Journal of Archaeological Method and Theory (2025) 32:28
https://doi.org/10.1007/s10816-025-09695-8

RESEARCH

"Open Sourcing" Workflow and Machine Learning Approaches for Attributing Obsidian Artifacts to Their Volcanic Origins: A Feasibility Study from the South Caucasus

Pavol Hnila[1,2] · Ellery Frahm[3,4] · Alessandra Gilibert[5] · Arsen Bobokhyan[6]

Accepted: 8 January 2025

Abstract

Traditionally, reliable obsidian sourcing requires expensive calibration standards and extensive geological reference collections as well as experience with statistical processing. In the South Caucasus — one of the most obsidian-rich regions on the planet — this combination of requirements has often restricted sourcing studies because few projects have geological reference collections that cover all known obsidian sources. To test an alternative approach, we conducted "open sourcing" using portable X-ray fluorescence (pXRF) analyses of geological specimens with three key changes to the conventional method: (1) commercially available calibration standards were replaced with a loanable Peabody-Yale Reference Obsidians (PYRO) set, (2) a comprehensive geological reference collection was replaced with a published dataset of consensus values (Frahm, 2023a, 2023b), and (3) processing in statistical packages was replaced with two semiautomated machine-learning workflows available online. For comparison, we used classification by-eye with JMP 17.2 statistical software. Furthermore, we propose a new method to evaluate calibrations, which streamlines comparisons and which we refer to as a symmetric difference ratio (SDR). The results of this feasibility study demonstrate that this "open sourcing" workflow is reliable, yet currently only in combination with classification by-eye. When the consensus values were combined with the machine-learning solutions, the classification results were unsatisfactory. The most encouraging aspect of our alternative "open sourcing" workflow is that it enables correct source identification without physically measuring reference collections, therefore surmounting an obstacle that, until now, has severely limited archaeological research. We anticipate that rapid developments in machine-learning will also soon improve the workflow.

Extended author information available on the last page of the article

Published online: 27 January 2025

Keywords pXRF analysis · Calibration assessment · Trendline comparisons · Peabody-Yale Reference Obsidians (PYRO) · SourceXplorer · AutoML for geochemistry

Abbreviations

LDA	Linear discriminant analysis
FUB	Freie Universität Berlin
PCA	Principal component analysis
PYRO	Peabody-Yale Reference Obsidians, a complete set consists of two parts: a calibration set and a check set
SDR	Symmetric difference ratio (see the "Symmetric difference ratio (SDR)" section for definition)
ML	Machine learning
Algorithmic classification	Classification by statistical prediction algorithms (usually LDA, PCA)
Classification by-eye	Attribution of unknown specimens to known geochemical groups based on visual inspection of 2D and/or 3D scatterplots (*i.e.*, overlaps, proximity/distance)
"knowns"	Dataset with specimens of known geological origin, used to train a machine-learning algorithm or to define standards to which the dataset(s) with unknown values should be compared
"unknowns"	Dataset with specimens of yet-to-be-determined geological sources

Introduction

For more than 60 years, archaeologists have used various means of geochemical characterization (primarily elemental analysis) to match obsidian artifacts to the geological origins of the volcanic glass (Cann & Renfrew, 1964). Commonly known as "obsidian sourcing" or "obsidian provenancing," this process was first successfully developed in the Near East and Aegean regions, where it revealed unexpectedly complex connections among early settlements via the exchange of obsidian. Since then, obsidian artifact sourcing has been applied to archaeological studies on every inhabited continent and throughout human history, ranging from the African Earlier Stone Age (Mussi et al., 2023) to nineteenth-century California (Silliman, 2005). For much of the past six decades, however, obsidian sourcing has largely remained the domain of relatively few specialists with the requisite resources and expertise. Specifically, conducting reliable volcanic source identifications of obsidian artifacts has traditionally required (1) specialized laboratories with highly precise analytical instruments, (2) expensive reference standards to calibrate those instruments and to assure reproducibility and inter-laboratory comparability of the resulting elemental data, (3) extensive geological reference collections for geochemical comparisons to artifacts, and (4) skill with statistical techniques to interpret the results and to

attribute artifacts to geological sources. Without each of these, it was not possible for archaeologists to conduct obsidian sourcing on their own.

In recent years, however, these four requirements have each seen noteworthy developments that allow expanded access to the geochemical means of obsidian sourcing, especially in the region where this method was introduced. First, the rise and proliferation of portable X-ray fluorescence (pXRF) instruments since the early 2000s has brought elemental analysis by X-ray spectrometry within reach of many archaeological projects (Frahm & Doonan, 2013; Kuzmin et al., 2020). Second, the creation of the Peabody-Yale Reference Obsidians (PYRO) calibration sets (Frahm, 2019) — loanable reference materials designed for obsidian analysis — reduced or even removed the need to purchase expensive calibration standards. Third, a database of consensus elemental values for obsidian sources from the Aegean to the South Caucasus (Frahm, 2023a, 2023b) minimized the requirement for a comprehensive reference collection of obsidian specimens from every possible geological source. Fourth, the growth of online statistical tools for geochemical classifications — such as SourceXplorer (McMillan et al., 2022) and AutoML for Geochemistry (Alferéz et al., 2022) — offered an alternative to the mastery of advanced statistical packages. Therefore, many hurdles that had prevented the practice of obsidian sourcing by non-specialists have since been either reduced or removed.

Inspired by the transformative potential of these developments, we sought to combine them into a new "open sourcing" and low-budget workflow for obsidian source identifications in the challenging setting of the South Caucasus, one of the most obsidian-rich regions on the planet. In practice, this means that we propose (1) using a pXRF instrument and calibrating its measurements with a loanable PYRO set, (2) comparing the geochemical results of measured artifacts with the published dataset of accurate consensus values, and (3) attributing the artifacts to sources with either statistical software or machine-learning tools (but, as we show, the latter will be a step for the future). Although proposing this new workflow is the primary aim of the paper at hand, novel assessment methods were developed as a secondary goal. To evaluate the results of a calibration, we propose a new method that combines the slope and intersect of the calibration equations into a single entity, which we refer to as a symmetric difference ratio (SDR; see the "Symmetric difference ratio (SDR)" section). After reliability of the instrument's calibration was established, we used published consensus values for obsidian sources as our set of geochemical "knowns" and our geological specimens with known volcanic origin as our set of "unknowns." The set of "knowns" was used to classify the "unknowns" by geochemical source using three different methods: the online machine-learning classification tools (1) SourceXplorer and (2) AutoML for Geochemistry and (3) the traditional approach of classification by-eye using a desktop statistical software package. Although R would be an ideal statistical software choice for an "open-source" workflow, out of convenience and familiarity with its graphical interface, we used JMP 17.2. The results of the source attributions were considered successful when the identified geochemical sources matched the recorded volcanic origins of our geological specimens. Figure 1 is a schematic representation of our workflow.

Our results establish that an "open sourcing" workflow is reliable based on the PYRO calibration and consensus values, yet currently only in combination with

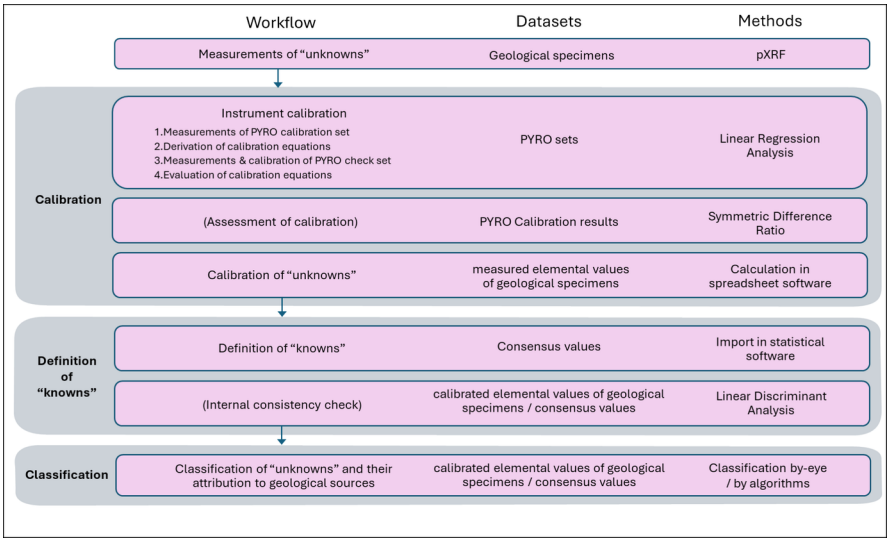

Fig. 1 Schematic representation of the proposed "open sourcing" workflow, datasets, and methods. Geological specimens from our assessment would be replaced with archaeological obsidian artifacts in a real-life application. Steps marked in brackets were necessary for our assessment but can be skipped once the instrument is calibrated

classification by-eye. At present, when we use consensus values to define sources and machine-learning solutions to assign the "unknown" geological specimens to sources, their resulting attributions are not reliable. We expect, however, that rapid developments within the field of machine learning will soon improve such a workflow and its results. Conducting the data analysis for obsidian source identifications may, in the coming years, look very different from the current procedures. The potential impact of such a workflow could be hardly overstated. It could radically diminish the need for chemical analyses of obsidian artifacts in a few specialized laboratories, and it could even be adopted for sourcing the complete obsidian assemblages on archaeological sites and, therefore, be applied to thousands of artifacts at a time. One of the most encouraging aspects of such a workflow is that it offers source identifications without having to possess expensive geological reference collections. In the South Caucasus, for geopolitical reasons, earlier studies were sometimes limited by which geological specimens were readily available to individual researchers for comparison to artifacts. The use of published consensus values for obsidian sources has, therefore, the added regional benefit of removing a major obstacle to this type of research.

Datasets

Our workflow used three datasets. The first, a Peabody-Yale Reference Obsidians (PYRO) set and the corresponding elemental measurements, was primarily used for calibration purposes. The PYRO set included, however, several geographically

relevant specimens, which also doubled as a subset of our second dataset: the geological specimens. In the third dataset consists of the consensus elemental values, which—contrary to the first two—do not reflect a physical set of specimens that we measured. Instead, it is a collection of published values for elemental concentrations for the region's many obsidian sources. The latter two datasets — the geological specimens and consensus values — served to test our workflow and assess its ability to correctly predict the sources of known obsidian specimens.

Peabody-Yale Reference Obsidians (PYRO) Set

The PYRO sets consist of well-analyzed geological specimens from around the world, and these sets are intended to be used on a collaborative basis for the calibration of EDXRF (including pXRF) instruments with a specific focus on obsidian sourcing (Frahm, 2019). PYRO's open-source status and its suitability as a reference collection especially tailored for calibrating obsidian analyses were also the main reasons why we adopted it for our calibration approach. Our workflow development and testing used one of these ten geologically identical sets intended for circulation among researchers.

A complete *PYRO set* is divided into two parts: 20 specimens belong to the *PYRO calibration set* and 15 to the *PYRO check set*. The division into two sets is intentional. It invites calibrating with one set and checking the calibrated values with the other set. The *PYRO calibration set* represents sources from East Africa, Central and South America, the Caucasus, Turkey, and Japan. The *PYRO check set* specimens have different origins than those in the calibration set, and they are intended to generally replicate an assemblage from the western USA (Frahm, 2019, Table 1). In other words, the 20 calibration specimens were chosen to represent the entire geochemical variation of obsidian sources around the globe, considering the minimum and maximum ranges of eight elements most relevant for obsidian sourcing with XRF (*i.e.*, Fe, Mn, Nb, Rb, Sr, Y, Zn, Zr), whereas the values of the check set cover a narrower range that might occur at a single archaeological site.

The published values of the PYRO sets are recommended values obtained through averaging data from multiple independent analytical techniques and laboratories to avoid inaccuracies or systematic error caused by any single method, laboratory, or other factors. Importantly — to prevent bias — pXRF measurements were not included in calculating these values (Frahm, 2019, 24–27). Our raw measurements of a PYRO set are presented in Online Resource 1. The benchmark elemental values were published in Frahm (2019: Table 15), and we repeat them here in Online Resource 2, together with the calibrated values of our measurements (for calibration methods, see the "Calibration" section).

Geological Specimens from the South Caucasus and Eastern Turkey

For our tests focused on the South Caucasus, we combined two sets of geological specimens: (1) regionally relevant specimens from the *PYRO calibration set* (see the previous section) and (2) field specimens collected by the Vishap archaeological

project (see the "Geological Field Specimens Collected by the Vishap Project" section). Although both subsets were collected by different people on different occasions, the geographic origins of these specimens are well documented, and all have been measured using the same instrument (see "Instrument and Measurement Settings") and calibration (see "Calibration"). Accordingly, both groups can be used for our comparative purposes in the same way. In our evaluation, geological specimens are primarily used as the group of "unknowns," which simulates a set of obsidian artifacts with sources yet to be determined. Using this geological dataset, rather than real artifacts, has the advantage that their geographic findspots can be used to cross-check the results for geochemical determinations of respective obsidian sources. For the sake of experimentation with machine-learning methods, we sometimes inverted the configuration and used our geological dataset as the "knowns," but such instances were clearly recorded as experimental. Altogether, our geological dataset is formed by 66 specimens representing at least 10 geochemical sources (*i.e.*, geochemically distinct volcanic events) from 10 volcanoes and volcanic complexes (for discussion of geographic vs. geochemical definition of sources, see Frahm, 2023a, p. 3). The calibrated elemental values for our set of geological specimens are presented in Online Resource 3.

A detailed description of our geological subsets follows below.

Geological Specimens from the PYRO Dataset

The first eight specimens of the *PYRO calibration set* (see the "Peabody-Yale Reference Obsidians (PYRO) Set" section) originate from the South Caucasus and eastern Turkey. Except for one trachyte from Hatis volcano, the other seven specimens belong to known obsidian sources: Aghvorik (Cal-1), Gutansar (Cal-2), Satanakar (Cal-4), Chikiani 1 (Cal-5), Meydan Dağ (Cal-6), Nemrut Dağ 6 (Cal-7), and Sarıkamış 1 (Cal-8) (cf. Frahm, 2019, Table 1). In addition to calibration, these seven specimens provided added value for this study because they effectively enlarged our group of geological field specimens.

Geological Field Specimens Collected by the Vishap Project

In 2017, while archaeologically surveying high mountain sites with dragon-stone steles (=vishaps) in Armenia and Georgia for the Vishap project, three of us — Arsen Bobokhyan, Alessandra Gilibert, and Pavol Hnila — collected geological obsidian specimens for comparative purposes (see Gilibert et al., 2012,; Bobokhyan et al., 2015; Hnila et al., 2019 for information about the Vishap project; Frahm, at this time, was surveying obsidian outcrops on Hatis volcano). Altogether, 59 obsidian specimens from five obsidian sources in Armenia and Georgia were collected by the Vishap team: Pokr Arteni, Mets Arteni, Gutansar, Hatis, and Chikiani (Table 1). Following the maps and coordinates published in Keller et al. (1996) and the online *Obsidatabase* (Varoutsikos & Chataigner, 2012), we preferably chose sampling locations at the first convenient point where obsidian was visible in surface outcrops or quarries. However, if no outcrops were identified in the available time, we collected dispersed obsidian pieces from the surface, alluvial deposits in riverbeds, or

Table 1 Locations of 59 geological samples collected directly in the field by the Vishap project (WGS84 coordinate reference system)

Label	Sampling location (field name by the Vishap project)	Alternative names (Obsidatabase, Frahm 2021, Frahm, 2023a, 2023b)	Geographic Source	Sampling context	Latitude	Longitude	Elevation (m a.s.l.)	Sample count
ALP1	Arteni lava perlite 1	Aragats flow	Pokr Arteni	Outcrop	40.34218	43.71274	1268	4
ALP2	Arteni lava perlite 2	Aragats flow	Pokr Arteni	Outcrop	40.34382	43.71316	1268	4
CH	Chikiani	Chikiani dome, Chikiani 1	Chikiani	Surface scatter	41. 49,706	43. 8759	2172	6
GF	Gutansar Fantan	Gutansar Fantan	Gutansar	Surface scatter	40.40668	44.70135	1811	4
GJ	Gutansar Jraber	Gutansar Djraber	Gutansar	Surface scatter	40.35947	44.65256	1832	4
GS	Gutansar Slope	Gutansar Djraber	Gutansar	Surface scatter	40.35947	44.65256	1832	4
GW	Gutansar Wadi	Gutansar dome	Gutansar	Erosion deposit	40.38164	44.68533	1903	4
HKKQ	Hatis Kotayk Kaput Quarry	Gutansar dome	Gutansar	Alluvial	40.38208	44.68593	1886	3
HR	Hatis Road	Hatis Beta (identified based on Frahm, 2021, Fig. 12)	Hatis	Outcrop	40.28821	44.67833	1537	4
MA	Mets Arteni	? (South flow)	Gutansar	Erosion deposit	40.33607	44.69757	1774	4
PA1	Pokr Arteni 1	Mets Arteni	Mets Arteni	Surface scatter	40.37908	43.77208	1605	10
PA2	Pokr Arteni 2	Pokr Arteni 1?	Pokr Arteni	Outcrop	40.34505	43.78211	1432	4
PA3	Pokr Arteni 3	Pokr Arteni 1?	Pokr Arteni	Outcrop	40.34731	43.78101	1458	4

erosional deposits exposed by slope-cuts. Although the risks of sampling obsidian sources from occurrences other than primary outcrops, especially surface scatters, have been recognized (Frahm, 2023a, p. 9, p. 20; see also "Geological specimen issues" section), for our analysis, the risks were limited: the geological specimens did not serve to define known sources but were instead used as "unknowns" whose attributions were to be determined. Their findspot information served only as a cross-check of the geochemical classifications.

All specimens from a single location were found less than 10 m apart, except for Mets Arteni, where we collected obsidian pieces scattered on its western slope over a ca. 500 × 200 m area. Since the Mets Arteni specimens were few and considerably worn (*i.e.*, with rounded edges and irregular surfaces showing hits and scratches), they may have rolled down from higher locations of the volcano during water-induced erosion processes. Concerning obsidian from the so-called "Aragats flow" at the Arteni complex (so named by Keller et al., 1996), where outcrops mainly were embedded in white perlite matrix, we provisionally followed the latest research, which disputed their previous association with Mets Arteni and instead argued that they must have belonged to Pokr Arteni (Frahm, 2023a, p. 9) — a conclusion later confirmed during our analysis (see results sections below).

The size of our geological specimens varied (typically 3–6 cm in largest dimension, 1–2 cm thick) since they were intended to replicate an archaeological assemblage with irregular surfaces. However, we rigorously avoided sampling very thin specimens (corresponding to blades or knapping debris), because pXRF measurements of certain elements can be less accurate when analyzing thin specimens (Frahm, 2016). None of our specimens were encrusted with sediments, and while some specimens had an eroded surface on one side, those sides were avoided during measurements.

It should be noted that the geological specimens were collected and measured before the availability of the *PYRO* sets. They were initially intended for direct comparisons with archaeological obsidian artifacts found by the Vishap project. None of the geological specimens were marked with labels to avoid potential elemental contamination.

Consensus Values of Obsidian Sources from Eastern Turkey and the Caucasus

Instead of measuring a geological reference collection to serve as the "knowns" in our workflow (see "Instrument and Measurement Settings"), we used a dataset of consensus element values (see Online Resource 4). This dataset is based on the comprehensive list of Frahm (2023a), who compiled and summarized previously published and unpublished compositional data for chemically distinct 58 obsidian sources in eastern Turkey and the Caucasus. Frahm's supplementary tables (Frahm, 2023a, S1-S60) contained the measurements from individual analytical laboratories and statistical summaries of those data (*i.e.*, mean, mean ± 1 std dev, median, first quartile, third quartile). Because the individual measurements were obtained by various analytical methods (*i.e.*, pXRF, EDXRF, WDXRF, EMPA-WDS, NAA, several ICP-MS techniques, ICP-AES, PIXE, SEM–EDS), calibrated following

procedures that may have varied between different facilities, we used the statistical summary data, which should accurately represent the geochemical composition of any given obsidian source. When available, we defined a geochemical source through all six summary values. In cases where the median and quartile values were not calculated (due to a small sample size), the geological groups were defined by the other three summary values (mean, mean + 1 std dev, mean-1 std dev). In the case of two sources, Bartsratumb and Süphan Dağ 1, for which no matching artifacts were reported and summary statistics were unavailable due to insufficient data, we defined groups based on all of their individual measurements, independent of the analytical method. Altogether, the dataset of consensus values prepared this way consists of 299 entries.

The published list of consensus values covers obsidian sources throughout the entire Caucasus and eastern Turkey region, so it offers a much broader base for identifications than the set of geological specimens treated as "unknowns" in our tests. It is a logical choice to utilize the consensus values for source identification purposes, provided that the calibrated values of our measurements are accurate and compatible with the published consensus values.

Methods

Instrument and Measurement Settings

Our measurements of the PYRO and other geological specimens (our "unknowns") were done in the Laboratory for Physical Geography of the Institute for Geographical Sciences, Freie Universität Berlin, with their pXRF instrument: a Thermo Scientific Niton XL3t GOLDD + model (serial number: XL3t-63525, large-area Si drift detector, software version CPU 7.1D, factory-calibrated on 9 March 2016). This instrument, hereafter referred to as the "FUB Niton analyzer" in this paper, was mounted in a portable SmartStand with a shielded lid for optimal safety and physical stability. To prevent sharp obsidian edges from piercing the measurement window film, a dedicated Mylar 6.0-µm thin XRF film was additionally used between the analyzer and analyzed specimens. As confirmed by repeated tests with a reference specimen (*Pokr Arteni 3.1*), the presence or absence of this additional thin film had no measurable influence on the elements relevant to obsidian sourcing.

All specimens were measured with the following default parameters: type = Mining Cu/Zn mode, units = ppm, reported measurement error = 1 sigma, and beam diameter = 8 mm. The measurement durations varied — all of our geological specimens were measured for 120 s, yet for calibration testing purposes on the PYRO sets, we also experimented with 10 s and 180 s durations. The XL3t GOLDD + instruments operate with a 2-W X-ray tube, Ag anode, and Si drift detector (SDD) with an energy resolution < 155 eV. In the instrument's "Mining" mode, each 120 s and 180 s analysis consisted of four sequential measurements of equal duration (30 s and 45 s, respectively) with different voltages and beam filters specified by the manufacturer as follows: (1) main beam conditions: 40 kV and an Al filter, (2) low: 20 kV

and a Cu filter, (3) high: 50 kV and a Mo filter, and (4) light: 6 kV and no filter. The measurements of 10 s duration used only the main beam conditions.

To minimize the drift of the energy calibration, we ran an automatic system-check ("Cal Check") of the instrument at the start of each measurement day. For consistency checks, we used the geological specimen *Pokr Arteni 3.1* as a control at the beginning and end of each measuring session, at least twice daily and, in cases of any interruptions, several times daily. For the PYRO dataset, this control specimen was remeasured as each tenth measurement. The use of this control specimen allowed us to check quickly whether our measurements were yielding consistent results.

Measured values were exported in a CSV (comma-separated values) file. The original file from the analyzer was saved in the Niton data transfer file format (with the NDT extension) and archived. For most elements, there were no differences between the CSV and the NDT files, and we converted the CSV file into an XLS (Microsoft Excel spreadsheet) file for both convenience and broad software compatibility. Differences arose only with Y, for which some measurements in the XLS file contained a non-numerical value "<LOD," standing for "below the limit of detection." Contrary to the XLS file, the NDT file contained numerical values, so we replaced the "<LOD" values in the XLS file with the original numerical values from the NDT file. Notably, some of the "<LOD" *Y* values retrieved from the NDT file were negative (due to counting statistics). We chose to use the negative numerical values instead of replacing the "<LOD" with either zeros, ones, or half of the calculated minimum detection limit (all of which are common practices) to observe any effects from further calibration.

Calibration

For accurate results, all measurement instruments should be calibrated. Using the manufacturer's built-in calibrations, however, may not be sufficient to guarantee data comparability between different techniques or different manufacturers. Instead, it is advisable to use dedicated protocols and reference standards suited for the analyzed material (*e.g.*, Frahm, 2019; Da Silva et al., 2023, Schauer et al., 2024). For our purposes, the FUB Niton analyzer was calibrated with a complete PYRO set (see 2.1), and linear regression analysis was used to compare our measurements to the benchmark elemental values for the PYRO specimens (Frahm, 2019).

Four Steps of a Calibration

Following the recommendations of Frahm (2019, p. 27), we used the *PYRO calibration set* to define the calibration equations and the *PYRO check set* to evaluate the calibration process. In practice, the calibration was divided into four steps. In the first step, we measured with factory settings all 20 *PYRO calibration set* specimens and compared them to the published benchmark elemental values. In the second step, we performed a linear regression analysis on those comparisons to derive calibration equations. In the third step, we measured all 15 *PYRO check set* specimens

and corrected them using the calibration equations from the previous step. In the fourth and final step, we performed again a linear regression analysis, this time on the calibrated values from the third step versus the published benchmark values. It should be stressed that the final step serves purely for evaluating the quality of a calibration. The calibration equations were calculated for the eight elements identified as relevant for obsidian analysis with pXRF analyzers and whose benchmark values were published for the PYRO sets: Fe, Mn, Nb, Rb, Sr, Y, Zn, Zr (Frahm, 2019, p. 26 and Table 15). Since the benchmark values are derived from multiple laboratories and analytical techniques (excluding pXRF), calibrating with them should secure the comparability of the Niton FUB analyzer measurements.

Although the Niton software allows for calibration through the insertion of slope coefficients and intercept values from linear regression equations as "User Factors," which subsequently would output calibrated elemental measurements (Frahm, 2019, 27), we opted to perform the calibration externally using standard spreadsheet software, which offered greater evaluation potential for our tests.

We measured the *PYRO sets* specimens three times, each time as a series with different duration: the first series with 10 s (main filter only), the second series with 120 s (all four filters, each for 30 s) and the third series with 180 s (all four filters, each for 45 s). The purpose of experimenting with the different durations was not to produce average values for the calibration, but to compare the effect of measurement duration on precision and uncertainty specifically for our instrument. Although previous studies proved that longer measurement times yield more precise results, the field-centric approach described by Frahm (2014) suggested that 10 s measurements can be sufficient to discern among well-distinguished sources. We were keen to quantify the effects of measurement duration on calibration because numerous archaeological artifacts that we excavated in Armenia were first screened using a measurement duration of 10 s (to be documented in a separate article). Moreover, we wanted to test whether raw values measured with a short duration can be improved by calibrating them with a linear regression trendline derived from measurements with a longer duration.

Evaluating Calibration Quality

The calibrations, expressed as linear equations, can be considered and compared either graphically or numerically. Both outputs have their own advantages and challenges, which inspired us to devise a new evaluation method, described in detail below.

A graphical comparison offers a straightforward way to perceive the quality of any calibration: on a dedicated chart, one can observe how close a calibration trendline comes to an ideal trendline in terms of orientation and shift and how far away the individual measurements stray from a calibration trendline. An ideal trendline would be a perfectly linear, 1:1 relationship — in this instance — between the measurements obtained by the FUB Niton analyzer and the published benchmark elemental values of the PYRO sets. The closer a calibration trendline comes to an ideal trendline, the less severe the correction that needs to be applied by the calibration,

and the closer the individual measurements fall along the calibration trendline, the more reliable and consistent the calibrated results will be. Less straightforward is how to graphically compare quality between a series of calibrations (*e.g.*, different measurement times for the same element, different calibration approaches). After a few calibrations, each with its own trendline, it becomes too challenging to represent and compare them on a visual basis.

A numerical comparison is more readily summarized, and it is a reasonable alternative to visual evaluation, yet it also presents challenges. Specifically, the character and quality of linear regression equations can be numerically expressed with three components — slope, intercept, and R^2 (coefficient of determination). Of these three numerical components, only the slope and R^2 already express ratios in numerical values that can be directly compared or averaged. The intercept, however, expresses a shift in absolute numbers, which cannot be directly compared or averaged because the concentration ranges of individual elements (in parts per million) differ vastly. Ignoring the intercept and comparing trendlines based solely on their slope is not a good option. If, for example, a pXRF instrument has a systematic shift, its calibration trendline might be equal to one and, thus, seem the finest possible in terms of slope evaluation. Yet, all of its measured values would be inaccurate without considering the actual shift, expressed via the intercept. Despite its importance for calibration, the traditionally used statistical methods (Student's *t*-test, ANCOVA, *etc*.) concentrate on comparisons of slopes. Moreover, they operate based on sophisticated mean values, whose correct implementation and interpretation are challenging even for experts (see Andrade & Estévez-Pérez, 2014).

Symmetric Difference Ratio (SDR)

To overcome the issues with numerical comparisons, we propose a new method considering the intercept and the slope within a single geometric entity: an area. Operating with areas might sound counter-intuitive in the context of linear regression trendlines. Yet, because any single trendline is only assessed for a specific range for calibration purposes, we can use that range to delimit a pertinent area. The minimum and maximum values of the range will delimit its area on the left and the right. On the top, its area will be delimited by the trendline itself, and on the bottom by the x-axis (Fig. 2). In this way, we can define areas pertinent to both the ideal trendline (polygon *AEFC* in Fig. 2) and the evaluated trendline (polygon *BEFD* in Fig. 2). Their symmetric difference can be expressed as a ratio. This ratio offers — we propose — a simple, understandable, and solid basis for such comparisons, both geometric and numerical. The smaller the symmetric difference ratio between the two areas, the closer a calibration trendline is to an ideal one in terms of the combined slope and shift.

The symmetric difference is a mathematical concept from set theory (*e.g.*, Boschini et al., 2022; Ghosh & Deguchi, 2008). A symmetric difference between two sets is defined as the set of elements present in either set, but not in both (*i.e.*, it equals the union of both sets minus their intersection). Among others, it has found applications in graph theory (Flament, 1963) and geographic information systems

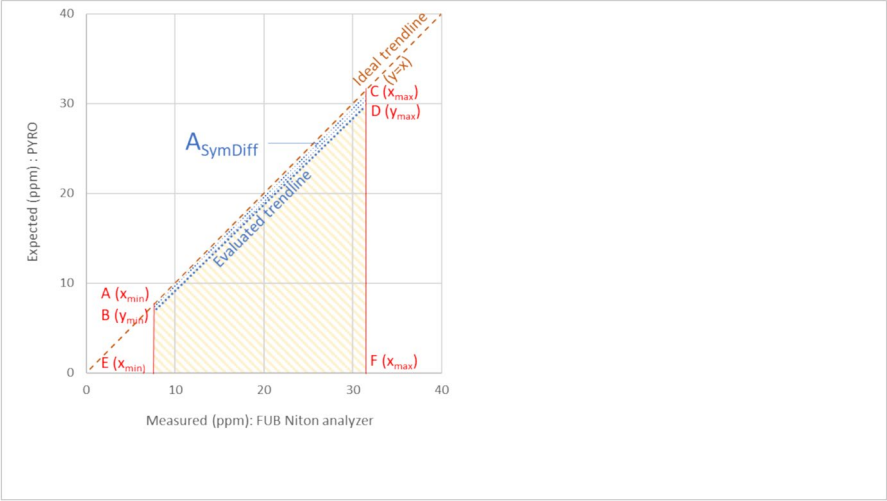

Fig. 2 Symmetric difference ($A_{SymDiff}$) between an area pertinent to the ideal trendline (polygon AEFC) and an area pertinent to the evaluated trendline (polygon BEFD) defined down to the x-axis. In cases, when the ideal trendline does not intersect with the evaluated trendline in the evaluated range, the symmetric difference of their pertinent areas (polygon ABDC) will be equal to the arithmetic difference of the area of the smaller polygon being subtracted from the area of the larger polygon

(*e.g.*, Korstanje, 2022 & Löwe et al., 2022). To our knowledge, this approach has not yet been applied to calibrations of geochemical data, as we outline here.

In geometric terms, the symmetric difference corresponds to the parts belonging to either one or the other area, outside the overlapping space of those areas. In the case of two trendlines delimited by the same minimum and maximum values on the x-axis, the symmetric difference is the area encompassed between both trendlines.

A calculation of symmetric difference is equal to a difference of a simple numeric subtraction only if one of the two pertinent trendline areas is wholly encompassed within the other — in such cases, a simple subtraction of the smaller area from the larger area will resolve the calculation task (Fig. 2). However, as soon as one area is not entirely overlapped by the other, there will be overlaps combined with protrusions of one area outside the other and vice versa (Fig. 3). Consequently, the two areas cannot be simply subtracted because a subtraction difference does not distinguish between overlaps, protrusions, and counter-protrusions. If, for example, the two areas were congruent in size but not in orientation, their overall numeric subtraction difference would be zero, while their overall symmetric difference would depend very much on the orientation and the size of their overlap.

The spatial component is therefore essential for the calculation of a symmetric difference. The symmetric difference as an area can be broken down into basic geometric shapes of triangles and rectangles — all of which can be calculated with appropriate formulae in a spreadsheet software. One can choose from several computational approaches. For our purposes — given that calibration trendlines are rarely parallel or identical to the ideal trendlines but still mostly intersect — the best approach was to sum two similar triangles defined by an intersect point between the evaluated and the ideal

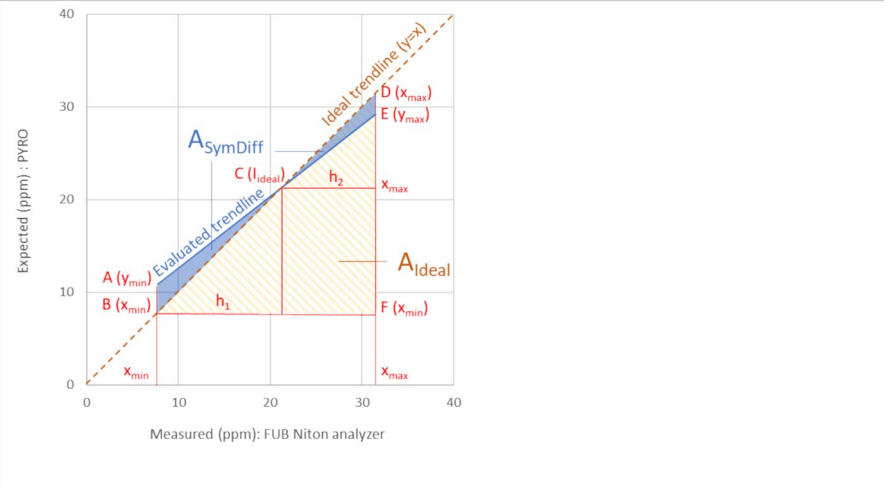

Fig. 3 Comparison of symmetric difference area (A_{SymDiff}) with the ideal triangle surface area (A_{Ideal}). When the evaluated trendline intersects with the ideal trendline within the evaluated range, the A_{SymDiff} can be calculated as the sum of areas of two opposite triangles delimited by both trendlines, by their intersection, and by the evaluated range ($\Delta ABC + \Delta CED$). The ideal surface area is defined by the triangle BFD

trendline (triangles ABC and CDE in Fig. 3). Each triangle area is calculated by multiplying its respective base (AB, DE on Fig. 3) with its respective height (h_1, h_2 on Fig. 3) and dividing the results by two. The bases and heights are defined by the range limits on both axes and by the intersect between both trendlines. Because the direction of the deviation from the ideal trendline is irrelevant, absolute values are used for the subtractions that define the bases of these triangles. In mathematical notation, the formula to calculate the total area of both triangles can be expressed as follows:

$$A_{\mathrm{SymDiff}} = \frac{\left|y_{\min} - x_{\min}\right|.\left(I_{\mathrm{ideal}} - x_{\min}\right) + \left|y_{\max} - x_{\max}\right|.\left(x_{\max} - I_{\mathrm{ideal}}\right)}{2}$$

Where A_{SymDiff} is the area of the symmetric difference, $x_{\max}$ is the maximum value on the x-axis, $x_{\min}$ is the minimum value on the x-axis, $y_{\max}$ is the result of the evaluation equation (linear regression equation behind the evaluated trendline) for $x = x_{\max}$, $y_{\min}$ is the result of the evaluation equation (linear regression equation behind the evaluated trendline) for $x = x_{\min}$, and I_{ideal} is the intersect of the evaluated trendline with the ideal trendline, calculated from the slope and intercept values of the evaluated trendline as $I_{\mathrm{Ideal}} = \frac{\mathrm{Intercept}}{1-\mathrm{Slope}}$.

Compared with other computational approaches, the main advantage of our approach is its almost universal utility. It works as a single formula for most (not all) scenarios since it does not depend on whether the intersect is below the evaluated range, within it, or above it. When the ideal trendline and the evaluated trendline intersect within the evaluated range, the symmetric difference of both areas will have the shape of two similar triangles with vertically opposite angles — that

is, sharing the same vertex (=the intersect) but having their bases on the opposite sides. The formula will sum their areas together. When the ideal trendline and the evaluated trendline intersect outside the evaluated range, the symmetric difference will have the shape of a trapezoid with two parallel sides. Yet, the shorter of these sides will also serve as a base for a triangle, whose third vertex is the intercept and both the trapezoid, and the triangle will be encompassed within a bigger triangle defined by the intercept and the larger base of the trapezoid. In the latter case, the area of the smaller triangle will return negative values, so, in our formula, it will be numerically detracted from the larger triangle, leaving the trapezoid area of interest as the result. Only in the unlikely case if there is no intersection because both trendlines either run entirely in parallel or are coincident would the proposed equation return an error, and an alternative approach would be inevitable. In the former case, when both trendlines run in parallel, the area of symmetric difference will have the shape of a parallelogram and the alternative calculation approach could accordingly multiply the parallelogram's side by its height; in the latter case, when both lines are coincident, there is zero symmetric difference and thus no need for calibration.

For assessment purposes, the raw values of symmetric difference are still unsuitable for direct comparisons, because the concentration ranges (in parts per million, ppm) of different elements vary so considerably (*e.g.*, Nb or Y are in the lower tens of ppm, while Fe is in the tens of thousands of ppm, so even a tiny geometric difference in Fe would produce a numerically very high difference in terms of area). To make the symmetric difference intercomparable between different elements, the A_{SymDiff} results must be detached from the measurement values (in ppm) of their respective elements and instead expressed in ratio to some ideal surface.

The definition of the ideal surface, however, is not as straightforward as it might seem. If the ideal surface was defined down to the x-axis ($y=0$) — as we have done for the concept of symmetric difference — elements whose range starts far from zero (*e.g.*, ca. 4000 ppm for Fe in the *PYRO check set*) would produce ideal areas too elongated in comparison to elements with ranges starting close to zero. Similarly, if the ideal surface were delimited by the lowermost y as its lower axis, its shape and size would depend on whether y_{min} is below or above the ideal trendline, because for all $y_{\mathrm{min}} > x_{\mathrm{min}}$ the ideal surface would be triangle-shaped and delimited by x_{min}, yet for all $y_{\mathrm{min}} < x_{\mathrm{min}}$, the triangle would be enlarged by a rectangle defined by x_{min}-y_{min} as its vertical side. Consequently, evaluated areas of the same size but placed below or above the trendline would produce slightly different ratios, which is undesirable. To avoid both mentioned distortions, the best candidate for an ideal surface is a triangle area delimited by the ideal trendline as its hypotenuse, $y=x_{\mathrm{min}}$ as its lower side and $x=x_{\mathrm{max}}$ as its right side. It is the smallest possible surface encompassed by the ideal trendline in any given range. We refer to it as the ideal triangle surface area (A_{ideal}), which can be calculated as:

$$A_{\mathrm{ideal}} = \frac{\left(x_{\mathrm{max}} - x_{\mathrm{min}}\right)^2}{2},$$

where A_{ideal} is the area of a triangle pertinent to the ideal trendline, x_{max} is the maximum value on the x-axis, and x_{min} is the minimum value on the x-axis.

Once the area of the symmetric difference and the area of the ideal triangle surface are calculated, we can express them as a ratio, by dividing the symmetric difference (A_{SymDiff} as the dividend) with the ideal triangle surface area (A_{ideal} as the divisor) and by multiplying the resulting quotient with 100. The obtained result expresses the percentage ratio of similarity between the evaluated area and the ideal triangle area. We refer to this ratio as the symmetric difference ratio or *SDR* coefficient. In mathematical notation, the *SDR* coefficient calculation is expressed as:

$$SDR = \frac{A_{\mathrm{SymDiff}}}{A_{\mathrm{ideal}}} \times 100,$$

where A_{SymDiff} is the area of the symmetric difference (see the first equation) and A_{ideal} is the triangular area pertinent to the ideal trendline, delimited by minimum and maximum values on the x-axis (see the previous equation).

For most of our assessment purposes, the *SDR* coefficient was used as the first step and the single most meaningful criterion, encompassing both the slope and intercept. Moreover, it encompasses them over their entire evaluated range, so the value itself is not a mean but a numerical representation of the entire range. Since it is always a positive number, it cannot be arithmetically "neutralized," as could happen when averaging two identical slopes running in opposite directions. Thus, *SDR* enables a quick evaluation of the overall calibration quality among multiple listings in a table. Also, and most importantly, it can be used for comparative purposes to calculate an average among variously defined groups. Because it quantifies the proportion in which an area pertinent to an evaluated trendline differs from that pertinent to an ideal one, the lower the *SDR* value, the less deviation from the ideal trendline, and, accordingly, the more accurate a calibration is.

For detailed evaluation, each calibration should, in addition to the *SDR,* consider the R^2 values, which reflect the instrument's precision. Moreover, each comparison of a series should include the *standard deviation (=SD.P)* values — to express how well calibration models represent the underlying measurement data and how far away from the mean the other values in a series deviate. To streamline the comparisons, we attributed high accuracy to all SDR values $\leq 5\%$ and high precision to all R^2 values ≥ 0.98.

When interpreting the results, one cannot renounce detailed calibration graphs, given that they visualize not only the calibration and ideal trendlines but also the distribution of one's individual measurements along trendlines for each element. In this way, one can readily note factors contributing to shifts, distortions, and variance and, therefore, interpret what is usually hidden behind the numerical expressions of each calibration.

Internal Consistency Checks by LDA

All datasets were checked for internal consistency with linear discriminant analysis (LDA) in the JMP 17.2 statistical software. The eight elements served as covariates, while the reported source served as the categorical variable. Each dataset was

used in its entirety, first as the test/training/defining set and second, as the set whose sourcing values were to be determined by the algorithm. The aim was to determine how the algorithmically generated classification of obsidian sources corresponded to the recorded geological sources of the same specimens. In an optimal case, the correspondence should be one to one. Any misclassification suggests inconsistencies in the geochemical groupings, such as geochemical overlaps between two or more geological sources, sources defined by an insufficient number of specimens, and/or statistically significant differences between specimens attributed to the same source. Consequently, misclassified geological sources must be treated with extra caution when classifying archaeological artifacts of unknown origins. Specimens flagged as suspicious by internal consistency checks may need to be double-checked or even excluded from further analysis.

Standard LDA vs Open LDA

LDA is typically employed to classify unknowns into groups through a defined set of already-known categories. Its resulting number of classes — in our case, obsidian sources — is thus fewer than or equal to those from the defining set. However, we additionally experimented with the chance that some specimens may have belonged to a currently unrecognized, "other" source. By permitting the algorithm to consider this alternative, we expected that inconsistent groups will be revealed not only through misclassifications, but also through a higher number of outliers classified as unknown sources. Moreover, it can be useful to deliberately leave the possibility of "other" sources open for archaeological applications, because despite decades of archaeological research in the South Caucasus and surrounding areas, the current number of recognized obsidian sources cannot be considered final. Recently a new source was identified in northwestern Iran, both as a geological outcrop and among artifacts at nearby archaeological sites (Orange et al., 2021). We expect that more obsidian sources might also be identified in eastern Turkey. Without allowing the option of "other" sources, potential artifacts from unknown sources would be forcefully attributed to one of the currently recognized sources with which the LDA was trained and, thus, would be misclassified as a result.

For implementation, we usually set the prior probability of the "other" group to 0.01 (=1%) in the experimental feature "consider new features" of JMP 17.2. Only in exceptional cases in which we experimented with non-standard configurations and the defining set contained too few sources, did we set the prior probability higher, proportional to the number of missing sources when compared with the dataset of all consensus values.

In the absence of a generally accepted terminology, we refer to the LDA reflecting a chance of an unknown source as the "open LDA" and to the one with sources restricted to known categories as the "standard LDA." In the following two sections, we describe in detail the procedure for each of the two datasets.

Internal Consistency Check of the Dataset with Geological Samples

The dataset of geological specimens extracted from the *PYRO calibration set* and measured with 120 s duration (see "Geological Specimens from the PYRO Dataset") was combined with the geological field specimens collected by the Vishap project (see "Geological Field Specimens Collected by the Vishap project"). These two datasets were joined because both comprise geological specimens with known origins and were measured with the same instrument, the FUB Niton analyzer. Afterwards, the joint geological dataset was calibrated, and the calibrated elemental values were checked using LDA (both standard and open) for internal consistency concerning the definition of obsidian sources.

Internal Consistency Check of the Dataset with Consensus Values

The dataset of consensus elemental values based on Frahm (2023a) (see the "Consensus Values of Obsidian Sources from Eastern Turkey and the Caucasus" section) was checked using LDA (both standard and open) for internal consistency of the geochemical groupings, following the same methodology as described above (the "Internal Consistency Checks by LDA" section).

Attribution to Geochemical Groups

Due to the high number of 58 candidate origins for potential identification, attributing specimens to volcanic obsidian sources in eastern Turkey and the South Caucasus is challenging. With so many geochemical groups, the ranges for one or more of the eight relevant elements are more likely to overlap, which might result in ambiguities. We tested attributions to geochemical groups using four different classification methods, always with our measurements of 120 s duration. The first three classification methods are guided by statistical algorithms, while the fourth method is the traditional procedure: by-eye examination of scatterplots and subsequent manual attributions of "unknowns" to likely sources.

Algorithmic Classification of Geological Samples using LDA with the Consensus Values

Primarily we ran standard LDA using the published consensus elemental values for the regional obsidian sources (the "Consensus Values of Obsidian Sources from Eastern Turkey and the Caucasus" section) as the set of "knowns." To improve the statistical assessment of the discriminant analysis, we selected all median values to serve as an evaluation set, while the remaining values served as the training set. Because some entries in the dataset of consensus values do not contain all eight elements relevant for our analysis, and because some sources being represented by insufficient number of entries do not have a median, our evaluation set contains 42 entries (15%), while the training set contains 240 entries (85%). As the test set, we used the data for the geological obsidian specimens (see the "Geological specimens

from the South Caucasus and eastern Turkey" section) — that is, they served as the "unknowns."

In JMP, the LDA is performed on a single table, so we produced a unified table by appending the geological specimens to the dataset of consensus values. The eight elements reported for the PYRO specimens were used as covariates to quantitatively define sources, while the "site" reported for the consensus sources served as the categorical variable. To enforce descriptive statistical output for a comparison between the test set and the training + evaluation sets in JMP, values from the "source" field of our dataset of geological specimens were written to the "site" field of the unified table. The original separation of provenance information into two fields with distinct naming conventions was intentional and a means to enforce that JMP algorithms did not use the provenance information from the geological specimen data to fine-tune the prediction model (as it indeed seems to happen when a separate evaluation set is missing, and thus the test set is treated as an evaluation set). The division of entries from our unified table between training, evaluation, and test sets in JMP was obtained by numeric coding in an extra added field. The resulting predictions were output in a dedicated column and compared with the findspot of the geological specimens.

We also experimented with running open LDA, in which there was a 1% prior probability of an unknown source. Although our "unknown" dataset should not contain any unrecognized sources not in the "known" dataset, this setting should account for any incompleteness of the consensus dataset if applied to an archaeological assemblage, so we wanted to test its behavior for future studies.

Additionally, we experimented with an inverted configuration, in which the geological specimens served as the defining corpus of "knowns" and the consensus elemental values were treated as the "unknowns" in LDA. This kind of prediction would only work correctly for obsidian sources among the ten included in our geological dataset, and all other sources would be forcibly (and inaccurately) attributed to one of those ten. To account for the extreme incompleteness of sources in this inverted configuration, we furthermore used open LDA in which we set the prior probability of "other" sources very high — to 82.76%, reflecting the proportion of the sources missing from the defining set (*i.e.*, 48 out of the total 58 sources present in the published set of consensus elemental values). Given the small size of the "knowns" set, we did not divide it into training and validation sets. The entire set of "unknowns" served as the training set. To ensure that JMP algorithms do not use the test set to fine-tune the prediction model, the defining field "source" was left empty for the consensus values.

Algorithmic Classification with SourceXplorer

SourceXplorer is an automated software tool specialized for geochemical groupings of obsidian or other toolstone sources (*e.g.*, the original paper tested felsic raw materials from the Pacific Northwest of North America; McMillan et al., 2022). It can be used either as a stand-alone application or online (https://sourcexplorer.org/, accessed on 13 April 2024). Its developers intended to facilitate and standardize interpretations of pXRF measurements, making them reproducible and trustworthy

even for non-specialists (McMillan et al., 2022). Primarily, LDA and principal component analysis (PCA) are bundled to define the source groups, whereas convex hull areas and confidence intervals are utilized to calculate whether specimens or artifacts fall inside or outside these defined groups.

For SourceXplorer to work, it is necessary to have two datasets: (1) a set with defined geological sources and (2) another set with the unknowns that need to be classified. In our case, the consensus values (the "Consensus Values of Obsidian Sources from Eastern Turkey and the Caucasus" section) were the defined sources (the knowns), while the geological specimens were the unknowns (the "Geological Specimens from the South Caucasus and Eastern Turkey" section). Both sets were prepared as CSV files according to the specifications needed by the SourceXplorer and exemplified by the template files downloadable from its online version. In this respect, it is important to note two peculiarities for file preparation. First, the sources file must contain latitude and longitude because the analysis only works for those sources visible on the map interface. Second, the dataset cannot contain certain special characters (otherwise the LDA fails to plot, and the entire analysis fails). Once successfully imported, only the entries with all field values will be considered, so if one of the elements is missing, the entire entry will be ignored. Of 299 entries in our dataset of consensus values, 282 from 56 sources were used by SourceXplorer for analysis. From the dataset of geological specimens, the software considered all of the entries.

Three possible sourcing outcomes are offered by SourceXplorer based on where each specimen falls with respect to the convex hull and confidence interval of LDA or PCA groups: (1) "basic" (within either the convex hull or confidence interval of either LDA or PCA), (2) "standard" (within either the convex hull or confidence interval of both LDA and PCA), and (3) "robust" (within both the convex hull and confidence interval of both LDA or PCA). If any specimen falls outside all of these combinations, it will be classified as belonging to an "Unknown" source. If a specimen's origin is classified to a certain LDA group but with a probability below a user-defined threshold (default value set to 70%), the sourcing outcome will result as "Ambiguous" — a statistical solution that was adopted to account for overlapping sources.

Algorithmic Classification with AutoML for Geochemistry

The second automated online system that we tested is AutoML for Geochemistry. AutoML stands for "automated machine learning pipeline for geochemical analysis" (Alférez et al., 2022). The tool is available online (http://216.249.119.85:8889/, the address published in the article and accessed on 15 April 2024; accessible via https://geochem.cs.southern.edu/, personal communication of Dr. Alférez from 16 October 2024). It provides a guided step-by-step interface with data preparation, and it offers both supervised and unsupervised learning methods. We assessed this software using the datasets of consensus values and our geological specimens.

Once a file is uploaded, the algorithm offers a choice to eliminate entries with missing values or replace them with the Extra Trees Regressor imputation method.

In our case, we opted for eliminating the entries, given that the measurement values of elements cannot be convincingly interpolated. After eliminating entries with missing values, 282 records remained in the dataset of consensus values (as was the case for SourceXplorer). Our dataset for the geological specimens did not have any missing values (which was also the case for SourceXplorer).

At this point, either supervised or unsupervised learning can be chosen. When the supervised method is chosen, one must upload the training dataset containing the classified values — in our case, this was the consensus values dataset. The program then proceeds in six steps. In the first step, the variables need to be chosen. In our case, the "Source" was the variable for class (dependent variable), and all eight elements were chosen as variables for the features (independent variables). In the second step, one can choose which classes are used to train the model and whether the data are balanced by upsampling. Upsampling serves to equalize classes so that none of them dominates the algorithm. We chose all sources as our classes and checked "apply" for upsampling. In the third step, the data are split between the training and the testing sets. A value for the proportion of the training set must be chosen from 70%, 75%, or 80% (and the proportion for the testing set is automatically updated accordingly). We opted for the highest proportion of 80% for the training purposes in order to obtain the highest possible accuracy. In the fourth step, the algorithm performs a classification based on five models: k-nearest neighbors, decision trees, support vector machines, logistical regression, and multilayer perceptron. It gives the results for all of them, yet it chooses the one it considers best. Once our dataset of geological specimens was uploaded, it was classified in the fifth step, and the results were shown. The results are downloaded in a CSV format in the final sixth step.

Under unsupervised machine learning, one can only upload a single file for evaluation, so we evaluated the consensus values dataset separately from the geological specimens. Once the file is uploaded, the algorithm performs PCA with selected numeric fields to reduce the dimensionality of the data. Based on the statistically summarized results from the PCA, one needs to specify the number of principal components on which the subsequent K-means cluster analysis is based. In our case, we chose all components covering variance > 1%. At the final stage, the K-means cluster analysis returns the number of resulting clusters — in this case, geochemical groups — anticipated to be present in the dataset.

Classification By-Eye

Classification by-eye refers to a visual inspection of scatterplots, which is often recommended as one of the most powerful methods for source assignments (*e.g.*, Shackley, 1995; Glascock, 2020). Although classification by-eye can be performed on plots combining various elements or dimension-reducing statistical components and we may have attributed all "unknown" specimens to geochemical groups using such scatterplots, it is timesaving to pre-group the "unknown" specimens using an LDA classification, especially since it was already done in one of the previous steps (see "Algorithmic Classification of Geological Samples Using LDA with the Consensus Values").

Our classification by-eye followed three steps, all performed in the JMP 17.2 statistic package. In the first step, we assigned a specific symbol to each obsidian source. Matching obsidian sources were represented in both datasets with the same symbol. A visual differentiation between predicted sources and consensus values was obtained by color: symbols of predicted sources were color-coded, whereas the consensus values were grey. In the second step, we performed PCA with all eight elements and used the same coding system for colors and symbols. In the third and final step, we visually inspected every specimen on the following scatterplots: Nb vs. Sr, Rb vs. Zr, Y vs. Fe, Mn vs. Zn, and the first two principal components, PC2 vs. PC1. These four scatterplot combinations cover all eight elements reported for the PYRO sets, while the first two principal components cover 79.5% of variance among the measurements in eight-dimensional space.

If any specimen stood out during the visual inspection because its color or symbol differed from the rest of the group, we investigated the reason. If some groups overlapped on one scatterplot, we double-checked their respective positions on the other scatterplots. Once we identified a convincing match between every single geological specimen and the geochemical groups of the consensus values, the identified source was recorded. After all of the specimens were attributed to geochemical groups, we compared the results of both datasets and checked for overlaps or shifts.

Results

In this section, we present our results following the sequence of methods from the section above. Given the complexity of different methods being combined with different datasets and different steps, we first include a tabular overview of their combinations, including direct references to where the results and their respective descriptive statistics are located (Table 2).

The subsections that follow present the results in more detail, while the discussion of results is later presented in the "Discussion" section.

Calibration Equations

The calibration results are presented in two sections. In the first, we present the equations obtained with the *PYRO calibration set*, and in the second section, we present the equations of the subsequent evaluation using the *PYRO check set.*

Calibration with the PYRO Calibration Set

The uncalibrated measurements of the complete PYRO set (calibration set + check set) and geological specimens are listed in *Online Resource 1*. For potential future comparisons, we report all measured elements, not just the eight elements whose benchmark values are currently published for the PYRO sets. All of our specimens were measured as a three series, each one with a different duration: 10, 120, and 180 s. The linear regression equations obtained in the second step of the calibration

Table 2 Tabular reference to results according to workflow steps, methods, and datasets used to define the "knowns" and "unknowns." For methods that use only one dataset, the "knowns" and the "unknowns" are the same. "Statist." numbers refer to the pages in Online Resource 6, which contains descriptive statistics for respective analyses

Step	Method	"Knowns"	"Unknowns"	Results	Statist
Internal consistency check in JMP	Standard LDA	Geological specimens	Geological specimens	"Internal consistency check of geological field specimens" section; Online Resource 3, Column K	13–16
Internal consistency check in JMP	Open LDA	Geological specimens	Geological specimens	"Internal consistency check of geological field specimens" section; Online Resource 3, Column L	17–21
Internal consistency check in JMP	Standard LDA	Consensus values	Consensus values	"Internal consistency check of the consensus values" section; Online Resource 4, Column D	81–85
Internal consistency check In JMP	Open LDA	Consensus values	Consensus values	"Internal consistency check of the consensus values" section; Online Resource 4, Column E	86–89
Algorithmic prediction in JMP	Standard LDA	Consensus values	Geological specimens	"Algorithmic classification of geological specimens using LDA and consensus values" section; Table 6; Online Resource 3, Column M	90–93
Algorithmic prediction in JMP	Open LDA	Consensus values	Geological specimens	"Algorithmic classification of geological specimens using LDA and consensus values" section; Table 7; Online Resource 3, Column N	94–97
Experimental inverted prediction in JMP	Standard LDA	Geological specimens	Consensus values	"Algorithmic classification of geological specimens using LDA and consensus values" section; Online Resource 4, Column F	133–143

Table 2 (continued)

Step	Method	"Knowns"	"Unknowns"	Results	Statist
Experimental inverted prediction in JMP	Standard LDA	Geological specimens w/o 2 outliers	Consensus values	"Algorithmic classification of geological specimens using LDA and consensus values" section; Online Resource 4, Column G	144–154
Experimental inverted prediction in JMP	Open LDA	Geological specimens w/o 2 outliers	Consensus values	"Algorithmic classification of geological specimens using LDA and consensus values" section; Online Resource 4, Column H	155–167
Algorithmic prediction in SourceXplorer	Standard LDA	Consensus values	Geological specimens	"Algorithmic classification with SourceXplorer" section; Online Resource 5, Column J;	-
Algorithmic classification in SourceXplorer	SourceXplorer (LDA + PCA + convex hulls + confidence intervals)	Consensus values	Geological specimens	"Algorithmic classification with SourceXplorer" section; Table 8; Online Resource 5, Column T	-
Algorithmic classification in AutoML for Geochemistry	Supervised classification without upsampling (K-nearest neighbor)	Consensus values (split 80:20 for training:testing)	Geological specimens	"Algorithmic classification with AutoML for Geochemistry" section; Online Resource 3, Column Q; Online Resource 5 (Register 5E)	98–107
Algorithmic classification in AutoML for Geochemistry	Supervised classification with upsampling (K-nearest neighbor)	Consensus values (split 80:20 for training:testing)	Geological specimens	"Algorithmic classification with AutoML for Geochemistry" section; Online Resource 3, Column R; Online Resource 5 (Register 5F)	108–119

Table 2 (continued)

Step	Method	"Knowns"	"Unknowns"	Results	Statist
Algorithmic classification in AutoML for Geochemistry	Unsupervised classification (Elbow method, K-Means, 5 PCA components)	Geological specimens	Geological specimens	"Algorithmic classification with AutoML for Geochemistry" section; Online Resource 3, Column S	120–123
Algorithmic classification in AutoML for Geochemistry	Unsupervised classification (Elbow method, K-Means, 5 PCA components)	Consensus values	Consensus values	"Algorithmic classification with AutoML for Geochemistry" section; Online Resource 4, Column I	124–132

process (for method, see "Four Steps of a Calibration") are numerically summarized in Table 3, where they are listed separately by duration.

In addition to this numerical overview, a graphical calibration summary for each element is given in the form of a detailed graph with the specimens used to define them (Fig. 4). These graphs reflect only the measurement series of 120 s (the graphs based on the 10 s and 180 s measurements are only marginally different and are not reproduced here due to space considerations, but they can be derived from the calibration equations in Table 3).

Evaluation with the PYRO Check Set

The original values of the *PYRO check set* measured by the FUB Niton analyzer are documented in Online Resource 1. As in the case of the calibration set, we report not only the eight elements of interest, whose benchmark values were published, but also our other measured elements (for the sake of potential future comparisons).

After their calibration via linear regression (see Table 3), the elemental values of the *PYRO check set* measurements were recalculated accordingly and hereafter are referred to as the calibrated values (Online Resource 2). Using these calibrated values, we again performed LDA, this time for evaluative purposes. The resulting equations are numerically summarized in Table 4, and a graphical evaluation summary — also for the 120 s series only — is illustrated in Fig. 5.

Our experiments, in which the *PYRO check set* series measured with shorter duration (10 s and 120 s) were calibrated using linear regression derived from the 180 s measurements of *PYRO calibration set*, are summarized in Table 5.

Internal Consistency Checks with Linear Discriminant Analysis

Internal Consistency Check of Geological Field Specimens

With the standard LDA check for internal consistency, a single specimen was classified outside of its geographic sampling location: MA1.7 from Mets Arteni was geochemically attributed without any doubt to Gutansar (Online Resource 3, column K). With the open LDA, accounting for a prior 1% probability of "other" sources (leaving 9.9% prior probability for each of the ten sources represented by the dataset), two specimens were classified outside their sampling location: MA1.7 as Gutansar and MA1.10 as "other" (Online Resource 3, column L).

Internal Consistency Check of the Consensus Values

Out of 299 records of the consensus values dataset, 17 were excluded from the evaluation because they lacked values for at least one of the eight elements. Consequently, they could not produce a result (*i.e.*, all three entries for Syunik Satanakar 2, two Syunik Satanakar 3, two Süphan Dağ 1, two Ptghni, two Kars Akbaba Dağ, two Erzurum Tambura, two Erzincan 2, and two Bartstratumb). Out of the remaining 282 entries, the standard LDA revealed only two minor misclassifications: both

Table 3 Calibration equations for the FUB Niton analyzer as derived through linear regression analysis of 20 samples from the PYRO calibration set. The expressions define the y-axis (the expected benchmark values) with the help of the measured values on the x-axis (FUB Niton raw values). SD(P) stands for standard deviation of a population defined by all eight elements. SDR stands for symmetric difference ratio (for calculation details see the "Symmetric Difference Ratio (SDR)" section)

		10 s measurements				120 s measurements				180 s measurements			
		Slope	Intercept	R^2	SDR (%)	Slope	Intercept	R^2	SDR (%)	Slope	Intercept	R^2	SDR (%)
Fe	$y=$	0.979x	+780	0.998	1.08	0.961x	+1145	0.998	2.00	0.961x	+1033	0.998	2.08
Mn	$y=$	0.943x	+129	0.990	4.44	0.943x	+133	0.989	4.86	0.916x	+149	0.987	4.71
Nb	$y=$	1.040x	+2.57	0.998	5.64	1.038x	+1.48	0.997	4.87	1.029x	+1.90	0.997	4.23
Rb	$y=$	1x*	−2.69	0.995	2.61	1.048x	−9.76	0.994	2.55	0.992x	−2.53	0.991	3.73
Sr	$y=$	1.045x	+1.74	0.999	5.34	1.079x	−0.16	0.999	7.84	1.072x	0	0.999	7.27
Y	$y=$	1.013x	+20.3	0.998	18.97	1.023x	+20	0.998	20.17	0.994x	+20.6	0.996	17.41
Zn	$y=$	0.951x	+10.2	0.996	2.52	0.952x	+10.7	0.998	2.39	0.976x	+8.80	0.996	1.68
Zr	$y=$	0.914x	+17.5	0.999	7.27	0.909x	+21.2	0.999	7.46	0.903x	+21.9	0.999	8.02
Mean		0.986x	+120	0.997	5.98	0.994x	+165	0.996	6.52	0.981x	+154	0.995	6.14
SD(P)		0.044x	+253	0.003	5.25	0.057x	+373	0.003	5.56	0.052x	+335	0.004	4.74

*All slope values in this table were rounded to a maximum of three decimal places for practical purposes. The Rb slope value of 1 is actually 0.99998862, and thus not completely parallel to the ideal trendline. Further computation in the software (*e.g.*, for SDR) always used the unrounded values

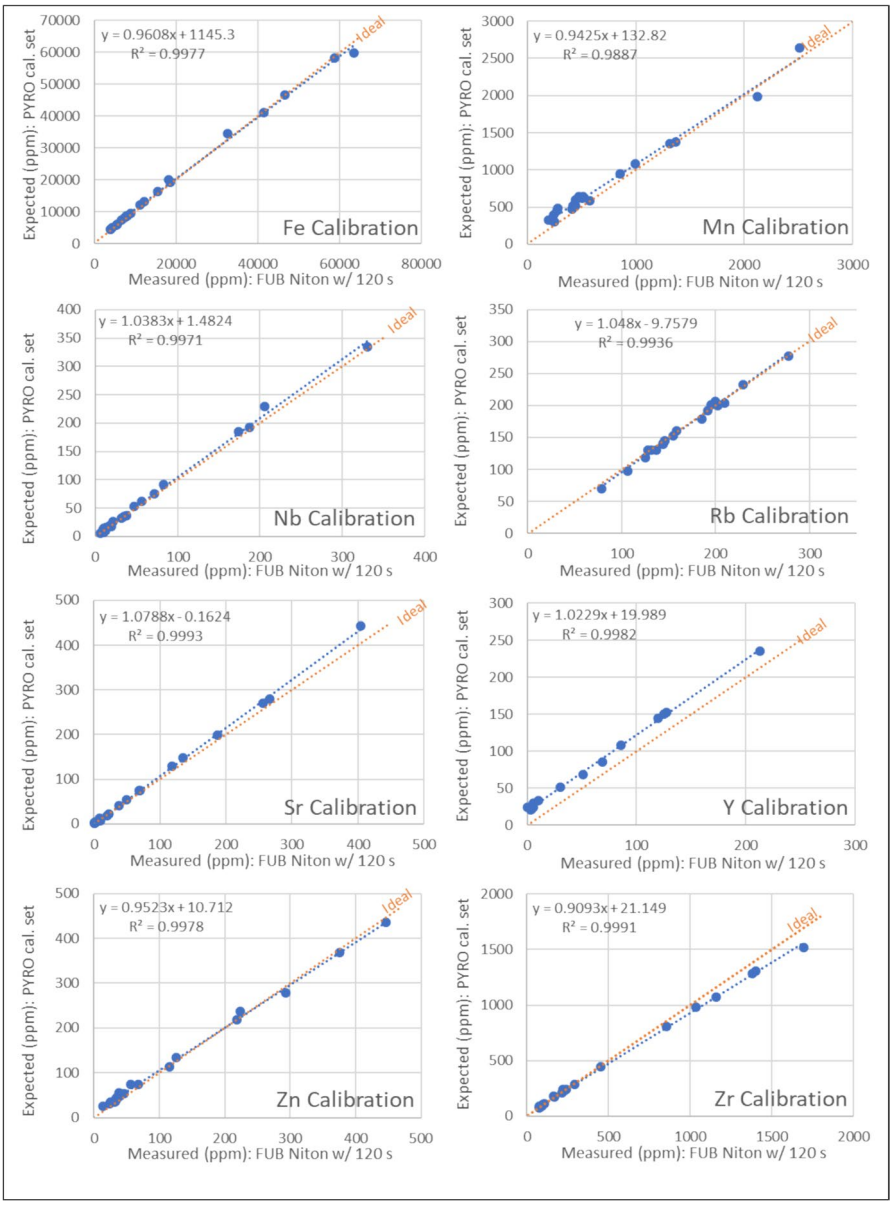

Fig. 4 Calibration charts of the FUB Niton analyzer with the PYRO calibration set for the 120 s measurements. The y-axis refers to the benchmark values of the PYRO calibration set, the x-axis refers to the raw measurements by the FUB Niton analyzer using the default settings, and the ideal trendline refers to the perfectly linear 1:1 relationship between x and y

Kelbadjar third quartile and *Kelbadjar+1 SD* were classified as Syunik Sevkar (Online Resource 3, Column D). The probability of prediction was, in all cases, very high (mostly 100%, the lowest being 73.9%). In only a few cases and with much

Table 4 Evaluation table of the calibrated results for the FUB Niton analyzer. The equations were derived by linear discriminant analysis of calibrated measurement values (x-axis) compared against the expected benchmark values (y-axis). SD(P) stands for standard deviation of a population defined by all eight elements. SDR stands for symmetric difference ratio (for calculation details see the "Symmetric Difference Ratio (SDR) section)

		10 s w/ linear regression for 10 s				120 s w/ linear regression for 120 s				180 s w/ linear regression for 180 s			
		Slope(x)	Intercept	R^2	SDR(%)	Slope(x)	Intercept	R^2	SDR(%)	Slope(x)	Intercept	R^2	SDR(%)
Fe	$y=$	1.016x	−414	0.996	5.64	1.078x	−907	0.999	4.73	1.049x	−672	0.996	4.32
Mn	$y=$	0.940x	+19.4	0.849	3.88	0.955x	+26.4	0.904	2.27	1.070x	−22.3	0.944	5.98
Nb	$y=$	0.961x	−0.51	0.844	10.83	0.947x	+0.13	0.936	8.53	0.982x	−0.36	0.949	5.59
Rb	$y=$	1.064x	−11.3	0.993	3.20	0.987x	+0.89	0.994	1.34	1.032x	−4.68	0.998	1.73
Sr	$y=$	0.973x	+1.19	0.993	1.44	0.959x	+3.32	0.997	2.30	0.965x	+3.22	0.998	2.19
Y	$y=$	0.993x	+1.73	0.987	3.72	1.011x	+1.99	0.987	6.49	1.017x	+1.56	0.993	6.04
Zn	$y=$	0.872x	+2.95	0.932	12.70	0.972x	+0.79	0.976	2.50	1.036x	−2.18	0.988	2.00
Zr	$y=$	0.963x	+0.84	0.998	4.40	0.997x	−6.01	0.998	2.75	0.982x	−3.67	0.998	3.75
Mean		0.973x	−50.1	0.949	5.73	0.988x	−110	0.974	3.86	1.017x	−87.5	0.983	3.95
SD(P)		0.053x	+138	0.063	3.68	0.040x	+301	0.033	2.34	0.035x	+220	0.021	1.70

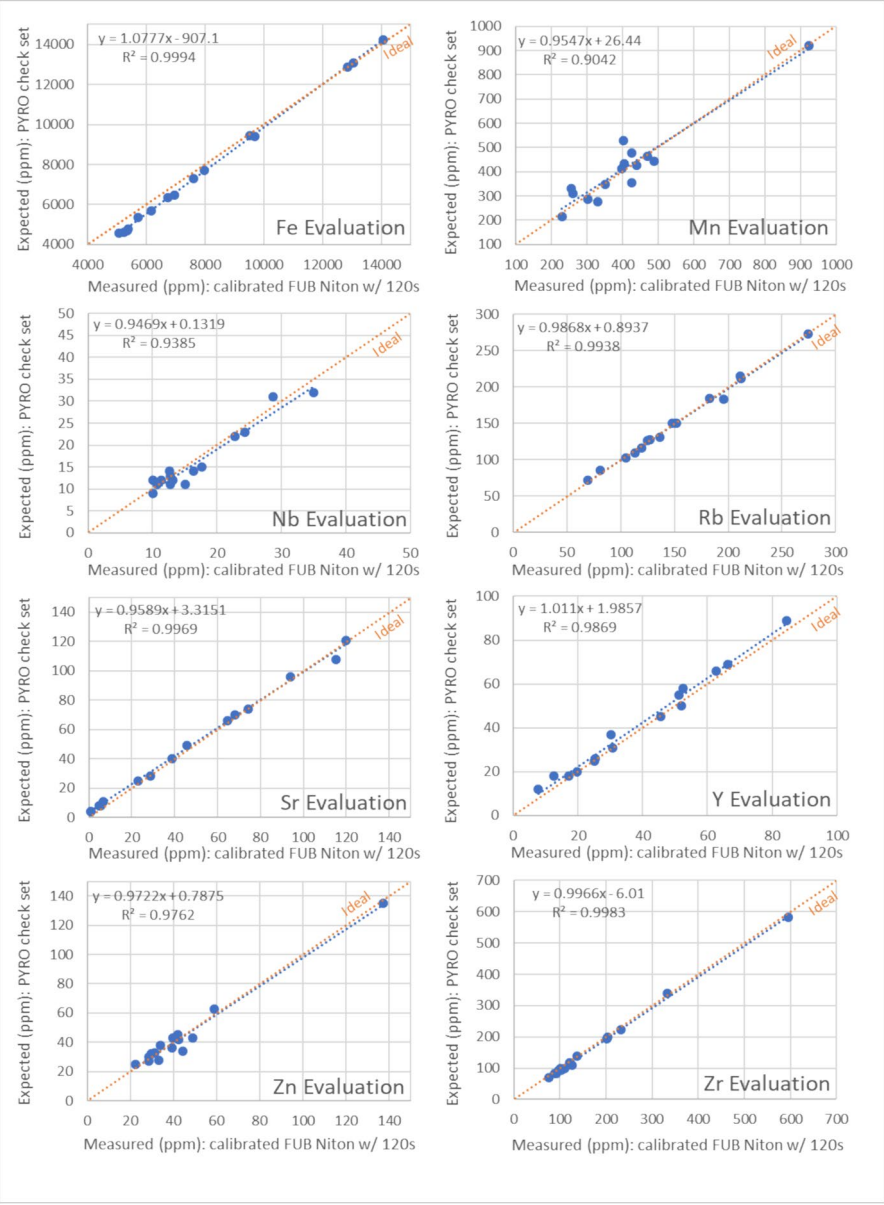

Fig. 5 Evaluation of calibrated measurements for FUB Niton analyzer with 120 s duration of measurements. The y-axis refers to the benchmark values of the PYRO check set, the x-axis refers to the measurements by the FUB Niton analyzer after their calibration, and the ideal trendline refers to the perfectly linear 1:1 relationship between x and y

Table 5 Evaluation table of the experimental sets, in which the measurements of the PYRO check set with lower duration (10 s and 120 s) were calibrated with the linear regression equation derived from the PYRO calibration set with 180 s duration. In the expressions, the calibrated measurement values (on the x-axis) are compared against the expected benchmark values (y-axis). SD(P) stands for standard deviation of a population defined by all eight elements. SDR stands for symmetric difference ration (for calculation details see the "Symmetric Difference Ratio (SDR)" section)

		10 s with 180 s linear regression				120 s with 180 s linear regression			
		Slope(x)	Intercept	R^2	SDR(%)	Slope(x)	Intercept	R^2	SDR(%)
Fe	$y=$	1.035x	−691	0.996	7.78	1.078x	−786	0.999	3.97
Mn	$y=$	0.968x	−3.42	0.849	5.71	0.983x	+6.98	0.904	1.09
Nb	$y=$	0.971x	+0.12	0.844	3.66	0.955x	−0.27	0.939	10.49
Rb	$y=$	1.073x	−11.4	0.993	3.69	1.043x	−6.1	0.994	2.28
Sr	$y=$	0.949x	+2.88	0.993	2.59	0.965x	+3.16	0.997	2.20
Y	$y=$	1.011x	+1.07	0.987	4.28	1.04x	+0.74	0.987	6.92
Zn	$y=$	0.849x	+4.39	0.932	13.38	0.948x	+2.86	0.976	3.03
Zr	$y=$	0.975x	−3.68	0.998	4.64	1.003x	−6.89	0.998	2.23
Mean		0.979x	−87.7	0.949	5.72	1.002x	−98.3	0.974	4.03
SD(P)		0.062x	+228	0.063	3.25	0.044x	+260	0.033	2.94

lower probability (from 10 to 26%), a secondary identification was proposed (*i.e.*, Chikiani 2 for Chikiani 1, Bartsratumb for Kelbadjar/Kechal Dağ, Nemrut Dağ 3 for Nemrut Dağ 4, Kelbadjar and Bartsratumb for Syunik Sevkar, Bartstratumb and Syunik Bazenk for Syunik Satanakar 1).

In open LDA, considering 1% prior probability of "other" sources, three entries were misclassified (*i.e.*, the two Kelbadjar entries were attributed to Syunik, one Nemrut Dağ 5 entry was attributed to an "other" source; Online Resource 3, Column E). When looking at the secondary identifications, apart from the examples noted by standard LDA, one Erzurum West 2 (14% probability) and an additional Nemrut Dağ 5 (10% probability) were classified as "other," yet clearly with a much larger probability of their correct identifications.

Attribution to Geochemical Groups

Algorithmic Classification of Geological Specimens Using LDA and Consensus Values

When the dataset of consensus values is used to define sources, as was our primary concern and assessment configuration, 11 out of the 66 geological field specimens (16.67%) differ regarding their recorded geographic origin and their geochemically predicted obsidian source. Pokr Arteni subsources were not differentiated among our geological specimens, and consequently, their classification as either Pokr Arteni 1 or 2 was considered correct by us, contrary to the algorithm assessments, which treated them as misclassifications due to non-identical names (Pokr Arteni vs. Pokr Arteni 1 or 2). Three of the diverging specimens have, at least, the expected source

as a secondary prediction; however, those secondary predictions have considerably lower probabilities (Table 6, Online Resource 3).

When using open LDA, with prior probability for "other" source set to 1% — against the equally distributed prior probability of 1.71% for any of the 58 obsidian sources represented in the dataset of consensus elemental values — the algorithm classified 12 of 66 geological specimens as coming from an "other" source (Table 7). This translates to "other" source posterior probability of 18%.

When we inverted the procedure and used our geological field specimens as the source-defining set (as an experiment to see if the sources would be correctly predicted), standard LDA classified the entries from the corresponding ten sources present in the check set of consensus values correctly, with the exception of five (of six) Mets Arteni entries (Online Resource 4, column F). Those Mets Arteni entries were misclassified as Pokr Arteni, the immediately adjacent source. Once we removed two of the Mets Arteni field specimens, which were revealed as problematic via internal consistency checks (see the "Internal Consistency Check of Geological Field Specimens" sections), all six Mets Arteni entries among the consensus values were correctly classified, and the correct classification of other entries in the same dataset did not change (Online Resource 4, column G). As expected, all entries for the 48 obsidian sources absent from the defining set (our field specimens) were wrongly attributed — the algorithm forcibly attributed them to one of the ten sources in the defining set (as those are the only ones from which the algorithm could choose).

When we applied open LDA on this inverted procedure, we excluded the two problematic field specimens from Mets Arteni, and we set a prior probability of "other" source to 82.76% (see "Algorithmic Classification of Geological Samples Using LDA with the Consensus Values"). In this scenario, we obtained a more complicated picture: most known sources were correctly identified, some known sources misclassified as "other," some sources were misclassified in the same general region (*e.g.*, Syunik Bazenk and Syunik Sevkar misclassified as Syunik Satanakar 1), unrepresented sources were most often correctly predicted as "other," and six of those unrepresented sources were misclassified as one of the known nearby sources (one Bartstratumb and two Kelbadjar attributed to Syunik Satanakar 1; three Chikiani 2 attributed to Chikiani 1; Online Resource 4, column H).

Algorithmic Classification with SourceXplorer

The results of the SourceXplorer classification with standard settings (LDA probability threshold: 70%, LDA model accuracy threshold: 80%) were not very encouraging. For the "robust" outcome, only a single specimen was attributed to a geological source, and six specimens were classified as ambiguous (Online Resource 5, column V). The remaining specimens were classified as "unknown." With the "standard" outcome (except for the same six specimens robustly classified as ambiguous), three were classified as Gutansar, and all remaining ones as unknown (Online Resource 5, column U). With the "basic" outcome (Online Resource 5, column T), the same six specimens as in the previous two outcomes were classified as ambiguous, while 22 specimens (33%) were attributed to known sources, and 38 specimens (58%) remained

Table 6 List of samples, for which their recorded origin diverges from the source predicted by the algorithmic classification with the Standard LDA

Label	Recorded site	Predicted source	Prediction probability (%)	Secondary predictions	Secondary predictions probability (%)
ALP2.2	Pokr Arteni	Gegham-Spitakasar/Geghasar-2	99.04	-	-
CH1.3	Chikiani 1	Chikiani 2	56.65	Chikiani 1	43.35
CH1.6	Chikiani 1	Chikiani 2	72.72	Chikiani 1	27.28
GS1.3	Gutansar	Erzincan 2	76.98	Hatis Beta	23
MA1.3	Mets Arteni	Pokr Arteni 1	100	-	-
MA1.7	Mets Arteni	Gutansar	100	-	-
MA1.8	Mets Arteni	Pokr Arteni 1	100	-	-
MA1.10	Mets Arteni	Gegham-Spitakasar/Geghasar-2	100	-	-
PA3.2	Pokr Arteni	Gegham-Spitakasar/Geghasar-2	95.31	-	-
PA3.4	Pokr Arteni	Gegham-Spitakasar/Geghasar-2	97.56	-	-
PYRO_Cal04	Syunik Satanakar 1	Syunik Bazenk	59.49	Syunik Satanakar 1	40

Table 7 List of samples classified as coming from "other" source by algorithmic classification with Open LDA. The LDA was trained on the dataset of consensus values with 1% prior probability for a source not present in the dataset

Label	Recorded site	Predicted source	Prediction probability (%)	Secondary predictions (with probabilities)
GF1.3	Gutansar	Other	98.66	Gutansar (1.34%)
GJ1.2	Gutansar	Other	99.88	Gutansar (0.12%)
GS1.2	Gutansar	Other	99.98	Gutansar (0.02%)
GS1.3	Gutansar	Other*	100	
HKKQ2.3	Hatis Beta	Other	94.75	Hatis Beta (5.25%)
HR1.2	Gutansar	Other	97.97	Gutansar (2.03%)
MA1.2	Mets Arteni	Other	100	
MA1.3	Mets Arteni	Other*	99.95	
MA1.5	Mets Arteni	Other	92.81	Mets Arteni (7.19%)
MA1.8	Mets Arteni	Other*	100	
PA3.2	Pokr Arteni	Other*	99.61	
PA3.4	Pokr Arteni	Other*	100	

*Specimens that were classified as known sources by the standard LDA (see Table 6)

classified as unknown (Table 8). Not even all PYRO calibration specimens (which are smooth and flat) were correctly identified: PYRO_Cal04 from Satanakar-1 was classified as ambiguous, whereas PYRO_Cal02 (Gutansar), PYRO_Cal05 (Chikiani 1), and PYRO_Cal07 (Sarıkamış 1) were all classified as unknown sources.

At the very least, the LDA predictions by SourceXplorer correspond 1-to-1 to our standard LDA predictions done with JMP 17.2, which indicates that the default prior settings for this method are identical in both software packages.

Algorithmic Classification with AutoML for Geochemistry

Supervised learning, where the dataset of consensus values served to define classes, against which the geological dataset was subsequently checked, produced the following results. After splitting the dataset of consensus values (80% for training and 20% for testing) and, at the same time, ignoring the upsampling feature, the algorithm automatically chose the k-nearest neighbors model as the best solution (k-nearest neighbors accuracy: 0.93). When experimenting with the same split but with the upsampling feature turned on, the algorithm still chose k-nearest neighbors as the best model, with a slightly better accuracy score than in the previous case (k-nearest neighbors accuracy: 0.97). In both cases, there were few correct source identifications: 23 without upsampling (Online Resource 3, column Q), and 18 with upsampling (Online Resource 3, column R), out of a total of 66.

Unsupervised learning with the dataset of consensus values used the 282 entries where all eight elements were present; consequently, the total number of considered sources dropped from 58 to 56. PCA produced five components with variance larger

Table 8 Geological samples attributed by SourceXplorer to a geochemical source or more sources (ambiguous) with the following settings: included elements (Fe, Nb, Mn, Rb, Sr, Y, Zn, Zr), included sources from the dataset of consensus values ($n=56$), source observations ($n=282$), LDA model accurray (95.24%), LDA probability threshold (70%)

ID	Location	LDA prediction	LDA probability (%)	Provenance summary (basic)	Provenance summary (standard)	Provenance summary (robust)
ALP1.1	Pokr Arteni	Pokr Arteni 2	95.57	Pokr Arteni 2	Unknown	Unknown
ALP2.1	Pokr Arteni	Pokr Arteni 2	99.99	Pokr Arteni 2	Unknown	Unknown
CH1.1	Chikiani 1	Chikiani 1	80.35	Chikiani 1	Unknown	Unknown
CH1.2	Chikiani 1	Chikiani 1	65.2	Ambiguous	Ambiguous	Ambiguous
CH1.3	Chikiani 1	Chikiani 2	53.5	Ambiguous	Ambiguous	Ambiguous
CH1.4	Chikiani 1	Chikiani 1	56.7	Ambiguous	Ambiguous	Ambiguous
CH1.5	Chikiani 1	Chikiani 1	80.36	Chikiani 1	Unknown	Unknown
GF1.1	Gutansar	Gutansar	100	Gutansar	Gutansar	Unknown
GF1.2	Gutansar	Gutansar	100	Gutansar	Gutansar	Unknown
GF1.4	Gutansar	Gutansar	100	Gutansar	Unknown	Unknown
GJ1.1	Gutansar	Gutansar	100	Gutansar	Unknown	Unknown
GJ1.3	Gutansar	Gutansar	100	Gutansar	Unknown	Unknown
GJ1.4	Gutansar	Gutansar	100	Gutansar	Unknown	Unknown
GS1.4	Gutansar	Gutansar	100	Gutansar	Unknown	Unknown
GW1.1	Gutansar	Gutansar	53.14	Ambiguous	Ambiguous	Ambiguous
GW1.2	Gutansar	Gutansar	100	Gutansar	Unknown	Unknown
GW1.3	Gutansar	Gutansar	100	Gutansar	Unknown	Unknown
HR1.1	Gutansar	Gutansar	100	Gutansar	Unknown	Unknown
HR1.2	Gutansar	Gutansar	100	Gutansar	Unknown	Unknown
HR1.4	Gutansar	Gutansar	100	Gutansar	Gutansar	Gutansar
MA1.7	Mets Arteni	Gutansar	100	Gutansar	Unknown	Unknown
PA1.4	Pokr Arteni	Pokr Arteni 2	57.35	Ambiguous	Ambiguous	Ambiguous
PA2.1	Pokr Arteni	Pokr Arteni 1	100	Pokr Arteni 1	Unknown	Unknown
PA2.4	Pokr Arteni	Pokr Arteni 1	93.22	Pokr Arteni 1	Unknown	Unknown
PYRO_Cal01	Aghvorik	Aghvorik	100	Aghvorik	Unknown	Unknown
PYRO_Cal04	Syunik-Satanakar 1	Syunik-Bazenk	60.39	Ambiguous	Ambiguous	Ambiguous
PYRO_Cal06	Meydan Dağ	Meydan Dağ	100	Meydan Dağ	Unknown	Unknown
PYRO_Cal08	Sarıkamış 1	Sarıkamış 1	99.74	Sarıkamış 1	Unknown	Unknown

than 1% (PC1 with 65.23%, PC2 with 15.24%, PC3 with 8.57%, PC4 with 6.37%, PC5 with 3.39%). Based on these five principal components with the most variance, the algorithm divided all entries into only three clusters. According to the generated report, the Elbow method was used to define the number of clusters and the k-means algorithm to generate them. We experimented with using just two principal components with the most variance and with all eight principal components, yet the number of generated clusters was always three.

Unsupervised learning with the dataset of all 66 geological specimens also produced five components with variance larger than 1%. It, too, led to the definition of just three clusters, while cluster #3 only contained a single specimen: Nemrut Dağ 6 (Online Resource 3, Column S). According to the generated report, the Elbow method chose the number of algorithms (3) and the k-means algorithm was used to generate them. Again, the number of clusters did not change in relation to the chosen number of principal components.

Classification By-Eye

Visually, the calibrated values of geological specimens revealed clearly delineated groups in most scatterplots (Figs. 6, 7, 8). Only the Mn vs. Zn scatterplot produced groups that were vague, dispersed, and largely overlapping (Fig. 9). Attributing the clearly delineated groups to the respective obsidian sources was rather straightforward because the symbols of the sources from the geological specimens were, in most instances, very close to the symbols of the sources from the consensus values dataset. Especially instructive are the overlaps on the scatterplot of the two most significant principal components (Fig. 10). As indicated by a detailed comparison with visual inspection (Table 10), there are noticeable shifts for some elements, which may explain why certain semi-automated classification algorithms struggled with correct identifications of the obsidian sources.

Discussion

Calibration Issues

The calibrated measurements exhibit generally good correspondence with the benchmark element values. Nonetheless, some differences were observed between the *PYRO calibration set* and the *check set*, among the series with different measurement times, and among different elements. Although the differences are minor, they show particularities worth considering.

Differences Between the PYRO Calibration and Check Sets

The mean symmetric difference ratios (SDRs) for the *PYRO calibration set* during the calibration are higher than for the *PYRO check set* during the evaluation, which is consistent with what could be reasonably expected (cf. Tables 3, 4). The same is true for the standard deviation for the SDR, which ranges between 4.74 and 5.56 for the calibration set and between 1.70 and 3.68 for the check set (with the lower values reached by the series with the longest measurement duration, as expected). However, there are some notable individual exceptions: the Nb, Fe, and Zn trendlines stray further from the ideal in the *PYRO check set* than they did in the *PYRO calibration set.* Also, contrary to the SDR values, the mean slope and R^2 values — two criteria for evaluating the accuracy of the calibration — appear numerically worse for the *PYRO check set* than for the *PYRO calibration set*.

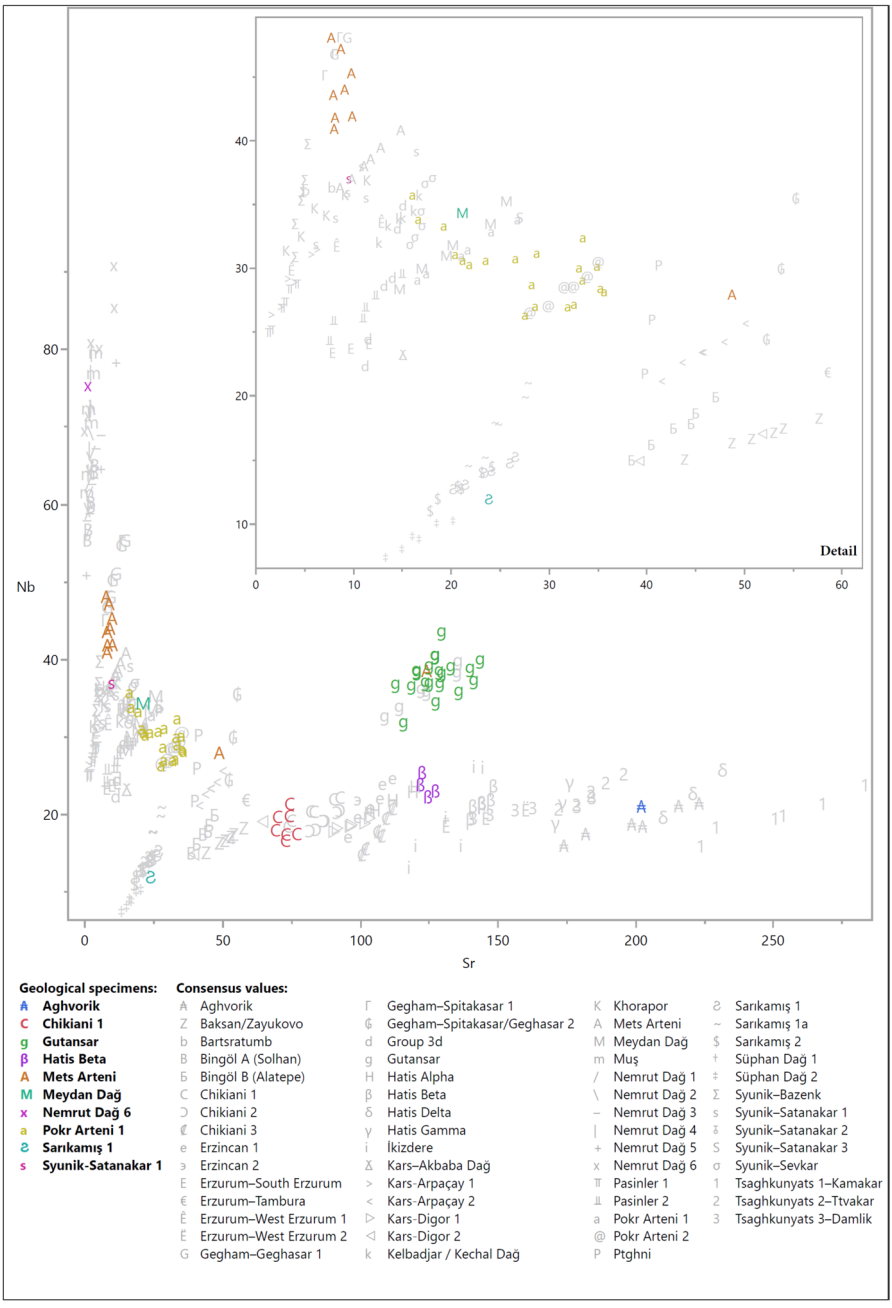

Fig. 6 Scatterplot Nb vs Sr from the classification by-eye. The grey symbols represent the dataset of consensus values, and the colored symbols represent the geological specimens

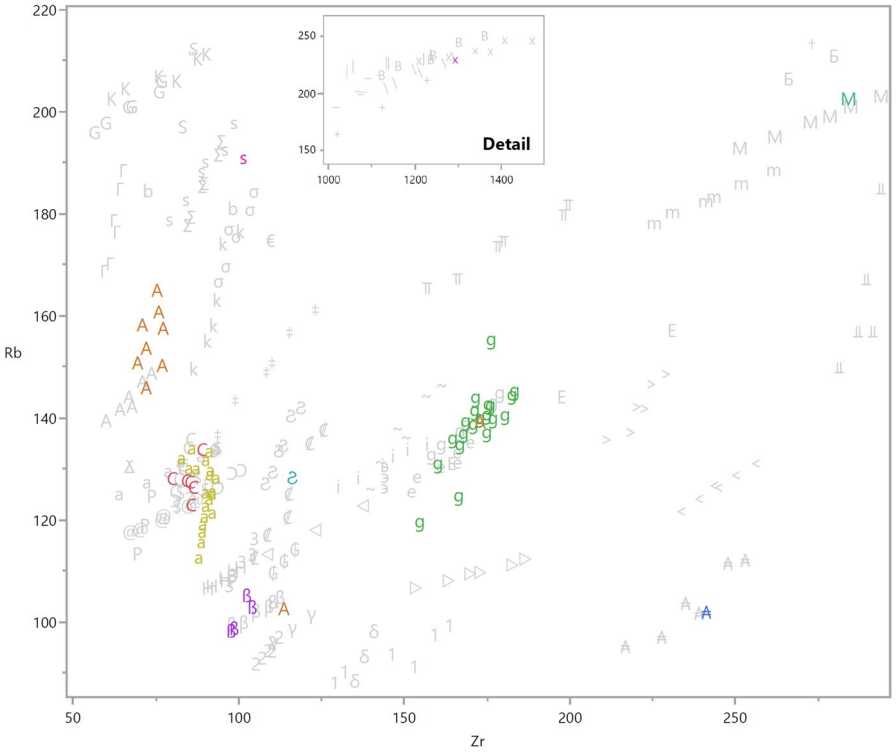

Fig. 7 Scatterplot Rb vs Zr from the classification by-eye. The grey symbols represent the dataset of consensus values, and the colored symbols represent the geological specimens. Symbols for sources are the same as in Fig. 6

However, these apparent contradictions are misleading. Apart from confirming the already mentioned fact that a slope without an intercept does not sufficiently reflect the accuracy of a calibration, we need to consider two more aspects. First, the two *PYRO* sets reflect different stages of the calibration process: (1) calibration of uncalibrated values versus (2) evaluation of calibrated ones. Second and most importantly, the two sets have different ranges — the value range of the *PYRO check set* is much more restrained than the value range of the *PYRO calibration set.* The latter encompasses the extreme values defined by nearly all known rhyolitic obsidian sources worldwide, while the former includes a limited selection of sources from North America. Accordingly, values that appear good and are tightly packed over a large range may appear less good and much further apart when observed within a "zoomed in" smaller range (*e.g.*, Fe). The results would appear to improve if the evaluation specimens were to cover larger sections of the calibration range or its entirety.

Keeping the range scales in mind, the slope and R^2 differences between the *PYRO calibration set* and the *PYRO check set* are much less surprising, and it is expected that the precision would be more limited when compared to the *PYRO calibration set.*

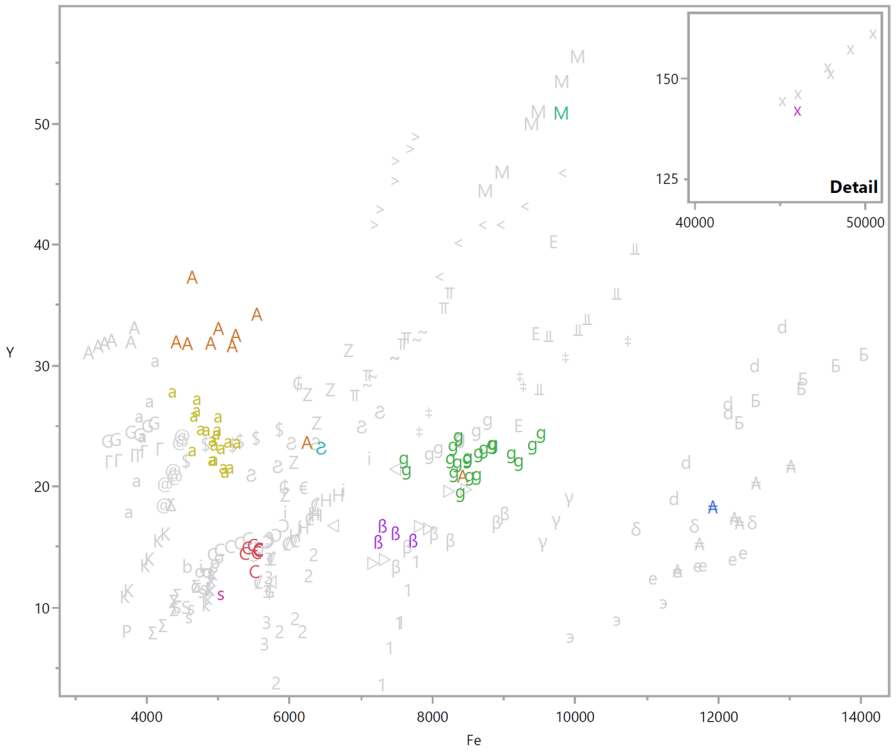

Fig. 8 Scatterplot Y vs Fe from the classification by-eye. The grey symbols represent the dataset of consensus values, and the colored symbols represent the geological specimens. Symbols for sources are the same as in Fig. 6

Differences Among Series with Different Measurement Times

Although this section focuses on evaluating different measurement times for the calibrated results of the PYRO check set, a few initial observations concerning the calibration of raw measurements of the *PYRO calibration set* deserve to be discussed.

Within the *PYRO calibration set*, the differences in mean SDR, slope, and R^2 values among the measurement series with 10, 120, and 180 s duration are generally minimal (cf. Tables 3, 4). The mean SDR values range from 5.98% (10 s) to 6.52% (120 s) — ca. 0.5% difference — which means that three series, each with different durations, require very similar calibrations. The mean slope values range from 0.981 (180 s) to 0.994 (120 s), although the individual values range more broadly (0.903 to 1.079). The mean R^2 ranges from 0.995 (180 s) to 0.997 (10 s). Surprisingly, the calibration of the 10 s measurement series produced exceptionally good slope and R^2 values. Its mean R^2 of 0.997 was the highest among all three series. Combined with the lowest standard deviation of 0.003, this attests to consistent measurements of the FUB Niton analyzer despite the short 10 s measurement duration. Even the standard deviation of the slope values

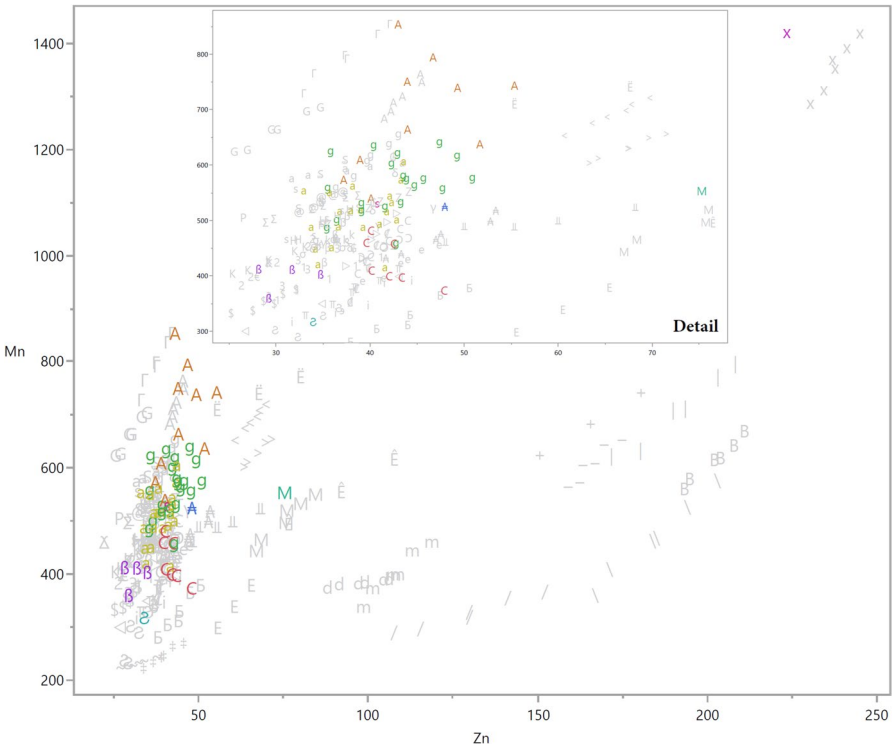

Fig. 9 Scatterplot Mn vs Zn, the only combination that produced rather vague, dispersed, and largely overlapping groups during the classification by-eye. The grey symbols represent the dataset of consensus values, and the colored symbols represent the geological specimens. Symbols for sources are the same as in Fig. 6

in the 10 s series was the lowest (0.044). Similarly surprising, the 120 s series reached slope (0.994) and R^2 (0.996) values better aligned with the ideal than the 180 s series. Based on counting statistics alone, one would expect the 10 s series to have the greatest scatter around an ideal trendline and that such scatter would decrease with more X-ray counts measured during the 120 s series and even more with the 180 s series. In fact, the 180 s series has, it would appear, no better precision than the 120 s series. Although it cannot be excluded that this has something to do with using only one voltage/beam filter setting in the 10-s measurements, as opposed to all filters in the 120/180-s measurements, we suggest that any advantage from greater measurement durations had statistically plateaued long before 180 s.

Compared to the *PYRO calibration set*, calibrations with different measurement times exhibited greater differences for the *PYRO check set*. The mean SDR values range from 3.86 (120 s) to 5.73 (10 s), while the individual SDR for elements varies between 1.34 and 12.70%. The mean slope values range from 0.973 (10 s) to 1.017 (180 s), while the R^2 ranges from 0.949 (10 s) and 0.983 (180 s). While there is a

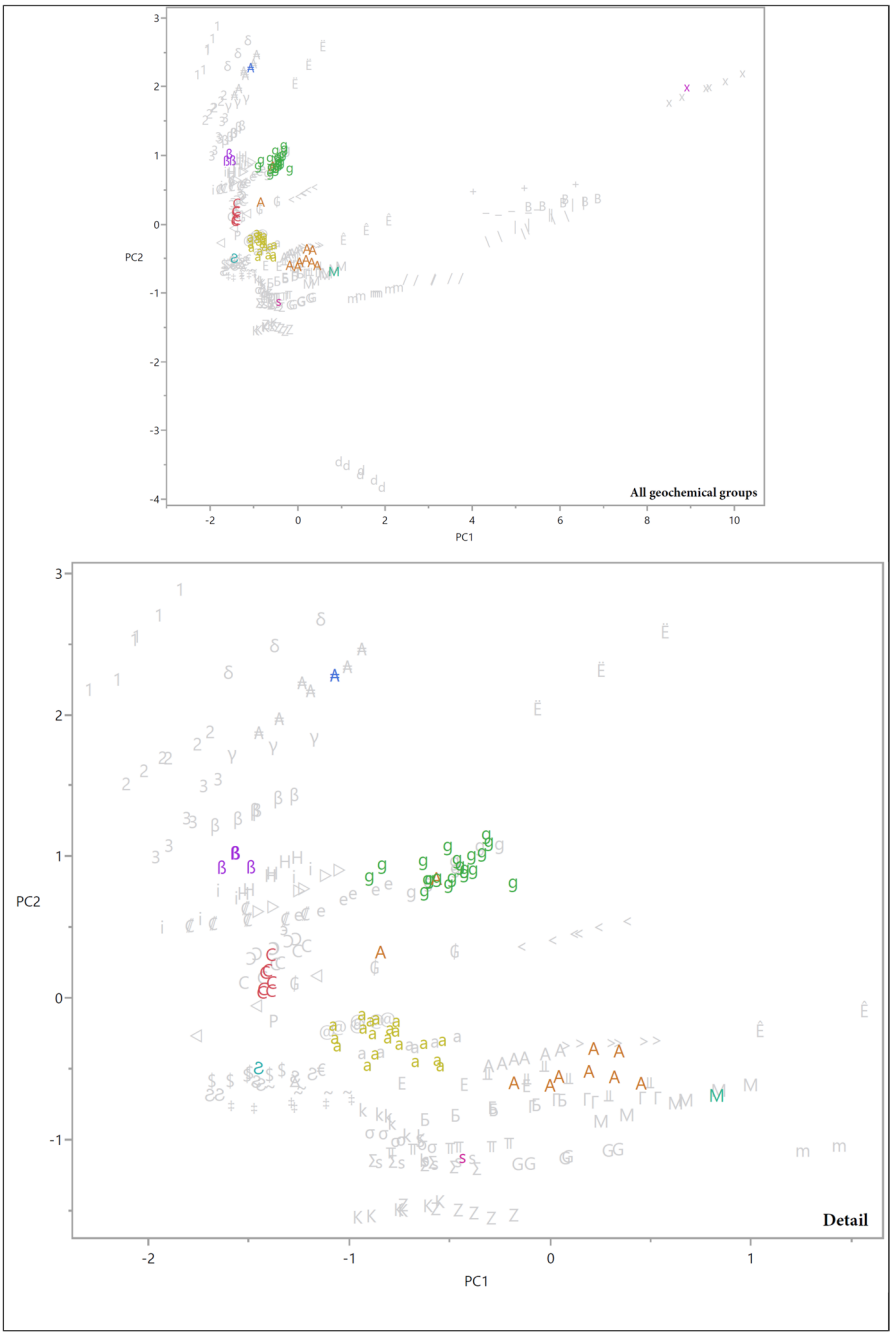

Fig. 10 Scatterplot of PC2 vs. PC1, the first two components of PCA based on all eight elements. The upper chart shows the computed proximity/distance among all geochemical groups represented in the consensus values dataset. The lower chart shows in more detail the densest area with overlaps. Symbols for sources are the same as in Fig. 6

tendency for the mean SDR, slope, and R^2 values to improve with increased duration (120 s and 180 s values are better than the 10 s values), again the mean 120 s values are slightly better than the 180 s ones (see Table 4 vs Table 3). It needs to be considered, however, that the individual values summarized by means vary. The level of variation is correlated with the duration of the measurements – the shorter the duration, the more variation is expressed by the R^2 (best exemplified on Mn and Nb, see Fig. 11) and standard deviation (=*SD(P)*). For example, two elements in the 10 s measurement series of the PYRO check set, Nb and Zn, produced SDR ratios higher than 10%, which explains why the standard deviation for this series was the highest one. The evaluation of all elements with respect to measurement duration is presented in the "Differences Among Elements" section.

When comparing the relative position of individual specimens on two- or three-dimensional plots for the published values of the *PYRO calibration set* and our 10, 120, and 180 s measurement series, we observed smaller differences. Although not substantial, these differences lead to positions that are offset in comparison to the consensus values. As can be observed in Fig. 12, the offsets among these series are not consistent in magnitude or direction. Because these offsets vary in different directions within a set, they signal that the relative positions might be potentially misleading when performing a classification by-eye on artifacts from unknown sources. However tempting a comparison of relative specimen positions on plots may sound in terms of chemical groupings, it cannot serve as a reliable guide when comparing results obtained with different measurement durations.

All in all, when considering the mean results and their standard deviations, the 10 s series can be considered the worst, while the 180 s series is the best calibrated in absolute terms. Nonetheless, the differences between 180 and 120 s series are so slight — the mean symmetric difference ratio is worse by just 0.09%, while the measurement time is 33% shorter — that, in terms of a combined evaluation, we consider the 120 s duration to be the best trade-off between measurement time and precision for the FUB Niton analyzer. The R^2 value of 0.974 for the 120 s series is still very good — sufficient to expect consistent and comparable results. It should be noted that time is not the primary factor here — it depends on the number of X-ray counts measured by an instrument, so with a faster instrument, the same number of counts will be obtained in a shorter time.

Experimental Calibration of 10 s and 120 s Measurements with Regression Equations Derived from 180 s Measurements

Most elements calibrated with regression equations from a different measurement duration yield worse results than with calibrations derived from measurements with the same duration (see Table 5 vs. Table 4). There was one exception: in the case of the 10 s series, the mean SDR was one-hundredth of one percent better with linear regression derived from 180 s duration series than with linear regression derived from the 10 s series (5.72% compared to 5.73% — for all purposes, identical). This result largely comes down to the more accurate Nb values, and all other elements finished with scores worse than those calibrated with equations based on the 10 s series. In the 120 s series calibrated with regression derived

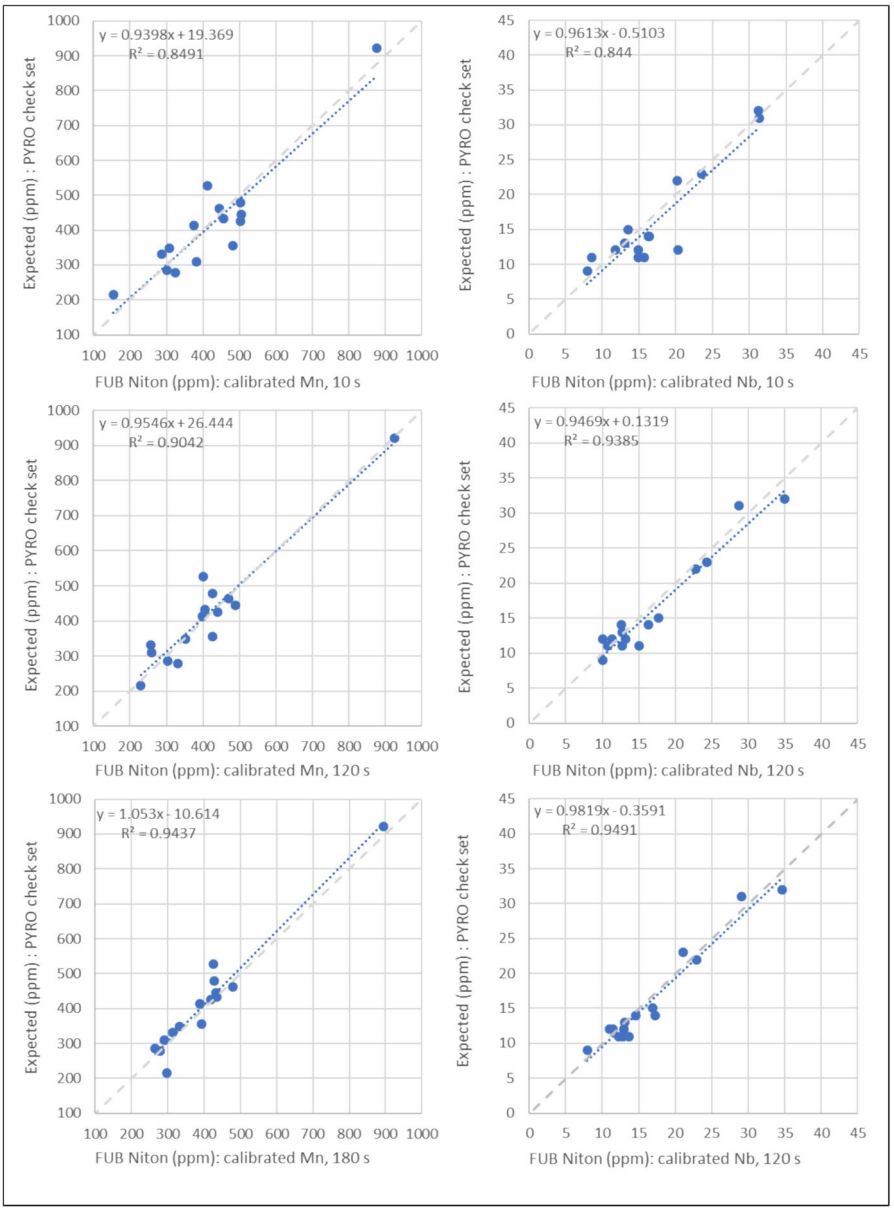

Fig. 11 Comparison of calibration evaluations with 10 s, 120 s, and 180 s measurement series for two elements with the lowest R^2: Mn (left) and Nb (right). Both elements show variance diminishing with the increasing measurement duration = individual samples are closer to trendline and thus more precise

from the 180 s series, the individual elements were more mixed. Yet the mean result of 4.03% is only slightly worse than 3.86% obtained via regression derived from the 120 s series.

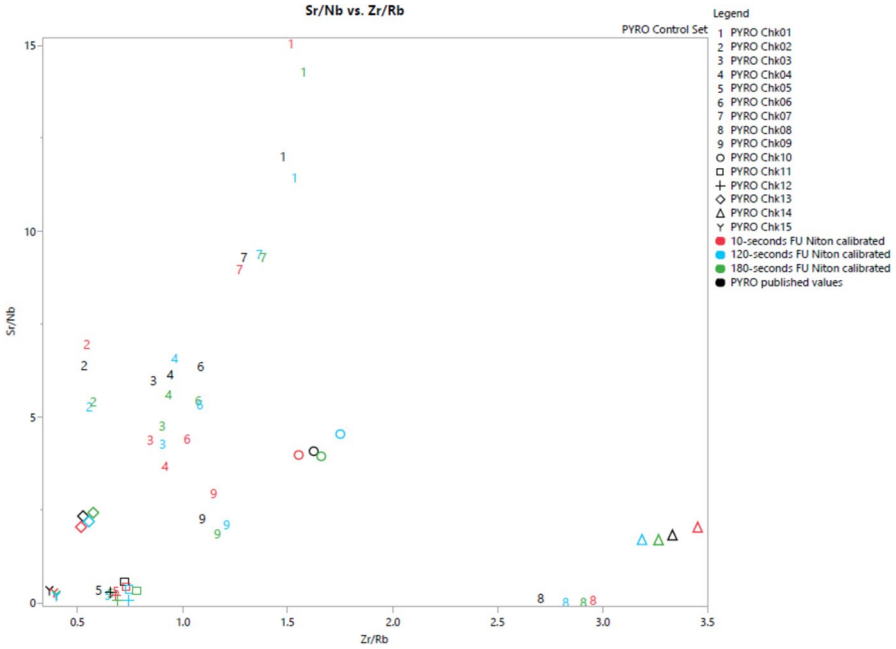

Fig. 12 PYRO check set — comparison of the published benchmark values with the calibrated values of Sr/Nb vs Zr/Rb for 10 s, 120 s, and 180 s duration measurements with the FUB Niton analyzer. Generally, the 180s values (green) are the closest and the 10s values (red) are the most far-off from the published benchmark values (black). Yet the offsets are not standard neither in their magnitude nor in their direction

Differences Among Elements

The necessity of adjusting the factory settings for the FUB Niton analyzer via added calibration depends very much on the element. On the one hand, the uncalibrated values for Fe, Rb, and Zn were so close to the benchmark values of the *PYRO* sets that their calibration seemed nearly redundant. On the other hand, the Mn, Nb, Sr, and Zr needed slight adjustments, whereas Y demonstrated a notable shift that required a significant correction. Because other projects may use other instruments by the same or a different manufacturer, these results confirm that custom instrument calibration continues to be essential to obtain intercomparable results.

Once the FUB Niton analyzer measurements were calibrated, all eight elements reported for the *PYRO* returned values comparable to the benchmark values. However, some are more reliable than others (Table 9). Those values combining high accuracy and precision are reliable. The others are less reliable, and added caution should be exercised, especially when grouping and interpreting results lacking in both precision and accuracy.

The calibration of Y demonstrated several particularities. It has the highest symmetric difference ratio values of all eight elements for the *PYRO calibration set.* Additionally, it is the only element that returned negative measurement values.

Table 9 Evaluation overview of the calibrated values of the PYRO check set measured by the FUB Niton analyzer. High accuracy is attested for SDR ≤ 5%, high precision is attested for $R^2 \geq 0.98$. For details see Table 4

	10 s duration measurements		120 s duration measurements		180 s duration measurements	
	High accuracy	High precision	High accuracy	High precision	High accuracy	High precision
Fe	-	+	+	+	+	+
Mn	+	-	+	-	-	-
Nb	-	-	-	-	-	-
Rb	+	+	+	+	+	+
Sr	+	+	+	+	+	+
Y	+	+	-	+	-	+
Zn	-	-	+	-	+	+
Zr	+	+	+	+	+	+

Yet, the nearly parallel position of its calibration trendline clearly indicates that this peculiarity must be due to a shifted factory setting for Y. Because the slope of the calibration trendline is, in all measurement series, very close to 1, nearly the entire symmetric difference is due to a systematic upward shift. The evaluation with the *PYRO check set* demonstrates that the shift is constant and, in turn, can be corrected. Accordingly, even the negative values — once calibrated — can be trusted, although a slight upward shift remained even after calibration.

Geological Specimen Issues

An internal consistency check based on LDA revealed that the geochemical sources of specimens MA1.7 and MA1.10 cannot be their recorded geographic findspots of Mets Arteni. This was confirmed by classification by-eye and algorithmic classifications identifying MA1.7 as Gutansar and MA1.10 as "Gegham/Spitakasar or Geghasar 2." Any laboratory issue or sampling switch between MA1.7 and MA1.10 is doubtful, if not outright excluded. Because both of these specimens were collected from a surface scatter on the slope of Mets Arteni volcano, not a geological outcrop, these outliers indicate that their field collection was the problem. Although eight of the collected obsidians from the same surface collection are indeed consistent with the published consensus values for Mets Arteni, the two outliers represent anthropogenic contamination. They must have been brought to the sampling locus from elsewhere. Most likely, what we collected was a scatter of archaeological origin. Since Gutansar is a clearly defined geochemical group, distinct from all the others, its presence on Mets Arteni cannot be explained in terms of a geochemical overlap. Although the "Gegham/Spitakasar or Geghasar 2" group might, in certain plots, geochemically overlap with Mets Arteni (see "Algorithmic Classification of Geological Specimens Using LDA and Consensus Values"), in classification by-eye they can be clearly separated (see "Classification with SourceXplorer"). Moreover, this group is attested already from palaeolithic sites. An archaeological scatter is consistent with the hill called Satani

Dar/Tapak Bloor, which lies next to the Mets Arteni peak, and which had once been erroneously interpreted as a separate obsidian source of the Arteni complex (Frahm, 2023a, 9). Similarly, Mount Ararat and Tendürek Dağ in Turkey were previously reported in the literature as obsidian sources, although no geological outcrops were ever found there — visitors had instead come across archaeological scatters (Frahm, 2023a, 20). Such ambiguities exemplify the danger of collecting scattered obsidians simply from the surface instead of locating the geological outcrops and sampling them directly.

Algorithmic Classification of Geological Specimens Using LDA and Consensus Values

Generally, when using the consensus elemental values as the defining dataset, the results of the algorithmic classification with standard LDA seem to imply several geochemical overlaps between a few sources (Table 6). Some of the noted overlaps are nearby sources or fall within the same source complex (PYRO Cal-04 from Syunik-Satanakar 1 predicted as Syunik-Bazenk; GS1.3 from Gutansar predicted as Hatis beta; two likely Chikiani 1 specimens predicted as Chikiani 2, two Mets Arteni predicted as Pokr Arteni), yet some others are between different geographic areas (three Pokr Arteni specimens and one Mets Arteni classified as "Gegham/Spitakasar or Geghasar 2"). While the former overlaps — although not optimal from the sourcing perspective — are archaeologically of little relevance, the latter is either more concerning or is perhaps revealing a key to the actual origin of that as-yet unclear obsidian source (*e.g.*, is it really part of the Arteni complex?). Algorithmic classification alone cannot reveal the reasons behind geochemical overlaps — it is necessary to consider the classification by-eye and look in detail at individual elements behind each classification.

For the inverted configuration — when geological sources were used as the defining dataset and the consensus values as a check set for standard LDA (especially after eliminating the two problematic specimens collected at Mets Arteni) — there are no overlaps. This attests that the calibrated values of the FUB Niton analyzer are accurate, and the geochemical definitions produced using the geological specimens always fall within the consensus values. Thus, the overlaps are almost surely caused by the consensus values, which are much broader due to reflecting multiple analytical methods.

Concerning the results of open LDA, which classified 24.24% of geological specimens as coming from "other" sources, it appears that geochemical groupings based on the consensus elemental values correspond to our geological specimens much less stringently than hoped. Again, this corresponds to the conclusion from the previous paragraph, meaning that it is still necessary to undertake a detailed analysis of elemental scatterplots.

The results of open LDA with the inverted configuration are less satisfactory because only 10 out of 58 obsidian sources are present in the defining set. Moreover, the geological sources were measured exclusively with a single method (pXRF) and a single instrument (FUB Niton analyzer), so they are more constrained than

the dataset of consensus values, which reflects multiple measurement methods by different laboratories. Thus, it is not surprising that, with the chosen very high prior probability of "other" sources, some known sources were cautiously classified as "other."

Classification with SourceXplorer

Although the low success rate (33%) of SourceXplorer — even in its most liberal "basic" mode — when classifying our geological specimens (using the consensus elemental values) seems at first sight disappointing, we need to realize that the tool was intended for comparison of datasets for both the sources and the unknowns measured using the very same instrument. This was not our case because we only measured the dataset of "unknowns." The source dataset, though, summarized both published and unpublished values, which were obtained using a wide variety of instruments and techniques and were subsequently averaged. The underwhelming performance with our dataset is unlikely the fault of the instrument or software. Rather it needs to be explained by the character of the consensus values, whose ranges of key elements will be broader due to their collection from different instruments and methods. Although the accuracy of the consensus values is arguably high, their overall precision is worse compared to what direct measurements of all geological sources with the FUB Niton analyzer would look like.

It also needs to be stressed that the SourceXplorer identifications never contradicted our previous consistency checks or predictions based on LDA, so it proved useful as a conservative confirmation of predictions obtained by less elaborate methods.

Classification with AutoML for Geochemistry

The unsupervised learning based on the dataset of consensus values identified three clusters. This is a very different result from the total number of 56 obsidian sources in the dataset. Given that some of the sources are represented by only one, two, or three entries, it seems unlikely that any algorithm would correctly classify all the sources as separate clusters. However, reducing the total number of clusters to just three is a particularly disappointing result. Moreover, the limits between those clusters look very deliberate and mechanical. Generally, they seem unconvincing upon a visual check — they occur right in the middle of the respective PCA groups, without any space separating them. The same disappointing result of only three clusters was also produced by unsupervised learning for the dataset of geological specimens.

Particularly concerning are the results of the AutoML supervised classification, which attributed all of the specimens to sources. However, only about one-third of these attributions were correct. The remaining specimens were attributed to incorrect sources, with no categories to distinguish between "ambiguous" or "unknown" sources.

Classification By-Eye

Traditional classification by-eye has proven to currently be the most powerful tool, which can both enhance and be enhanced when combined with PCA and LDA. On one hand, it enables us to understand better what is behind certain algorithmic classifications and thus check their results. On the other hand, the algorithmic classifications can pre-classify the specimens for classification by-eye and post-process problematic identifications. Our classifications yielded very clear results, which can be differentiated and further analyzed on the level of each element and combinations of elements. Generally, the calibrated FUB Niton analyzer values compare well with the published consensus values — only minor issues in terms of accuracy and precision were observed. The lightest elements in the set had the most issues (*i.e.*, Mn, Fe, Zn). The takeaways from the classification by-eye can be summarized into three points:

1) All geological specimens were correctly identified at the geographic source level, although some of them only thanks to visual inspections that revealed shifts exhibited in certain scatterplots. The shifts were relatively minor and never systematic. On the contrary, they seemed to be of varying magnitude and limited to certain geochemical groups. These minor shifts were mostly caused by differences between the pXRF method and the statistical means of calculating consensus values, which depended on the prevalent analytical techniques, the number of specimens used for source definitions, and other variables. A summary of our observations concerning every group of geological specimens, including their fits and shifts, for the most significant scatterplots is offered in Table 10.

2) Some outcrops within the same geographic source area are challenging to discern geochemically when relying on the consensus elemental values (*e.g.,* Chikiani 1 and 2, Hatis alpha and beta, Sarıkamış 1 and 2, Syunik-Bazenk and Syunik-Satanakar 1, Pokr Arteni 1 and 2). Although Hatis can be correctly identified as a source with classification by-eye, it would be difficult to decide whether our Hatis specimens belong to the Hatis alpha or Hatis beta geochemical group. Even on a three-dimensional plot of the first three principal components or of Sr/Nb vs Zr/Rb vs Fe/Y, our geological specimens fall between the Hatis alpha and beta groups of consensus values. Only with the help of an additional LDA, where consensus values are placed together with the geological specimens, are these geological specimens unequivocally linked to Hatis beta. Similar problems arose with our geological specimens from Chikiani — the classification by-eye had difficulties in distinguishing between Chikiani 1 and Chikiani 2 as defined by the consensus values. When inspecting the finds on a plot of the first three principal components, the geological specimens overlap with Chikiani 1, although they are still rather close to Chikiani 2. Surprisingly, with an LDA limited to Chikiani specimens and trained with Chikiani consensus values, the geological specimens were classified as Chikiani 2 with 100% probability. The Sarıkamış 1 specimen from PYRO also poses difficulties for the geochemical groups Sarıkamış 1 and Sarıkamış 2. Although the differentiation difficulties are less severe in this example, the classification by-eye to either the first or second Sarıkamış outcrop would leave some doubts. Lastly, in the case of Pokr Arteni, a distinction between

Table 10 Classification by-eye: summary showing how the groups of geological samples are positioned in respect to the consensus values on respective scatterplots. Fits, overlaps, and shifts refer to ellipses with coverage of 95% defined around respective sources from the dataset of consensus values

Source ($n=66$)	Rb vs. Zr	Nb vs. Sr	Y vs. Fe	Zn vs. Mn	PC2 vs. PC1
Aghvorik ($n=1$)	Fit	Slight shift of 2 ppm towards higher Nb	Slight shift of 2 ppm towards higher Y; close to Hatis delta	Slight shift of c. 5 ppm towards lower Zn; overlap with Hatis gamma	Fit
Chikiani 1 ($n=6$)	Partial fit, overlap with Chikiani 2, Pokr Arteni 1, Sarıkamış 2; slightly different orientation	Partial fit, different orientation; close to Chikiani 2 (without overlap)	Slight shift of 2 ppm towards lower Y and subtly higher Fe, overlap with Chikiani 2 and Ikizdere	Partial fit, strays much further, overlaps with Aghvorik, Chikiani 2 and 3, Bingöl B, Erzincan 1, Ikizdere, Kars Digor 1 and 2, Pasinler 1 and 2	Fit, yet different orientation, overlap with Chikiani 2, Kars Digor 2
Gutansar ($n=20$)	Fit; overlap with Erzincan 1	Fit, yet pXRF strays more	Partial fit, partial diagonal shift to slightly lower values; strays more; overlap with Kars Digor 1, Pasinler 2	Fit, but strays much further, overlap with Baksan Zayukovo, Chikiani 1, 2 and 3, Gegham-Spitakasar/Geghasar 2, Hatis alpha, beta, gamma, and delta, Kars Digor 1, Pasinler 2, Pokr Arteni 1 and 2, Syunik Bazenk, Satanakar 1, 2, and 3	Fit, yet strays further, overlap with Erzincan 1
Hatis beta ($n=4$)	Fit; touch with lower limit of Chikiani 3	Diagonal shift of c. 10 ppm towards lower Sr and 3–5 ppm towards higher Nb; overlap with Hatis alpha and Ikizdere	Partial fit, partial diagonal shift towards slightly higher Y and Fe, overlap with Kars Digor 1	Partial fit, partial notable shift towards lower Zn and lower Mn; overlap with Erzurum Tambura, Khorapor, Tsaghkunyats 1, 2, 3, Sarıkamış 2	Slight shift towards lower PC2, nearly touching Hatis alpha

Table 10 (continued)

Source ($n=66$)	Rb vs. Zr	Nb vs. Sr	Y vs. Fe	Zn vs. Mn	PC2 vs. PC1
Mets Arteni ($n=8$)	Partial fit, slight shift towards higher Rb	Shift of c. 5 ppm towards higher Nb, overlap with Gegham-Spitakasar 1 and Gegham-Spitakasar/Geghasar 2	Shift of 1500–2000 ppm towards higher Fe values	Fit, but strays much further; overlap with Erzurum West 2, Gutansar, Hatis delta, Pokr Arteni 1 and 2, Syunik Bazenk, Satanakar 1, 2 and 3	Slight shift towards lower PC2, overlap with Erzurum South, Kars Arpaçay 1, Pasinler 2
Meydan Dağ ($n=1$)	Fit, minimally offset towards higher Rb (for the given Zr)	Shift of c. 2 ppm towards higher Nb	Almost fit, very slight offset	Slight shift of c. 10 ppm towards lower Zn, overlap with Pasinler 2	Fit, yet close to Gegham-Spitakasar 2 and Pasinler 2
Nemrut Dağ 6 ($n=1$)	Fit; overlap with Nemrut Dağ 2	Fit, overlap with Muş and Nemrut Dağ 4	Almost fit, very slight offset	Diagonal shift towards lower Zn and higher Mn	Fit
Pokr Arteni 1 or 2 ($n=20$)	Partial fit; perpendicular orientation cutting both Pokr Arteni groups; overlap with Chikiani 1, Chikiani 2, Sarıkamış 2, Süphan Dağ 2	Partial fit, overlap between both Pokr Arteni groups, Meydan Dağ, Syunik-Sevkar, Kelbadjar, Group 3d	Fit with Pokr Arteni 1, slight overlap with Pokr Arteni 2, overlap with Sarıkamış 2	Fit, but strays much further and orientation differs from Pokr Arteni 1 and 2; overlap with Gutansar, Hatis delta, Mets Arteni, Syunik Bazenk, Satanakar 1, 2, 3	Fit, tiny overlap with Pasinler 2
Sarıkamış 1 ($n=1$)	Diagonal shift of c. 10 ppm towards higher Zr	Shift of c. 2 ppm towards lower Nb	Fit, overlap with Baksan-Zayukovo	Fit, overlap with Ikizdere	Slight shift towards higher PC2, overlap with Sarıkamış 2
Syunik Satanakar 1 ($n=1$)	Diagonal shift of c. 10 ppm towards higher Zr; proximity to both Syunik Bazenk and Satanakar 1	Fit; overlap with Khorapor and Mets Arteni	Slight diagonal shift, closer to Kelbadjar	Slight shift toward higher Zn, overlap with Baksan Zayukovo, Hatis delta, Gegham-Spitakasar/Geghasar 2	Fit

Table 10 (continued)

Source ($n=66$)	Rb vs. Zr	Nb vs. Sr	Y vs. Fe	Zn vs. Mn	PC2 vs. PC1
Outliers ($n=2$)	- MA1.7 slight shift from Gutansar; - MA1.10 closest to (list in sequence of proximity): Hatis beta, then Hatis Gamma, then Gegham-Spitakasar/Geghasar 2	- MA1.7 slight shift from Gutansar; - MA1.10 slight shift from Gegham-Spitakasar/Geghasar 2, and Kars-Arpaçay 2	- MA1.7 overlap with Kars-Digor 1, shift from Gutansar (yet within limits of Gutansar geol. specimens); - MA1.10 overlap with Baksan/Zayukovo, Gegham-Spitakasar/Geghasar 2, and Sarıkamış 1	- MA1.7 overlap with Gutansar - MA1.10 overlap with Hatis alpha, Gegham-Spitakasar/Geghasar 2, Syunik-Satanakar 1	MA1.7 overlap with Gutansar; MA1.10 very close to Gegham-Spitakasar/Geghasar 2

Pokr Arteni 1 and 2 proved impossible with most of our geological specimens, except for specimens from the perlite matrix that tend to match Pokr Arteni 2.

3) Two specimens collected from the surface scatter on Mets Arteni stand out on all elemental and PCA scatterplots and due to their overlaps can be convincingly attributed to two different geochemical sources: MA1.7 to Gutansar and MA1.10 to Gegham-Spitakasar/Geghasar 2.

Conclusions

Our feasibility study demonstrates that combining two "open sourcing" tools — the PYRO sets and a database of consensus elemental values — can lead to convincing, comprehensive, and reproducible results. Fine calibration continues to be a necessity with pXRF instruments — no different than other analytical techniques — yet it can be adequately resolved using the PYRO sets. The consensus values dataset still poses minor challenges with geochemically overlapping groups, yet it proved accurate for the locations we tested. Not only were all geological specimens correctly classified to their respective sources, but intrusive archaeological artifacts, present in surface scatters of sampling sites, were also recognized and their sources identified. As an unintended secondary result, we may have identified an archaeological site or deposit on the western slopes of Mets Arteni volcano.

We conclude, however, that convincing identifications with these tools were only obtained using classification by-eye. The algorithmic classification methods and online machine-learning methods that we tested are, in combination with the consensus elemental values, either too conservative, which leads to too few identifications (SourceXplorer), or overly optimistic, which leads to numerous false identifications (AutoML for Geochemistry). The differences in performance between classification by-eye and by these algorithms are due to the human observer's ability to account for small shifts between the dataset of consensus values and the calibrated pXRF measurements. Algorithms, on the other hand, strictly apply such a comparison, without considering such shifts. These small shifts are most likely caused by different measurement methods — the consensus values were averaged from multiple analytical techniques and laboratories, whereas our pXRF measurements were produced by a single instrument. Consequently, the averaged values may be "bent" towards a dominant technique or calibration protocol used to characterize each geological source. Moreover, because of their averaging, the consensus values will be more concentrated and will not scatter as far as the complete measurement series on which they are based. In comparison, individual pXRF measurements will, by default, tend to scatter more than the multiple averaged measurements calculated from other techniques.

It should be stressed that our assessment was neither designed nor intended to address the overall capacity of machine-learning tools to yield convincing sourcing results when comparing known and unknown specimens measured with the same instrument. Nor do we dispute that direct measurements of a geological reference collection and the unknown obsidian artifacts using the same instrument can lead to more precise geochemical groupings and, thus, to less ambivalent results than comparing the unknowns with a dataset of consensus elemental values. One

key question was whether possessing a complete reference collection with all 58 geochemical sources of the eastern Turkey-South Caucasus region is an absolute requirement and a workable solution for most archaeological projects. The few specialized laboratories, which could eventually possess all comparative specimens, cannot handle the sheer number of archaeological artifacts that are excavated annually and that could instead be quickly tested and identified with pXRF using an "open sourcing" workflow. Such a workflow (see a graphical representation in Fig. 13), as we have discussed, has low costs and little need of extensive geological collections, yet the need for specialized knowledge remains as it is still essential for successful calibration and correct attribution of the artifacts to obsidian sources.

At the very least, we propose that the use of consensus elemental values can largely alleviate the requirement for direct measurements of all geological source specimens. Targeted analyses of source specimens remain most necessary for potentially overlapping chemical groups. A way to address this issue — and to further improve the "open workflow" — would be to circulate localized sets (perhaps as supplements to PYRO), containing the specimens from any geochemically overlapping groups in each macro-region. Indeed, the future outlooks are positive. The increasing precision of pXRF instruments and numbers of obsidian analyses will enable the corpus of consensus values to substantially grow in quantity and quality, so eventual updates will be even more accurate and better adapted for use with pXRF (or other future analytical tools), perhaps eventually eliminating the need for investigators to have their own exhaustive reference collections altogether. Ideally, the entire calibration process would be fully automated and included in user-friendly

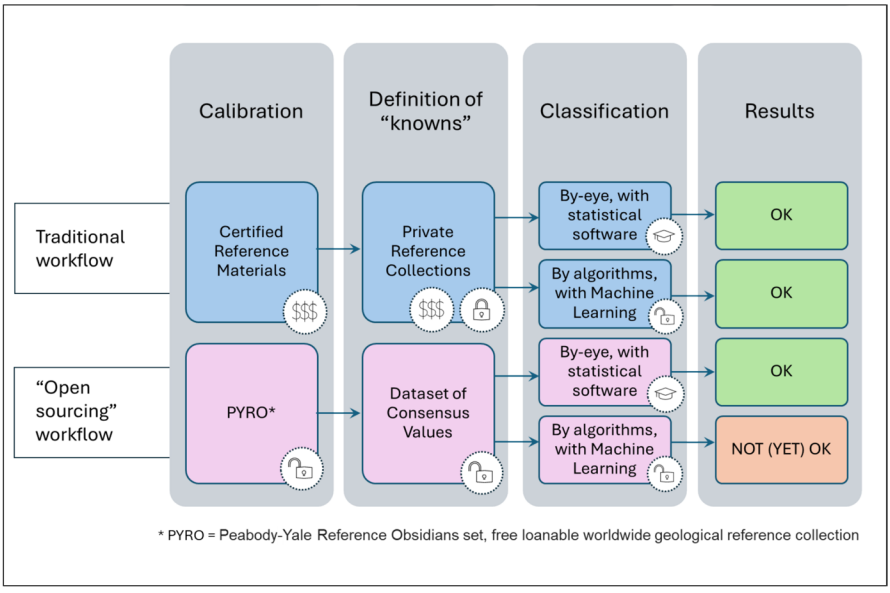

Fig. 13 Graphical comparison between a traditional workflow and an "open sourcing" workflow after the measurements phase. Requirements in terms of costs, openness/closeness, and expertise are summarized with the respective icons

online tools. The structured character of geochemical datasets, calibration processes, and sourcing algorithms also make them an ideal subject for machine learning, so that once any overlap issues are sufficiently resolved, we anticipate that machine learning and other tools in the AI repertoire will work even more effectively in the future.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10816-025-09695-8.

Acknowledgements For the study at hand, geological field specimens were measured by Michael Rummel in 2017 and all PYRO samples by Darius Danne in 2022. Dr. Philipp Hoelzmann was very helpful with organizing the measurements in the Laboratory for Physical Geography of the Institute for Geographical Sciences, Freie Universität Berlin. We warmly thank four anonymous reviewers, who offered very helpful input, comments, and suggestions, which considerably improved and clarified the arguments presented in the paper.

Author Contributions Conceptualization: Pavol Hnila, Ellery Frahm; Methodology: Pavol Hnila, Ellery Frahm; Formal analysis and investigation: Pavol Hnila; Writing—original draft preparation: Pavol Hnila, Ellery Frahm; Writing—review and editing: Arsen Bobokhyan, Alessandra Gilibert, Ellery Frahm, Pavol Hnila; Funding acquisition: Pavol Hnila; Sampling in field: Arsen Bobokhyan, Alessandra Gilibert, Ellery Frahm, Pavol Hnila; Supervision of laboratory measurements: Alessandra Gilibert, Pavol Hnila.

Funding Open Access funding enabled and organized by Projekt DEAL. The research leading to these results was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) — project no. 462731233.

Data Availability The authors confirm that all data generated or analyzed during this study are included in this published article and its electronic supplementary material.

Declarations

Competing Interests The authors declare no competing interests.

References

Alférez, G. H., Esteban, O. A., Clausen, B. L., & Ardila, A. M. M. (2022). Automated machine learning pipeline for geochemical analysis. *Earth Science Informatics, 15*(3), 1683–1698. https://doi.org/10.1007/s12145-022-00821-8

Andrade, J. M., & Estévez-Pérez, M. G. (2014). Statistical comparison of the slopes of two regression lines: A tutorial. *Analytica Chimica Acta, 838*, 1–12. https://doi.org/10.1016/j.aca.2014.04.057

Bobokhyan, A., Gilibert, A., & Hnila, P. (2015). Archaeology of vishap stones. In A. Petrosyan & A. Bobokhyan (Eds.), *The Vishap Stone Stelae* (1st ed., pp. 269–396). Yerevan: «GITUTYUN» publishing house (in Armenian).

Boschini, C., Hansen, A., & Wolf, S. (2022). *Discrete mathematics*. Zürich: vdf Hochschulverlag AG. https://doi.org/10.3929/ethz-b-000548053. Accessed 13 Oct 2023

Cann, J. R., & Renfrew, C. (1964). The characterization of obsidian and its application to the Mediterranean Region. *Proceedings of the Prehistoric Society, 30*, 111–133. https://doi.org/10.1017/S0079497X00015097

Da Silva, A. C., Triantafyllou, A., & Delmelle, N. (2023). Portable x-ray fluorescence calibrations: Workflow and guidelines for optimizing the analysis of geological samples. *Chemical Geology, 623*, 121395. https://doi.org/10.1016/j.chemgeo.2023.121395

Flament, C. (1963). *Application of graph theory to group structures*. Prentice-Hall.

Frahm, E. (2014). Characterizing obsidian sources with portable XRF: Accuracy, reproducibility, and field relationships in a case study from Armenia. *Journal of Archaeological Science, 49*, 105–125. https://doi.org/10.1016/j.jas.2014.05.003

Frahm, E. (2016). Can I get chips with that? Sourcing small obsidian artifacts down to microdebitage scales with portable XRF. *Journal of Archaeological Science: Reports, 9*, 448–467. https://doi.org/10.1016/j.jasrep.2016.08.032

Frahm, E. (2019). Introducing the Peabody-Yale Reference Obsidians (PYRO) sets: Open-source calibration and evaluation standards for quantitative X-ray fluorescence analysis. *Journal of Archaeological Science: Reports, 27*, 101957. https://doi.org/10.1016/j.jasrep.2019.101957

Frahm, E. (2023a). The obsidian sources of eastern Turkey and the Caucasus: Geochemistry, geology, and geochronology. *Journal of Archaeological Science: Reports, 49*, 104011. https://doi.org/10.1016/j.jasrep.2023.104011

Frahm, E. (2023b). Obsidian sources from the Aegean to central Turkey: Geochemistry, geology, and geochronology. *Journal of Archaeological Science: Reports, 52*, 104224. https://doi.org/10.1016/j.jasrep.2023.104224

Frahm, E., & Doonan, R. C. P. (2013). The technological versus methodological revolution of portable XRF in archaeology. *Journal of Archaeological Science, 40*(2), 1425–1434. https://doi.org/10.1016/j.jas.2012.10.013

Frahm, E., Martirosyan-Olshansky, K., Sherriff, J. E., Wilkinson, K. N., Glauberman, P., Raczynski-Henk, Y., et al. (2021). Geochemical changes in obsidian outcrops with elevation at Hatis volcano (Armenia) and corresponding Lower Palaeolithic artifacts from Nor Geghi 1. *Journal of Archaeological Science: Reports, 38*, 103097. https://doi.org/10.1016/j.jasrep.2021.103097

Ghosh, P. K., & Deguchi, K. (2008). *Mathematics of shape description: A morphological approach to image processing and computer graphics*. Singapore: John Wiley & Sons.Glascock, M. D. (2020). A systematic approach to geochemical sourcing of obsidian artifacts. *Scientific culture*, *6*(2), 35–47. https://doi.org/10.5281/zenodo.3724847

Gilibert, A., Bobokhyan, A., & Hnila, P. (2012). Dragon Stones in context. The discovery of high-altitude burial grounds with sculpted stelae in the Armenian mountains. *Mitteilungen der deutschen Orient-Gesellschaft, 144*, 93–132.

Glascock, M. D. (2020). A systematic approach to geochemical sourcing of obsidian artifacts. *Scientific Culture, 6*(2), 35–47. https://doi.org/10.5281/zenodo.3724847

Hnila, P., Gilibert, A., & Bobokhyan, A. (2019). Prehistoric sacred landscapes in the high mountains: The case of the vishap stelae between Taurus and Caucasus. In B. Engels, S. Huy, & C. Steitler (Eds.), *Natur und Kult in Anatolien: viertes Wissenschaftliches Netzwerk an der Abteilung Istanbul des Deutschen Archäologischen Instituts* (pp. 283–302). Istanbul: Ege Yayınları.

Keller, J., Djerbashian, R., Karapetian, S. G., Pernicka, E., & Nasedkin, V. (1996). Armenian and Caucasian obsidian occurrences as sources for the neolithic trade: Volcanological setting and chemical characteristics. In Ş. Demirci, A. M. Özer, & G. D. Summers (Eds.), *Archaeometry 94: The proceedings of the 29th International Symposium on Archaeometry, Ankara, 9–14 May 1994* (pp. 69–86). Presented at the International Symposium on Archaeometry, Ankara: TÜBİTAK.

Korstanje, J. (2022). *Machine learning on geographical data using Python: Introduction into* geodata *with applications and use cases*. Berkeley, CA: Apress. https://doi.org/10.1007/978-1-4842-8287-8

Kuzmin, Y. V., Oppenheimer, C., & Renfrew, C. (2020). er, C., & Renfrew, C. (2020). Global perspectives on o. *Quaternary International, 542*, 41–53. https://doi.org/10.1016/j.quaint.2020.02.036

Löwe, P., Anguix Alfaro, Á., Antonello, A., Baumann, P., Carrera, M., Durante, K., et al. (2022). Open Source – GIS. In W. Kresse & D. Danko (Eds.), *Springer Handbook of Geographic Information* (pp. 807–843). Cham: Springer International Publishing. https://doi.org/10.1007/978-3-030-53125-6_30

McMillan, R., Waber, N., Ritchie, M., & Frahm, E. (2022). Introducing SourceXplorer, an open-source statistical tool for guided lithic sourcing. *Journal of Archaeological Science, 144*, 105626. https://doi.org/10.1016/j.jas.2022.105626

Mussi, M., Mendez-Quintas, E., Barboni, D., Bocherens, H., Bonnefille, R., Briatico, G., et al. (2023). A surge in obsidian exploitation more than 1.2 million years ago at Simbiro III (Melka Kunture, Upper Awash, Ethiopia). *Nature Ecology & Evolution, 7*(3), 337–346. https://doi.org/10.1038/s41559-022-01970-1

Orange, M., Abedi, A., Le Bourdonnec, F.-X., Vosough, B., Ebrahimi, G., Razani, M., & Marro, C. (2021). Consuming local: The new obsidian source of Ideloo (Northwestern Iran) and first evidence of use by neighbouring prehistoric communities. *Geoarchaeology, 36*(2), 266–282. https://doi.org/10.1002/gea.21829

Schauer, M., Siegmund, F., Helfert, M., & Drake, B. L. (2024). The Munich Procedure – Standardising linear regression documentation in p-XRF research. *Software Impacts, 21*, 100660. https://doi.org/10.1016/j.simpa.2024.100660

Shackley, M. S. (1995). Sources of archaeological obsidian in the greater American Southwest: An update and quantitative analysis. *American Antiquity, 60*(3), 531–551. https://doi.org/10.2307/282264

Silliman, S. W. (2005). Obsidian studies and the archaeology of 19th-century California. *Journal of Field Archaeology, 30*(1), 75–94.

Varoutsikos, B., & Chataigner, C. (2012). Obsidatabase: Collecter et organiser les données relatives à l'obsidienne préhistorique au Proche-Orient et en Transcaucasie. In O. Henry (Ed.), *Archéologies et espaces parcourus* (pp. 11–21). İstanbul: Institut français d'études anatoliennes. https://doi.org/10.4000/books.ifeagd.987

Authors and Affiliations

Pavol Hnila[1,2] · Ellery Frahm[3,4] · Alessandra Gilibert[5] · Arsen Bobokhyan[6]

✉ Pavol Hnila
pavol.hnila@fu-berlin.de

1 Institute of Ancient Near Eastern Studies, Freie Universität Berlin, Berlin, Germany

2 Institute of Prehistory, Protohistory and Near-Eastern Archaeology, Heidelberg University, Heidelberg, Germany

3 Council On Archaeological Studies, Department of Anthropology, Yale University, New Haven, USA

4 Anthropology Division, Peabody Museum of Natural History, Yale University, New Haven, USA

5 Dipartimento Di Studi Umanistici, Università Ca' Foscari, Venice, Italy

6 Institute of Archaeology and Ethnography, National Academy of Sciences of the Republic of Armenia, Yerevan, Armenia