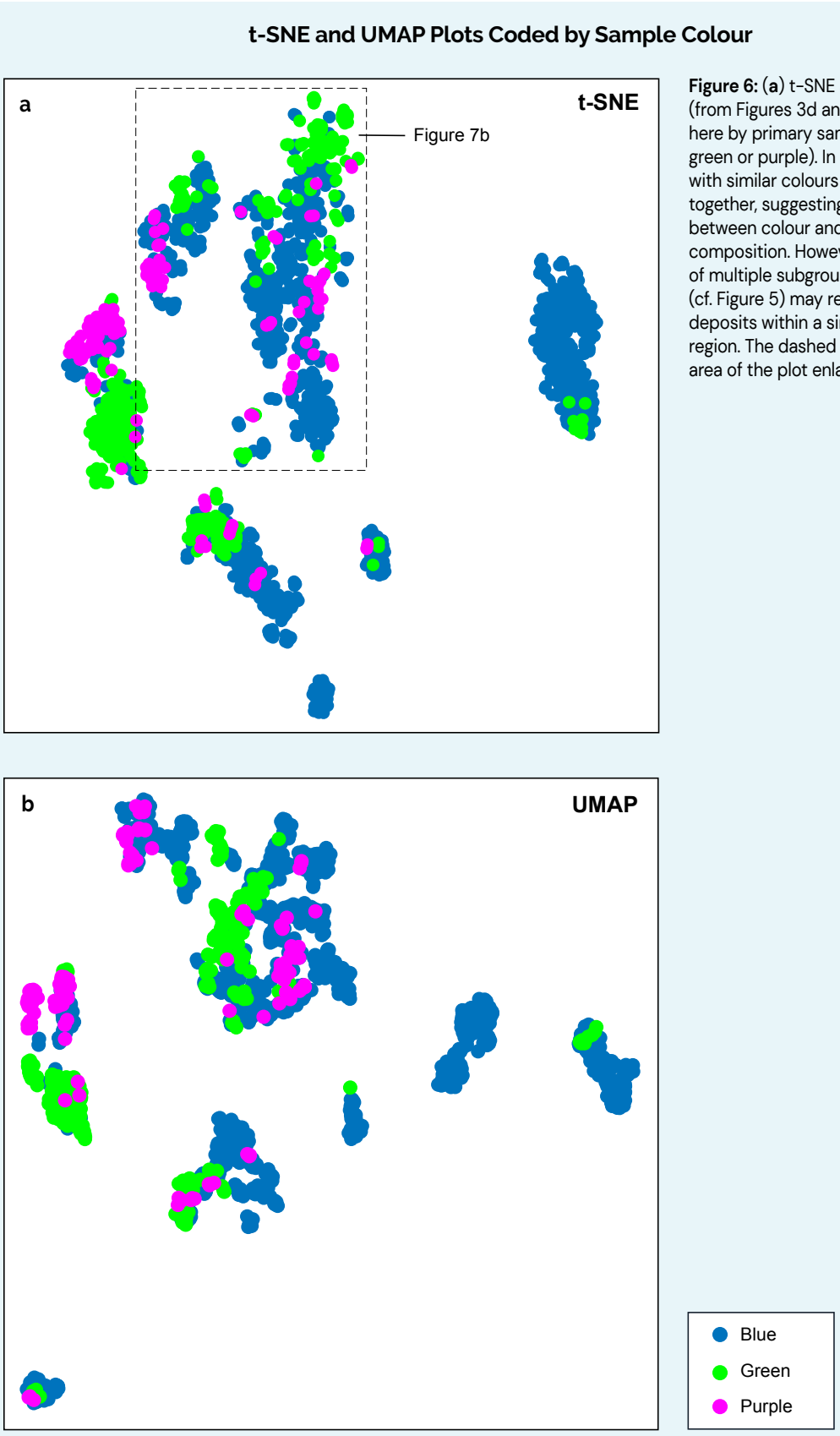

Figure 6: (**a**) t-SNE and (**b**) UMAP plots (from Figures 3d and 3e) are coded here by primary sample colours (blue, green or purple). In both plots, samples with similar colours tend to cluster together, suggesting some relationship between colour and elemental composition. However, the presence of multiple subgroups per locality (cf. Figure 5) may reflect multiple deposits within a single geographic region. The dashed box indicates the area of the plot enlarged in Figure 7b.

states, this interpretation remains speculative.

The presence of multiple ML subgroups per locality raises the question of whether these subgroups reflect multiple deposits within a single geographic region. For instance, Brazil hosts several known Cu-bearing tourmaline deposits, including São José da Batalha, Alto dos Quintos and the Mulungu mine. However, examination of a single zoned multicoloured Brazilian sample, described below, seems to suggest otherwise.

Colour and Composition Zoning in Cu-bearing Tourmaline

We performed further analyses of a multicoloured Cu-bearing tourmaline (reportedly from São José da Batalha, Brazil) from the 469-sample dataset to demonstrate how t-SNE can be used to trace compositional and colour zoning within a single crystal. The sample colours ranged from purple in the core through blue to green at the rim, and 13 laser-ablation spots were analysed across these colour zones (Figure 7a). Spot 5 intersected an inclusion and yielded a significantly higher Cs concentration (10 ppm) than the average (0.06 ppm) of other spots. Similar Cs-rich phases have been reported in pink tourmaline-bearing pegmatites in Lower Austria (Walter *et al*. 2020), so this spot was considered an outlier and omitted from further analysis.

Figure 7b shows an enlarged view of the Brazilian cluster in the t-SNE plot (dashed box in Figure 6a) with the positions of the multicoloured sample's analyses indicated. Interestingly, the purple, blue and green colour zones of the sample corresponded to three distinct subclusters of the Brazilian group in the t-SNE plot. This corresponds to the link between colour zoning and compositional variation within tourmaline. However, despite these compositional differences, all data points from this sample remained within the overall Brazilian cluster, supporting the conclusion that they originated from a single geographic origin.

This finding suggests that certain subgroups in the t-SNE plot may reflect growth zoning within single crystals or local variations within one mine rather than distinct geographic deposits. Furthermore, because this sample was unheated, as the purple core suggests, the blue and green zones are likely also unmodified. These natural variations provide a valuable reference for distinguishing unheated and heat-treated samples. Studying multicoloured tourmalines from confidently known sources is essential to understand intra-crystal zoning and refine origin assignments, and may provide a compositional framework for identifying heat treatment of Cu-bearing tourmaline.

Implications for Heat Treatment Detection in Cu-bearing Tourmaline

The analysis of the multicoloured Brazilian sample (Figure 7) may offer a promising new approach to detect heat treatment in Cu-bearing tourmaline, complementary to conventional methods based on inclusions and spectral features such as the Mn^{3+} absorption band (Laurs *et al*. 2008). Because the colours of the sample's zones are believed to be natural and are linked to distinct clusters in the t-SNE plot, these clusters can be interpreted as elemental signatures of unheated material. For example, the purple core falls within a specific cluster in the t-SNE map, but some of the other samples in that cluster display a blue colour instead of

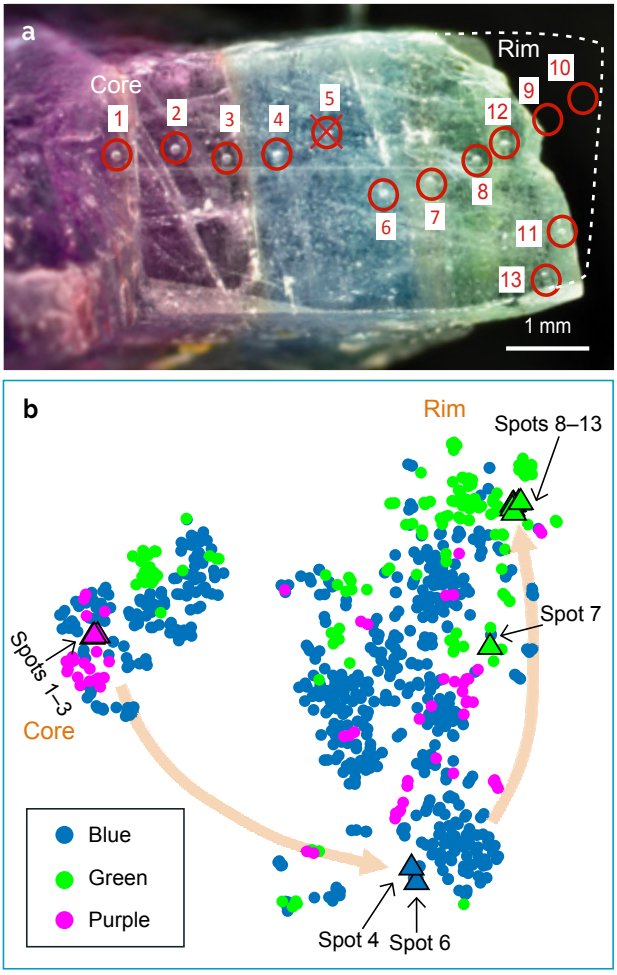

Figure 7: (**a**) An unheated multicoloured Cu-bearing tourmaline sample from Brazil (reportedly from São José da Batalha) shows distinct colour zoning, from purple in the core through blue to green at the rim. The 13 analytical spots are marked. Spots 9 and 10 were measured on a fragment that detached during sample handling (dashed outline in the photo). (**b**) An enlarged view of the Brazilian cluster (dashed box in Figure 6a) shows the relative positions of the 13 analysed spots (triangle symbols) within the t-SNE plot. Spot 5 likely captured a Cs-bearing inclusion, so it was omitted from data analysis. A comparison of Figures 6 and 7 shows that the observed subgroups in Figures 3–6 do not necessarily correspond to different deposits.

purple. This raises the possibility that those specimens may have undergone heat treatment to reduce purple and enhance blue colouration. Similarly, the lower cluster in the t-SNE plot includes the blue zone of the multicoloured sample. Other samples in this cluster are also blue, which suggests that this cluster may represent unheated blue tourmalines. These associations between composition and colour provide a new framework for flagging potential heat-treatment candidates.

This approach is especially valuable because many Cu-bearing tourmalines are of high clarity and lack inclusions that could provide evidence of heat treatment. More work is needed to validate this method, combining unsupervised ML with confidently unheated reference samples to strengthen the identification of treated stones. In addition, this framework may complement established geochemical indicators. For example, Okrusch *et al*. (2016) proposed that blue Cu-bearing tourmalines with $CuO/MnO_{tot}<0.5$ may have undergone heat treatment to reduce the reddish component associated with Mn^{3+}. Integrating such chemical ratios with ML-derived clustering and colour zoning could enhance the reliability of treatment detection.

Figure 8: An exceptional Paraíba tourmaline (about 5 ct) from Brazil, set in a ring from the Asta Collection, Hong Kong, illustrates gem-quality tourmaline for which origin determination is important. Photo by SSEF.

CONCLUSIONS

This study demonstrates that combining full-spectrum elemental analysis via LA-ICP-TOF-MS with unsupervised ML techniques offers a powerful and objective approach to determining the geographic origin of Cu-bearing tourmaline (e.g. Figure 8). Unsupervised ML techniques such as t-SNE and UMAP are more effective than traditional PCA in resolving complex relationships within high-dimensional multielement datasets.

Overall, mapping individual element distributions by t-SNE provided a powerful visual tool for interpreting the geochemical drivers of sample clustering. The element-specific overlays demonstrated that the t-SNE clustering was not arbitrary; rather, it reflected geochemical similarities across major, minor and trace elements. By linking elevated elemental zones with geographic clusters, we gained intuitive origin-specific elemental fingerprints and an interpretable framework for origin determination. Moreover, this technique may provide insight into the geochemical evolution and formation conditions of tourmaline deposits.

By combining t-SNE and UMAP (Figure 5), robust non-linear ML could resolve not only major country-of-origin groups but also intra-origin subclusters defined by subtle elemental patterns. These subclusters may reflect variations in geological conditions during crystal growth (e.g. different mining sites within a larger deposit area), and are of particular interest for ongoing mineralogical research. For example, the presence of relatively high-Cd Brazilian samples suggests unique formation conditions that require further investigation.

The ability of unsupervised ML algorithms to reveal natural groupings without relying on pre-assigned origin labels makes them particularly valuable in gemmology, where origin information is frequently uncertain and new sources continue to emerge. Moreover, such unsupervised approaches provide a crucial preliminary step before applying downstream ML or AI methods, helping to refine the dataset and augment the gemmologist's decision-making for geographic origin determination.

Beyond origin determination, unsupervised ML potentially provides insight into a stone's colour zoning and growth history, paving the way for more comprehensive gemmological investigations. In particular, the integration of known unheated reference samples into the workflows may support the development of a new method for detecting heat treatment in high-purity Cu-bearing tourmaline.

Overall, this study highlights the transformative potential of unsupervised ML in gem research and testing. It lays the foundation for more accurate, data-driven classification systems that can adapt to the evolving complexity of the global gem market.

REFERENCES

Abduriyim, A. & Kitawaki, H. 2005. Gem News International: Cu- and Mn-bearing tourmaline: More production from Mozambique. *Gems & Gemology*, **41**(4), 360–361.

Abduriyim, A., Kitawaki, H., Furuya, M. & Schwarz, D. 2006. "Paraíba"-type copper-bearing tourmaline from Brazil, Nigeria, and Mozambique: Chemical fingerprinting by LA-ICP-MS. *Gems & Gemology*, **42**(1), 4–21, https://doi.org/10.5741/gems.42.1.4.

Bendinelli, T., Biggio, L., Nyfeler, D., Ghosh, A., Tollan, P., Kirschmann, M.A. & Fink, O. 2024. Gemtelligence: Accelerating gemstone classification with deep learning. *Communications Engineering*, **3**(1), article 110, https://doi.org/10.1038/s44172-024-00252-x.

Chow, B. & Reyes-Aldasoro, C. 2021. Automatic gemstone classification using computer vision. *Minerals*, **12**(1), article 60, https://doi.org/10.3390/min12010060.

Dereppe, J.M., Moreaux, C., Chauvaux, B. & Schwarz, D. 2000. Classification of emeralds by artificial neural networks. *Journal of Gemmology*, **27**(2), 93–105, https://doi.org/10.15506/JoG.2000.27.2.93.

Dutrow, B.L., McMillan, N.J. & Henry, D.J. 2024. A multivariate statistical approach for mineral geographic provenance determination using laser-induced breakdown spectroscopy and electron microprobe chemical data: A case study of copper-bearing tourmalines. *American Mineralogist*, **109**(6), 1085–1095, https://doi.org/10.2138/am-2023-9164.

Fritsch, E., Shigley, J.E., Rossman, G.R., Mercer, M.E., Muhlmeister, S.M. & Moon, M. 1990. Gem-quality cuprian-elbaite tourmalines from São José da Batalha, Paraíba, Brazil. *Gems & Gemology*, **26**(3), 189–205, https://doi.org/10.5741/gems.26.3.189.

Giuliani, G. & Groat, L.A. 2019. Geology of corundum and emerald gem deposits: A review. *Gems & Gemology*, **55**(4), 464–489, https://doi.org/10.5741/gems.55.4.464.

Guillong, M., Hametner, K., Reusser, E., Wilson, S.A. & Günther, D. 2005. Preliminary characterisation of new glass reference materials (GSA-1G, GSC-1G, GSD-1G and GSE-1G) by laser ablation-inductively coupled plasma-mass spectrometry using 193 nm, 213 nm and 266 nm wavelengths. *Geostandards and Geoanalytical Research*, **29**(3), 315–331, https://doi.org/10.1111/j.1751-908X.2005.tb00903.x.

Hardman, M.F., Homkrajae, A., Eaton-Magaña, S., Breeding, C.M., Palke, A.C. & Sun, Z. 2024. Classification of gem materials using machine learning. *Gems & Gemology*, **60**(3), 306–329, https://doi.org/10.5741/gems.60.3.306.

Healy, J. & McInnes, L. 2024. Uniform manifold approximation and projection. *Nature Reviews Methods Primers*, **4**(1), article 82, https://doi.org/10.1038/s43586-024-00363-x.

Henn, U., Bank, H., Bank, F.H., Platen, H.V. & Hofmeister, W. 1990. Transparent bright blue Cu-bearing tourmalines from Paraiba, Brazil. *Mineralogical Magazine*, **54**(377), 553–557, https://doi.org/10.1180/minmag.1990.054.377.04.

Karampelas, S. & Klemm, L. 2010. Gem News International: "Neon" blue-to-green Cu- and Mn-bearing liddicoatite tourmaline. *Gems & Gemology*, **46**(4), 323–325.

Katsurada, Y. & Sun, Z. 2017. Cuprian liddicoatite tourmaline. *Gems & Gemology*, **53**(1), 34–41, https://doi.org/10.5741/gems.53.1.34.

Katsurada, Y., Sun, Z., Breeding, C.M. & Dutrow, B.L. 2019. Geographic origin determination of Paraíba tourmaline. *Gems & Gemology*, **55**(4), 648–659, https://doi.org/10.5741/gems.55.4.648.

Kemal, E. 2020. t-SNE clearly explained. Medium.com, https://medium.com/data-science/t-sne-clearly-explained-d84c537f53a, accessed 23 November 2025.

Klumb, A. 2018. Field trip to Paraiba, Brazil. *Facette Magazine*, No. 24, 8–9, https://ssef.ch/wp-content/uploads/2018/03/2018_SSEF_Facette_24.pdf.

Koivula, J.I. & Kammerling, R.C. 1989. Gem News: Paraíba tourmaline update. *Gems & Gemology*, **25**(4), 248–249.

Krzemnicki, M.S., Wang, H.A.O., Wälle, M., Lefèvre, P., Zhou, W., & Cartier, L.E. 2024. Gemmological characterisation of emeralds from Musakashi, Zambia, and implications for their geographic origin determination. *Journal of Gemmology*, **39**(4), 338–350, https://doi.org/10.15506/JoG.2024.39.4.338.

Kullback, S. & Leibler, R.A. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*. **22**(1), 79–86, https://doi.org/10.1214/aoms/1177729694.

Laurs, B.M., Zwaan, J.C., Breeding, C.M., Simmons, W.B., Beaton, D., Rijsdijk, K.F., Befi, R. & Falster, A.U. 2008. Copper-bearing (Paraíba-type) tourmaline from Mozambique. *Gems & Gemology*, **44**(1), 4–30, https://doi.org/10.5741/gems.44.1.4.

LMHC 2023. *LMHC Information Sheet # 6: Paraiba Tourmaline*. Laboratory Manual Harmonisation Committee, 2 pp., https://www.lmhc-gemmology.org/wp-content/uploads/2023/06/LMHC-Information-Sheet_6_V8_2023.pdf.

McCarthy, J., Minsky, M.L., Rochester, N. & Shannon, C.E. 1955. *A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence*. Dartmouth College, Hanover, New Hampshire, USA, 23 pp., https://raysolomonoff.com/dartmouth/boxa/dart564props.pdf.

McInnes, L., Healy, J. & Melville, J. 2018. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv*, arXiv:1802.03426v3, https://doi.org/10.48550/arXiv.1802.03426.

Milisenda, C.C. & Müller, S. 2017. REE photoluminescence in Paraiba type tourmaline from Mozambique. *35th International Gemmological Conference*, Windhoek, Namibia, 8–19 October, 71–73, https://www.igc-gemmology.org/wp-content/uploads/2023/12/IGC2017-web.pdf.

Okrusch, M., Ertl, A., Schüssler, U., Tillmanns, E., Brätz, H. & Bank, H. 2016. Major- and trace-element composition of Paraíba-type tourmaline from Brazil, Mozambique and Nigeria. *Journal of Gemmology*, **35**(2), 120–139, https://doi.org/10.15506/JoG.2016.35.2.120.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V. & Thirion, B. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, **12**, 2825–2830, https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.

Rossman, G.R. & Mattson, S.M. 1986. Yellow, Mn-rich elbaite with Mn-Ti intervalence charge transfer. *American Mineralogist*, **71**(3–4), 599–602.

Seneewong-Na-Ayutthaya, M., Sripoonjan, T., Wanthanachaisaeng, B., Leelawatanasuk, T., Lhuaamporn, T. & Suwanmanee, W. 2025. Machine-learning applications in gemmology: Classifying the geographic origin of ruby and sapphire using chemical data. *Journal of Gemmology*, **39**(7), 634–655, https://doi.org/10.15506/JoG.2025.39.7.634.

Smith, C.P., Bosshart, G. & Schwarz, D. 2001. Gem News International: Nigeria as a new source of copper-manganese-bearing tourmaline. *Gems & Gemology*, **37**(3), 239–240.

Turing, A.M. 1950. I.—Computing machinery and intelligence. *Mind*, **LIX**(236), 433–460, https://doi.org/10.1093/mind/LIX.236.433.

van der Maaten, L. & Hinton, G. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, **9**, 2579–2605.

van der Maaten, L., Postma, E. & van den Herik, J. 2009. Dimensionality Reduction: A Comparative Review. TiCC TR 2009–005, Tilburg University, Tilburg, The Netherlands, 35 pp., https://lvdmaaten.github.io/publications/papers/TR_Dimensionality_Reduction_Review_2009.pdf.

Walter, F., Auer, C., Bernhard, F., Bojar, H.-P., Brandstätter, F., Grill, J.A., Kiseljak, R., Knobloch, G. *et al*. 2020. Neue Mineralfunde aus Österreich LXIX. *Carinthia II*, **210**(130), 153–218, https://www.zobodat.at/pdf/CAR_210_130_0153-0218.pdf.

Wang, H.A.O. & Krzemnicki, M.S. 2021. Multi-element analysis of minerals using laser ablation inductively coupled plasma time of flight mass spectrometry and geochemical data visualization using t-distributed stochastic neighbor embedding: Case study on emeralds. *Journal of Analytical Atomic Spectrometry*, **36**(3), 518–527, https://doi.org/10.1039/d0ja00484g.

Wani, A.A. 2025. Comprehensive review of dimensionality reduction algorithms: Challenges, limitations, and innovative solutions. *PeerJ Computer Science*, **11**, article e3025, https://doi.org/10.7717/peerj-cs.3025.

Wattenberg, M., Viégas, F. & Johnson, I. 2016. How to use t-SNE effectively. *Distill*, https://doi.org/10.23915/distill.00002, 13 October, accessed 23 November 2025.

Wentzell, C.Y. 2004. Lab Notes: Copper-bearing color-change tourmaline from Mozambique. *Gems & Gemology*, **40**(3), 250–251.

Zang, J.W., da Fonseca-Zang, W.A., Fliss, F., Höfer, H.E. & Lahaya, Y. 2001. Cu-haltige Elbaite aus Nigeria. *Berichte Der Deutschen Mineralogischen Gesellschaft*, **13–14**, 202.

The Authors

Dr Hao A. O. Wang FGA[1], Dr Michael S. Krzemnicki FGA[1,2], Dr Markus Wälle[1] and Prof. Dr Rainer A. Schultz-Guttler[3]

[1] Swiss Gemmological Institute SSEF, Aeschengraben 26, 4051 Basel, Switzerland
Email: hao.wang@ssef.ch

[2] Department of Environmental Sciences, University of Basel, Bernoullistrasse 32, 4056 Basel, Switzerland

[3] Institute of Geosciences, University of São Paulo, CEP 05508 080, São Paulo, Brazil

Acknowledgements

We sincerely thank our colleagues at the Swiss Gemmological Institute SSEF for valuable discussions and technical support throughout the study. We acknowledge the Analytics team for assisting with part of the LA-ICP-TOF-MS measurements, and the Gemmology team for their careful work in geographic provenance determination. We are also grateful to Carlo Somma (Brazil), Sebastian Ferreira (Brazil Paraíba Mine, Brazil) and Nelson Oliveira (Brazil) for hosting the SSEF field trip in 2017 and for generously donating reference samples. Special thanks are extended to Peter Lyckberg (Luxembourg), Paul Wild Co. (Germany), Wild & Petsch Co. (Germany), Ekkehard Schneider (EFS Gems, Germany), Joshua Nassi (Mya Nassi Inc., USA), Pedro Oselieri (Oselieri-Racine SA, Switzerland), Azizi Hashmat (Azizi Enterprises, Thailand) and Horst Munch (Brazil) for kindly donating or providing well-documented samples. Their contributions were essential for establishing a robust and reliable dataset for this work. Prof. Leander Franz and Sarah Degen (University of Basel) are acknowledged for fruitful discussions and assistance. The new perspective on heat-treatment detection presented in this work arose from answering a question posed by Martin Julier (Bucherer AG, Switzerland).

The draft of this article was polished by ChatGPT. The authors reviewed, edited and revised the ChatGPT-edited text, and take ultimate responsibility for the content of this publication.