

How comparable are geochemical datasets really? and why it matters!

M. K. Traun, A. Renno, **Leander Kallas***, M. Willbold, T. Waight, D. Garbe-Schönberg & G. Wörner

Comparable to be compilable

Research using large geochemical datasets to understand Earth processes is usually based on compilations of smaller local or regional geochemical datasets. Assessing the comparability of two or more smaller datasets in a compilation is no simple task as a single value has a long tail of metadata.

Here we reference material offsets tools to assess the inter-study biases using GeoReM:

1. Checks if a measured value is within the expected distribution of values, according to the interquartile outlier criteria ie. "boxplot whiskers".
2. Checks if a measured value is within range of the based the Horwitz outlier criteria.
3. Data quality report tool that returns the inter-study bias as an absolute and relative offset of reference material values for user-provided paper identifiers like DOIs incl. metadata.

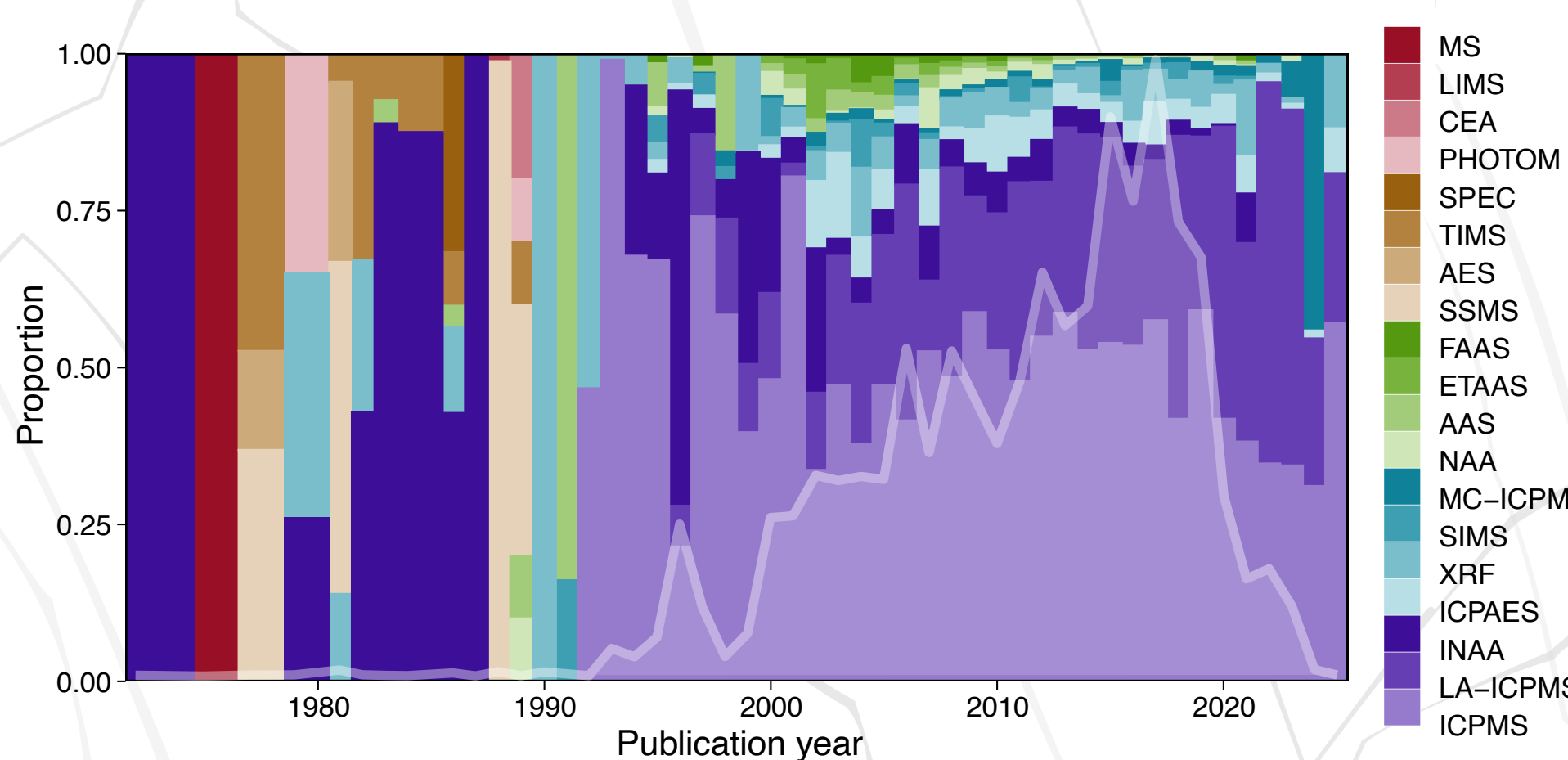
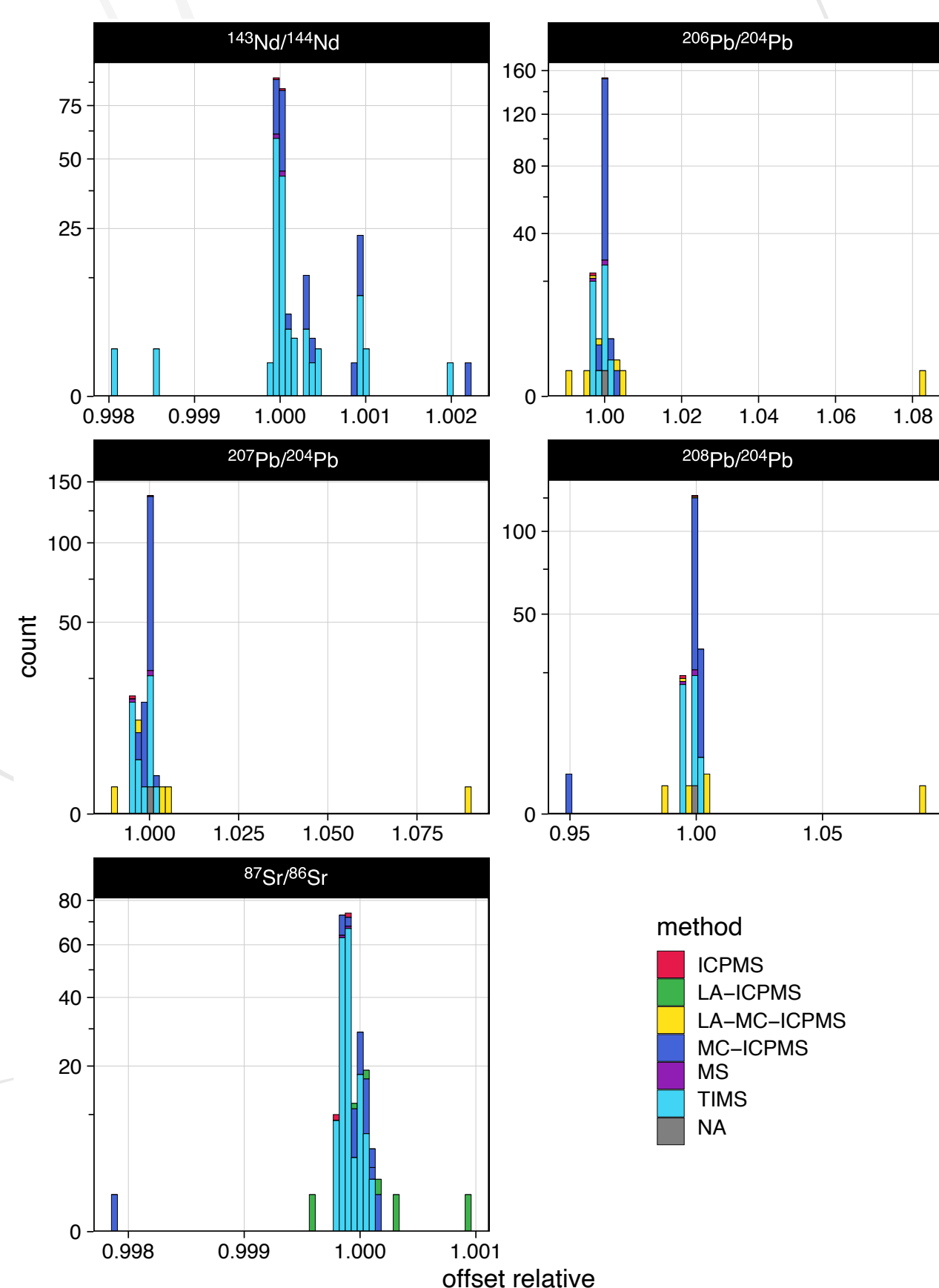
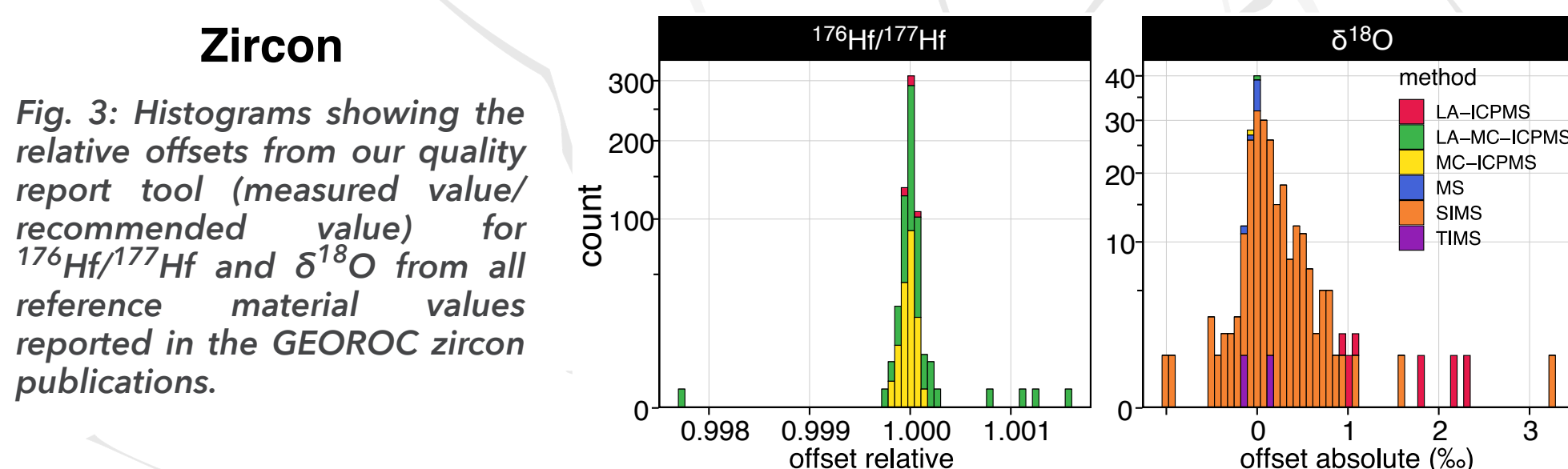


Fig. 1: The relative proportion of the most widely used methods to measure trace element concentrations ($n = 268447$) by publication year. Data source: GeoReM.



Mantle Zoo

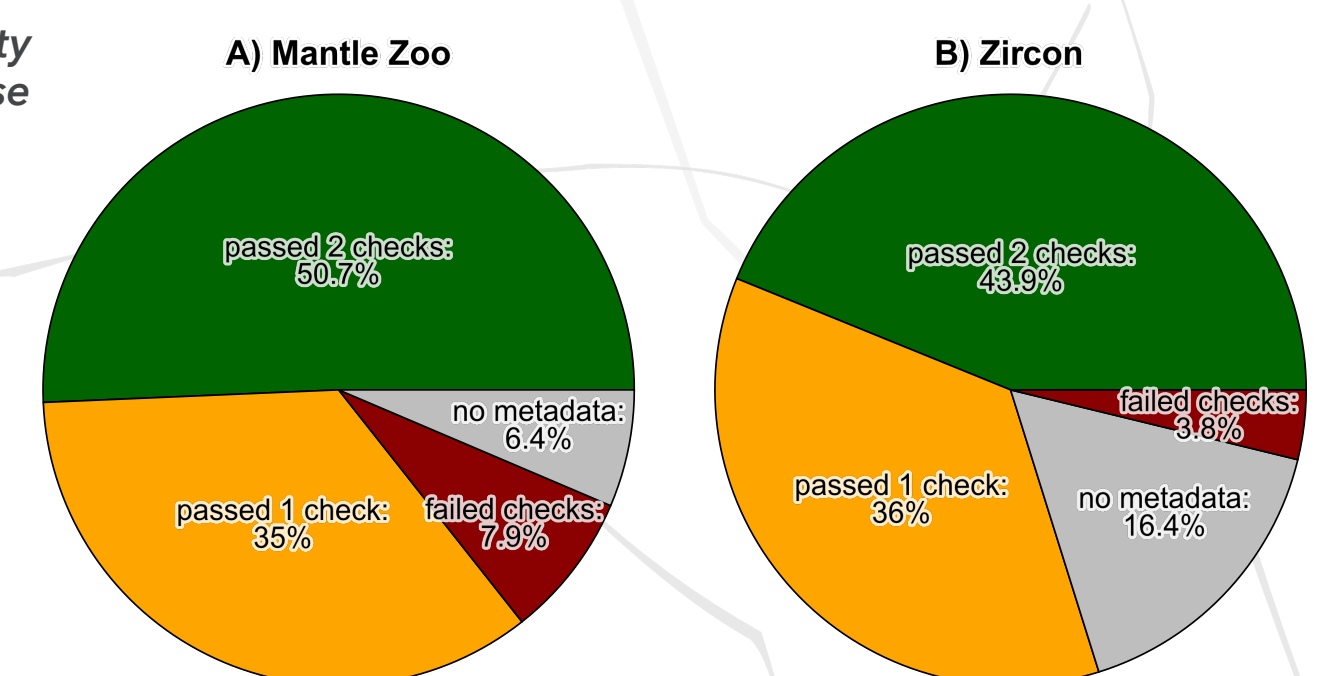
Fig. 2: Histograms showing the relative offsets from our quality report tool (measured value/recommended value) for the radiogenic isotope analytes from all reference materials reported in the Mantle Zoo publications. The y-axis is square root transformed to make the low count range easier to see.



Zircon

Fig. 3: Histograms showing the relative offsets from our quality report tool (measured value/recommended value) for $^{176}\text{Hf}/^{177}\text{Hf}$ and $\delta^{18}\text{O}$ from all reference material values reported in the GEOROC zircon publications.

Fig. 4: The result of the quality outlier checks for two use cases.



Unidentified bias of unknown impact

The tools are showcased for two famous geochemical big data research topics:

1. For the Mantle Zoo, there is a noticeable lack of metadata quality information available in the database for the EM-1. We also observe that the FOZO/PREMA end-member zone has a larger proportion of data that did not pass the reference material outlier checks for Pb-isotopes.
2. For zircons, there is notably less metadata coverage in the older range > 3 Ga in the database. Also, low scores often coincide with high $\delta^{18}\text{O}$ values at around 400 Ma, 1.9 Ga, 2.4 Ga and 3.3 Ga.

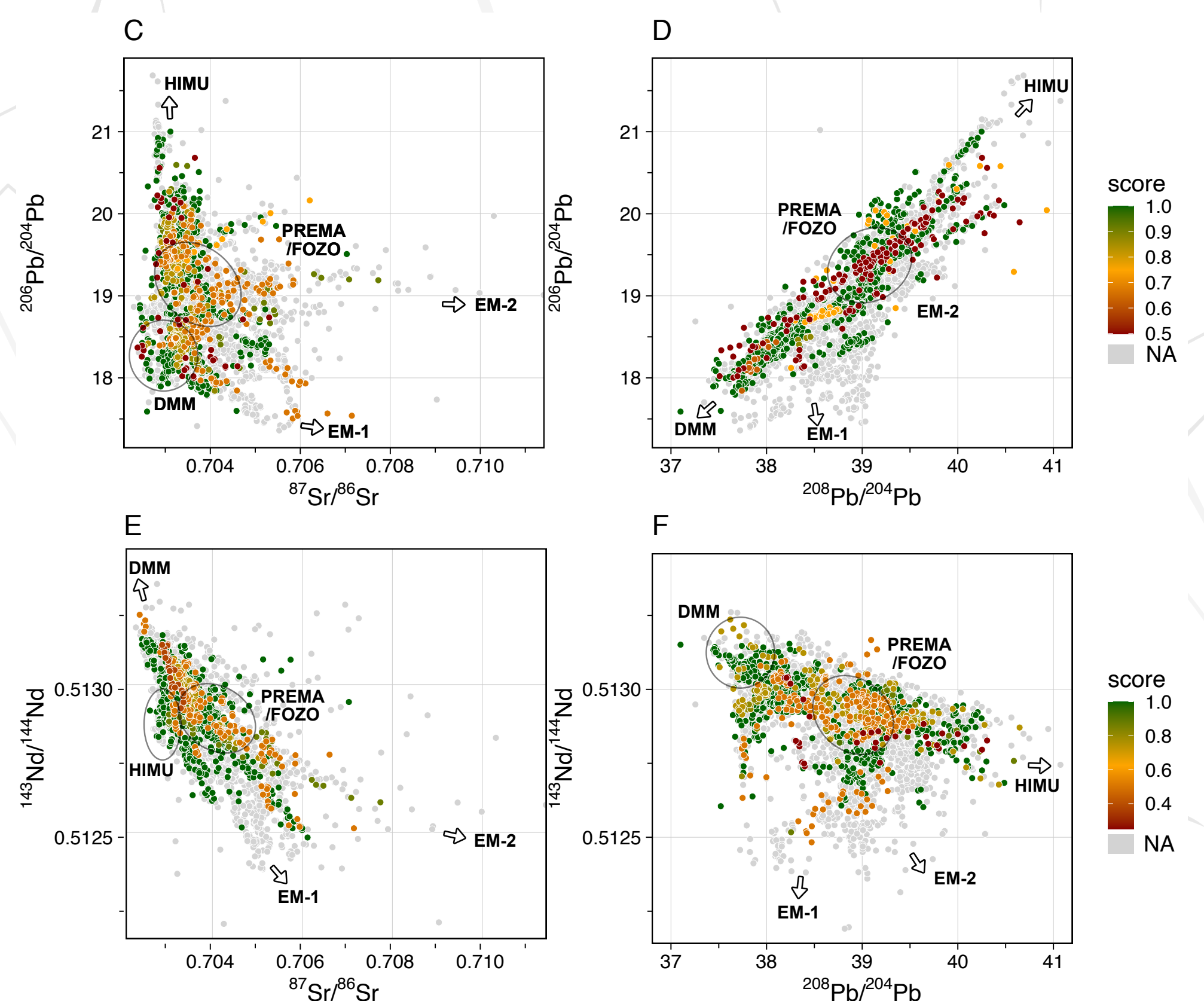
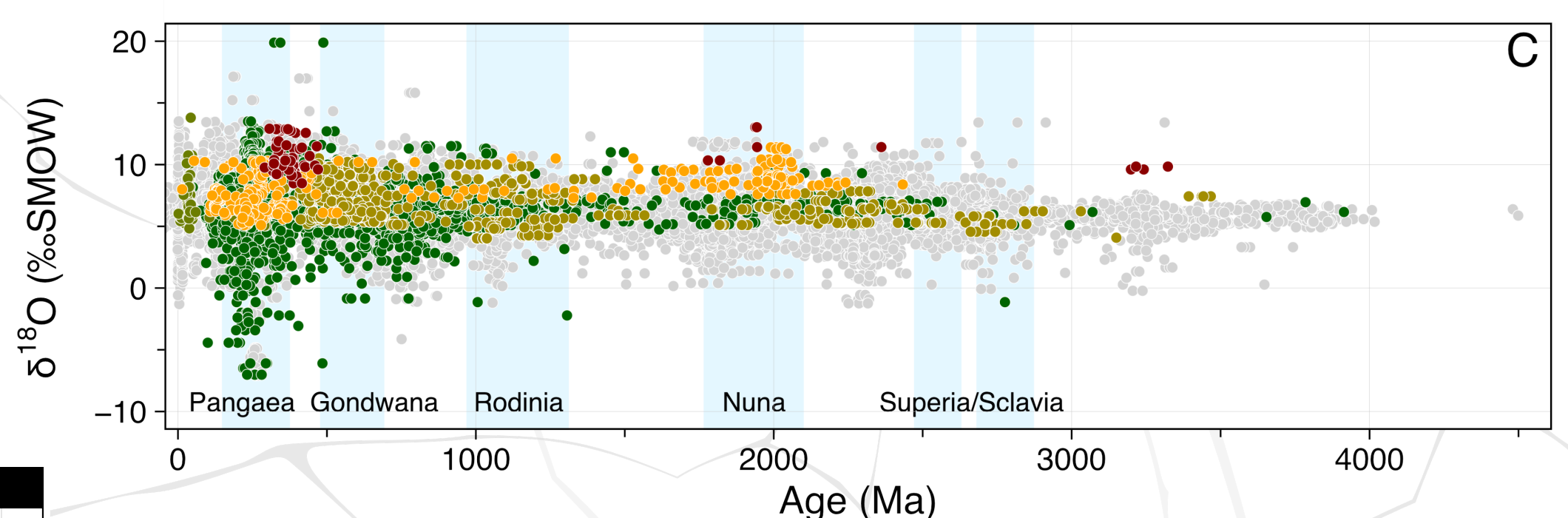


Fig. 5: "Classic" Mantle Zoo plots for oceanic basalts coloured by the offset score, defined as the proportion that passed our outlier checks. Data source: GEOROC.



Test your own compilations

How to use the tools through the new GeoReM API is demonstrated in Python Jupyter Notebooks and R Notebooks provided in the accompanying GitLab repository (QR code).

