

Supplementary information for:
Reconstructing Earth's Atmospheric Oxygenation History Using Machine Learning

Guoxiong Chen¹, Qiuming Cheng^{1,2,3*}, Timothy W. Lyons⁴, Jun Shen¹, Frits Agterberg⁵, Ning
Huang¹, Molei Zhao²

¹ State Key Laboratory of Geological Processes and Mineral Resources, China University of Geosciences, Wuhan, 430074, China.

² State Key Laboratory of Geological Processes and Mineral Resources, China University of Geosciences, Beijing, 10083, China.

³ School of Earth Science and Engineering, Sun Yat-sen University, Zhuhai, 51900, China.

⁴ Department of Earth and Planetary Sciences, University of California, Riverside, California, 92521, USA.

⁵ Geological Survey of Canada, 601 Booth Street, Ottawa, Ontario, K1A 0E8, Canada.

Corresponding author: Qiuming Cheng (qiuming.cheng@iugs.org)

This supplementary document contains the following information:

Supplementary Figure 1. Histogram of mafic igneous geochemical concentration.

Supplementary Figure 2. Filtering outliers in mafic igneous geochemistry data.

Supplementary Figure 3. Time series record of global mean mafic geochemistry composition.

Supplementary Figure 4. Cluster analysis of mafic geochemistry composition data.

Supplementary Figure 5. SOM analysis of mafic igneous geochemistry data.

Supplementary Figure 6. Principal component analysis of mafic igneous geochemistry data.

Supplementary Figure 7. Parameterization and optimization in SVR modeling.

Supplementary Figure 8. SVR modeling test with different uncertainties in training labels.

Supplementary Figure 9. SVR modeling test with constraints in boring billion and GOE.

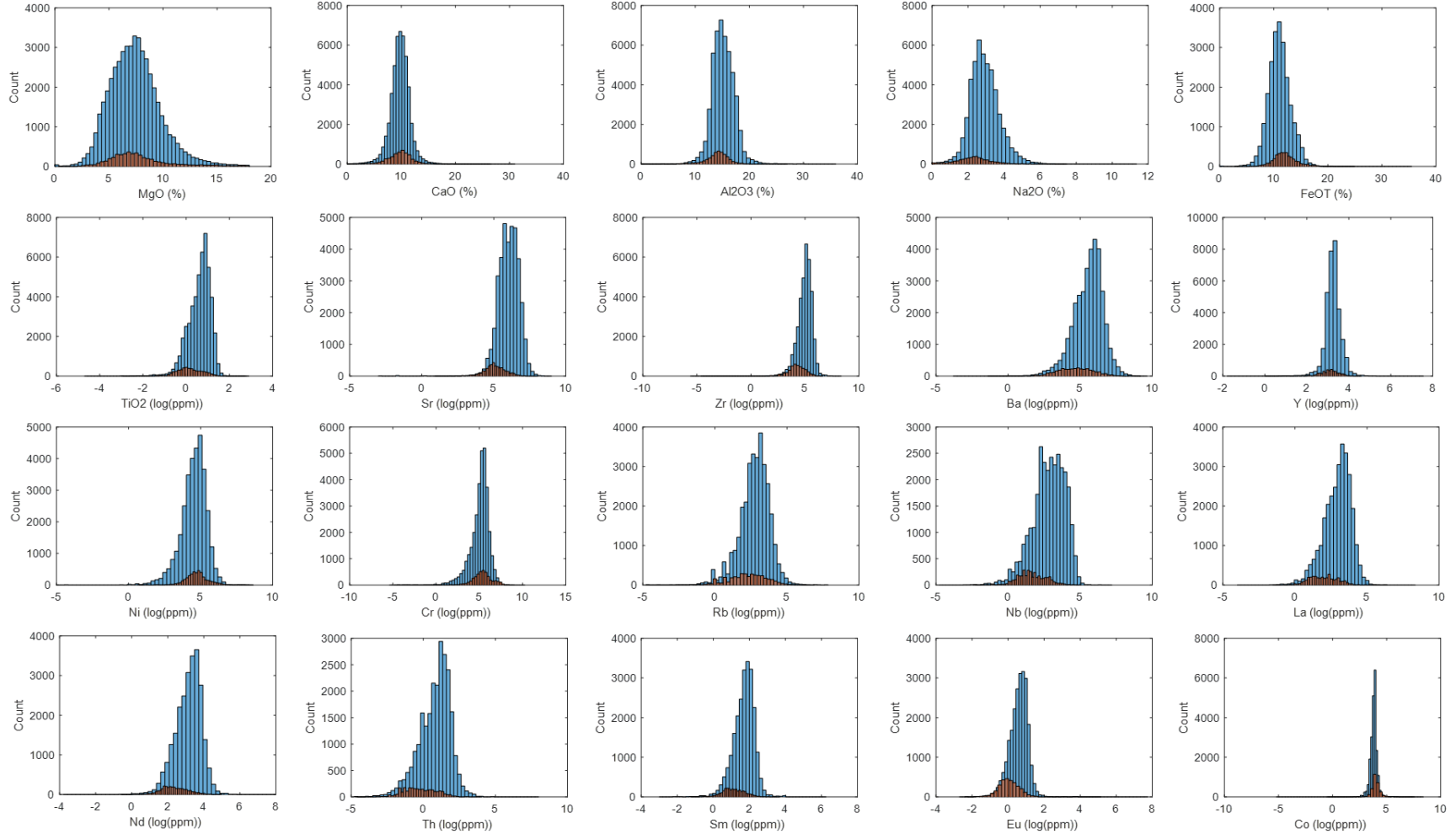
Supplementary Figure 10. SVR modeling test using different training datasets.

Supplementary Figure 11. PCA and SVR modeling of CLR-transformed composition data.

Supplementary Figure 12. O₂ modeling curves using different machine learning methods.

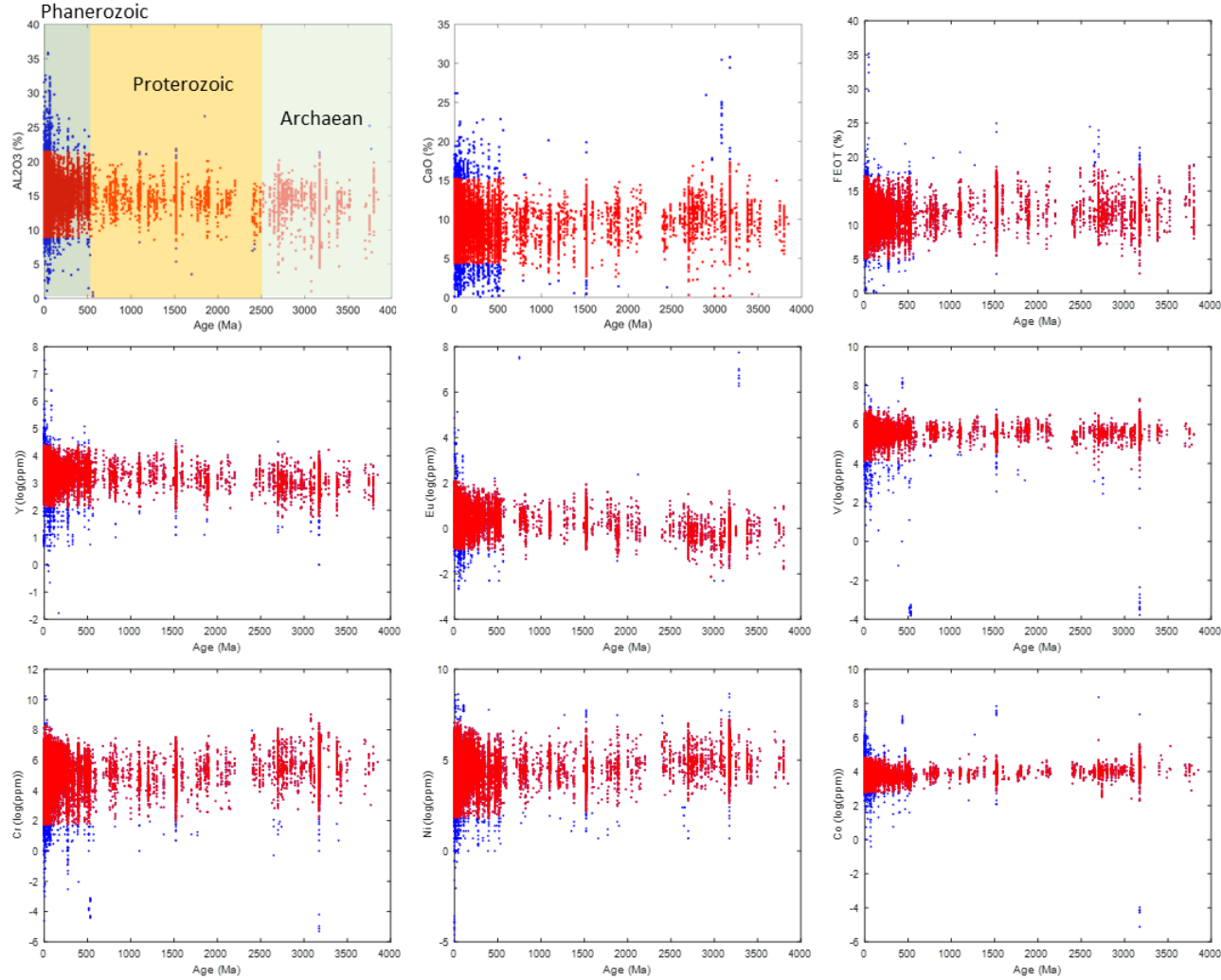
Supplementary Figure 13. Feature importance of all elements in RF modeling.

Supplementary Figure 14. Sedimentary-hosted paleo-oxybarometers through time.



Supplementary Figure 1

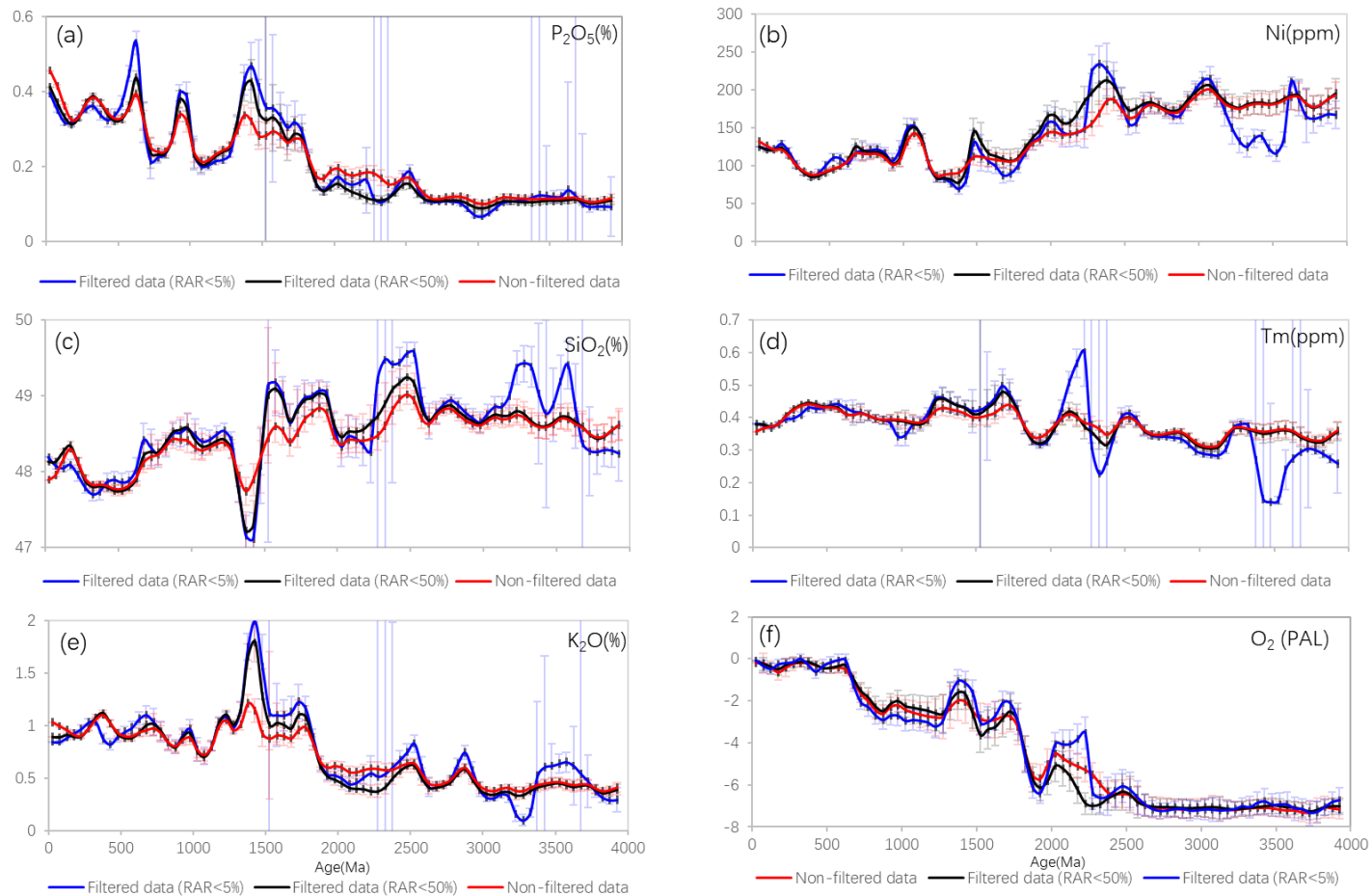
Histogram of mafic igneous geochemical concentration. Frequency distribution of mafic igneous geochemical concentration (few examples) during the Phanerozoic (blue) and Precambrian (brown) time. Trace elements often follow lognormal distribution, while major elements follow normal distribution.



Supplementary Figure 2

Filtering outliers in mafic igneous geochemistry data by using mean $\pm 3\delta$ thresholding. Logarithm transformation was applied to trace elements data for outliers filtering. The blue points indicate the filtered samples while red points are for calculation.

10

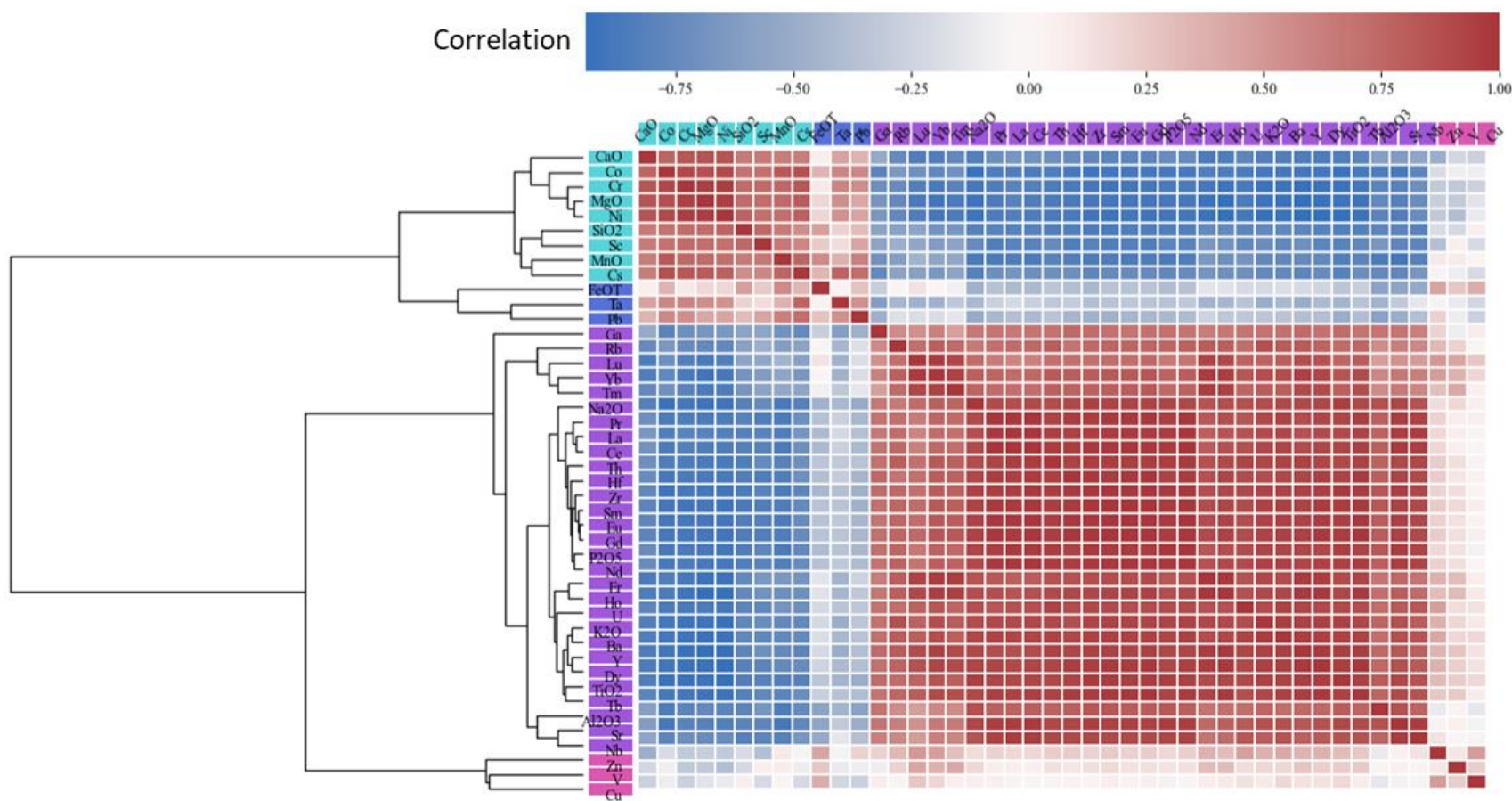


11

12

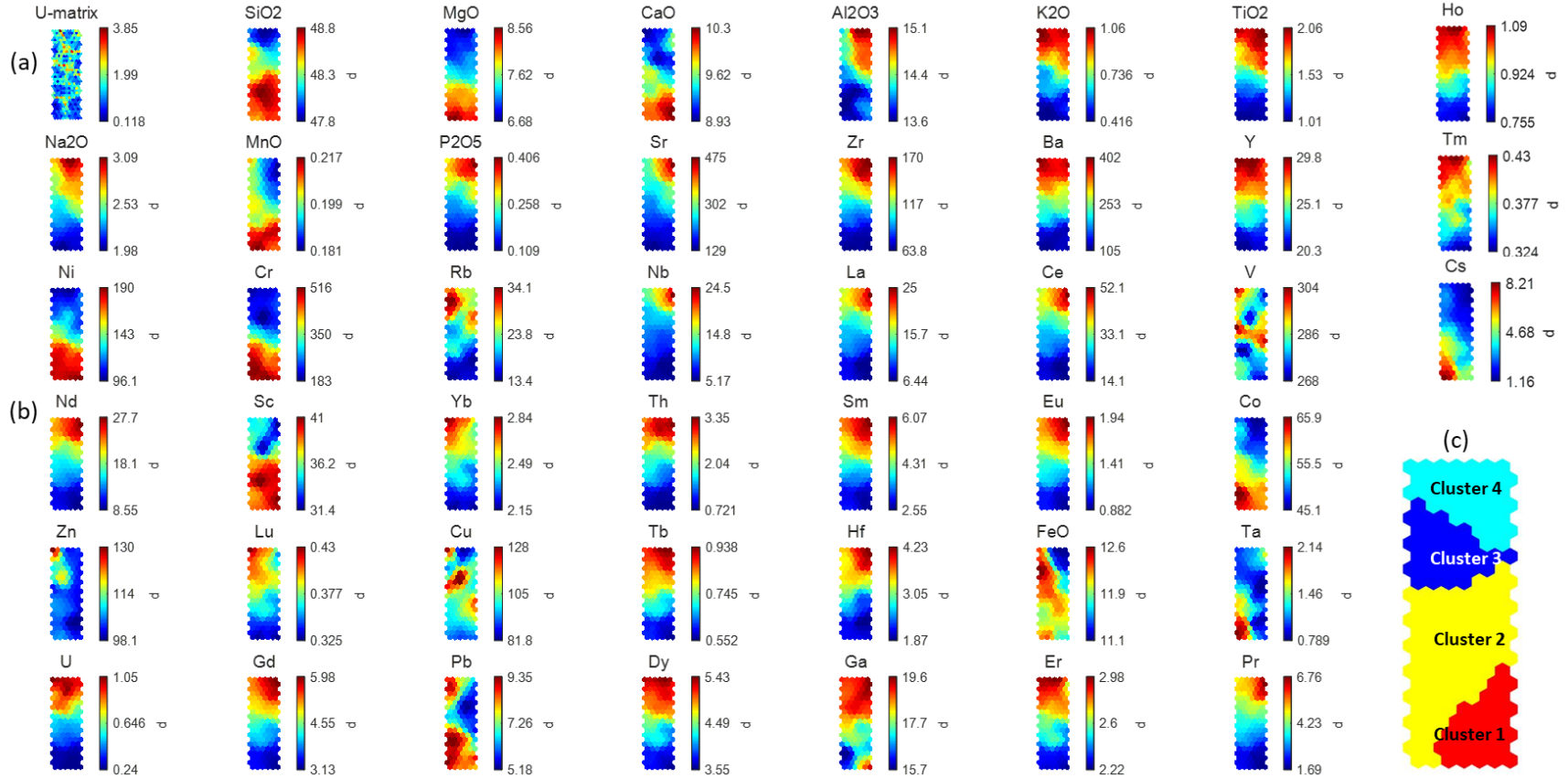
Supplementary Figure 3

13 Time series record of global mean mafic geochemistry composition. (a-e) Time series records of mean geochemical concentration
 14 obtained by the weighted bootstrap sampling method¹ from filtered data with RAR² <5% (blue line) and RAR<50% (black line), and
 15 non-RAR filtered dataset (red line). (f) Atmospheric O_2 modeling patterns obtained from filtered and non-RAR filtered datasets. The
 16 error bar shows 2 standard deviation (2-SD) uncertainties.



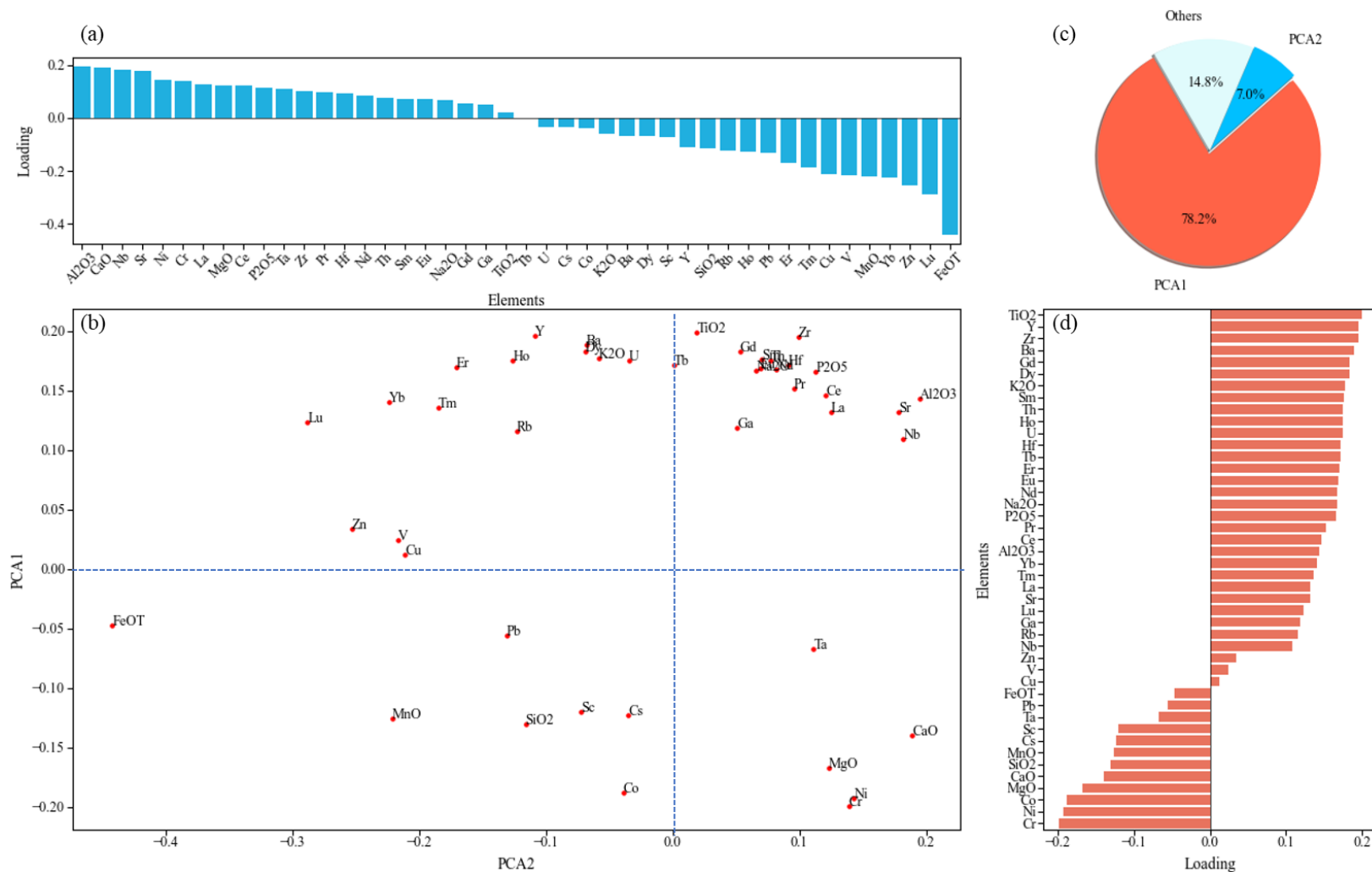
Supplementary Figure 4

Cluster analysis of mafic geochemistry composition data. Correlation matrix and cluster analysis of mafic composition spanning 44 elements from mafic igneous geochemical composition data; each square shows the Pearson's correlation coefficient for a pair of elements, and the intensity of the red or blue color indicates the value of correlation, as shown on the color bar; hierarchical clusters are represented by a cluster tree on the left panel.



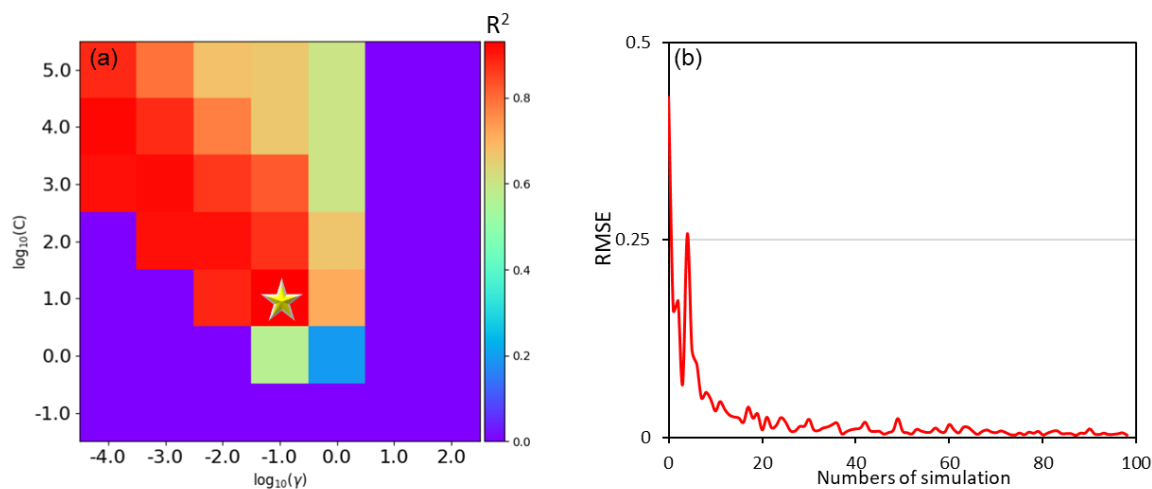
Supplementary Figure 5

SOM analysis of mafic igneous geochemistry data. (a) U-matrix of the trained SOM, uniform areas of low value indicate clusters while high values indicate cluster border, and color bar is the distance between the prototype vectors. (b) Component plots of 44 elements in SOM analysis. Component plots show the principal relations visually identified between the 44 variables. Warm colors represent high contribution to the SOM pattern, while cool colors represent low contribution. (c) SOM classification result of global mafic igneous geochemistry time series data, with colors coded in light blue, dark blue, yellow, and red for clusters 1, 2, 3 and 4, respectively.



Supplementary Figure 6

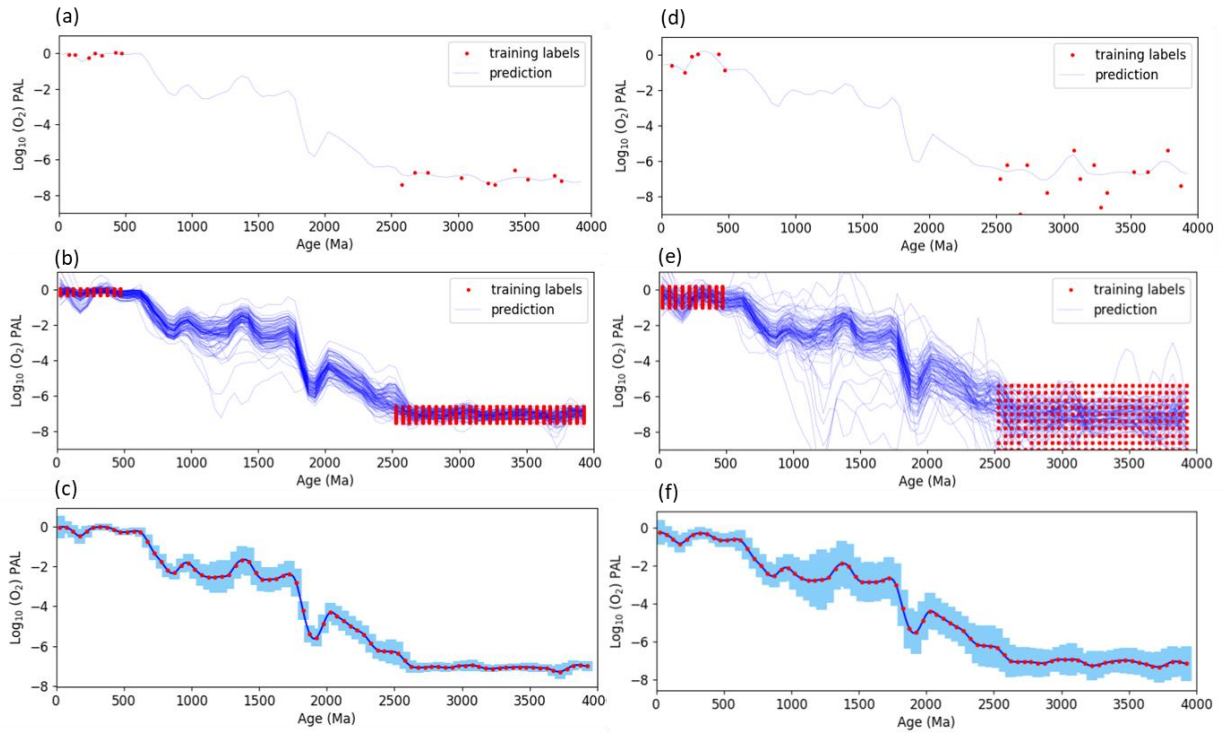
PCA analysis of mafic igneous geochemical data. (a) and (d) show the element loadings of PCA2 and PCA1, respectively. (b) Variance contribution of PCA1, PCA2 and rest principal components. (c) PCA1 loadings vs. PCA2 loadings for all elements.



Supplementary Figure 7

Parameterization and optimization in Support Vector Regression (SVR) modeling. (a) Illustration shows one example of prediction accuracy in variation with RBF function parameter γ and coefficient of penalty C for SVR, and the optimal parameters ($\gamma = 0.1$, $C = 10$) used in the calculation are determined by the maximum of $R^2 = 0.92$. (b) The root-mean-square-error (RMSE) curve of SVR modeling with Monte Carlo simulation.

48
49



50

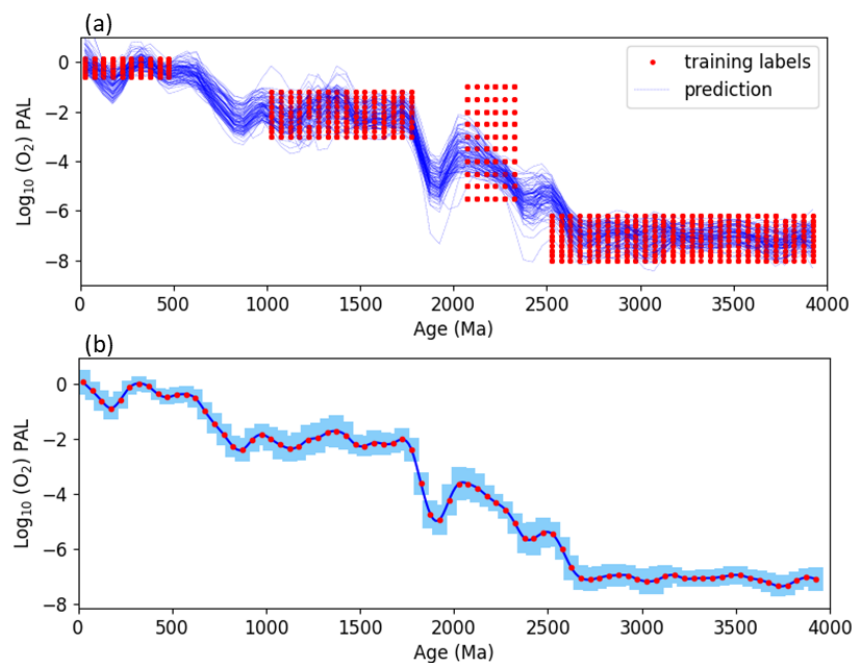
51

Supplementary Figure 8

52 SVR modeling test with different uncertainties in training labels. Left panels (a) to (c) show the
 53 prediction of atmospheric O_2 variation through Monte Carlo simulation, using random labels
 54 (including numbers and values) ranging from 0.5 to 1.2 PAL post 0.5 Ga and $10^{-6.5}$ to $10^{-7.5}$ PAL
 55 pre 2.5 Ga, while right panels (d) to (f) show the prediction using random labels ranging from 0.1
 56 to 2 PAL post 0.5 Ga, and 10^{-5} to 10^{-9} PAL pre 2.5 Ga. Illustrations (a) and (d) show one example
 57 of SVR predictions, while (b) and (e) show 100 curves of SVR predictions. (c) and (f) show the
 58 time series record of mean O_2 prediction through SVR and Monte Carlo simulation. The error bar
 59 shows 2 standard deviation (2-SD) uncertainties.

60

61



62

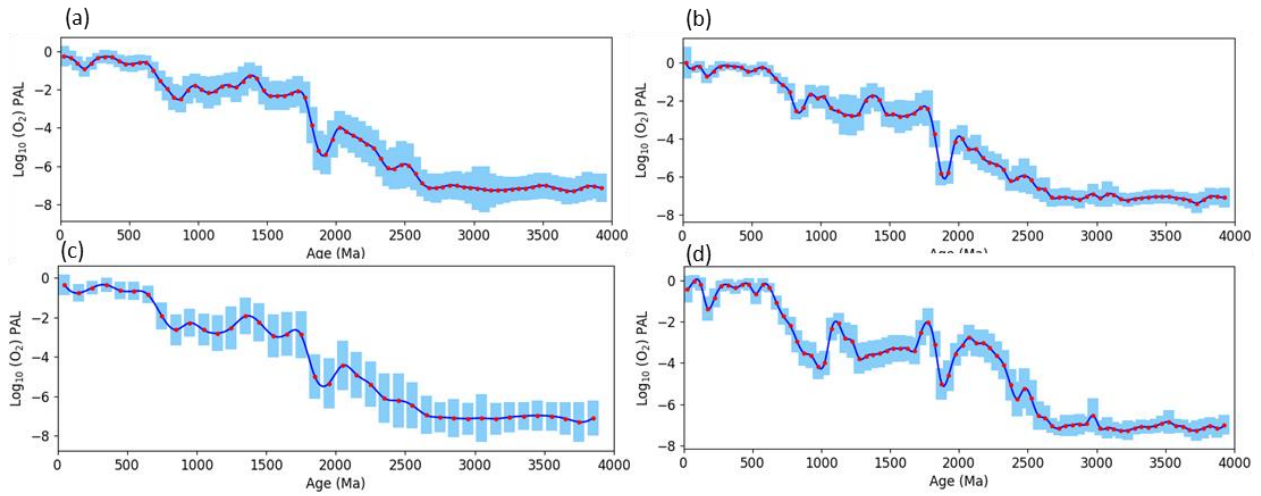
63

Supplementary Figure 9

64 SVR modeling test with constraints in boring billion and GOE. (a) All SVR modeling curves of
 65 atmospheric O_2 content using global mafic igneous geochemistry data through Monte Carlo
 66 simulations ($n = 100$) with additional informed labels during the “boring” billion and the “oxygen
 67 overshoot” of GOE. (b) Estimated atmospheric O_2 variation through time using mean value of
 68 1000 simulations, the error bar shows 2 standard deviation (2-SD) uncertainties

69

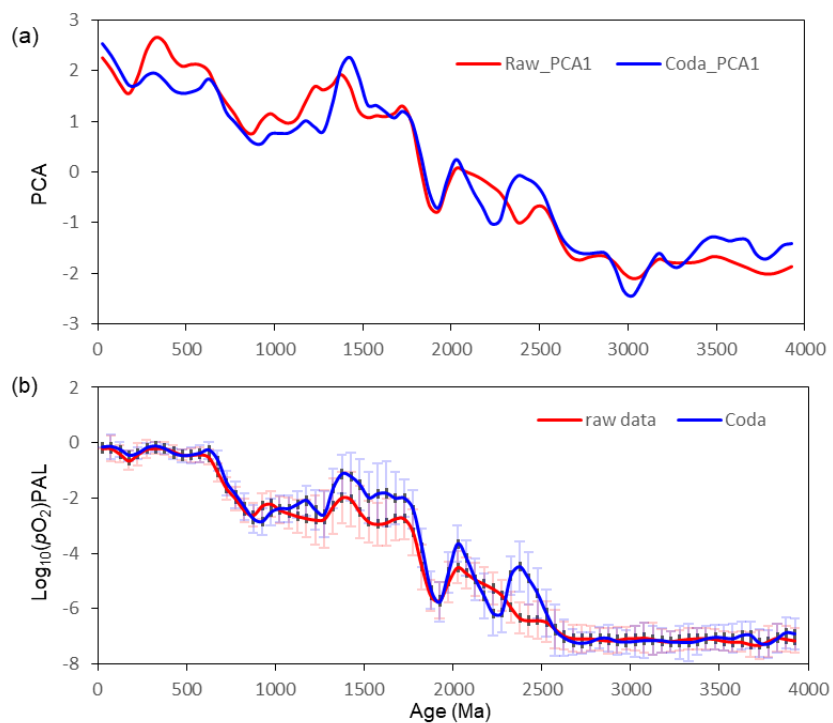
70



Supplementary Figure 10

SVR modeling test using different training datasets. (a) SVR modeling only using the highly correlated elements as suggested by cluster analysis in supplementary Fig. 4 (i.e., precluding FeO_T, Ta, Pb, Cu, V and Zn). (b) SVR modeling of atmospheric O₂ contents using mafic igneous geochemical time series data prepared with 50 Myr bin and uncertainty age = 50 Ma. (c) SVR modeling of atmospheric O₂ contents using mafic igneous geochemical time series data prepared with 100 Myr bin and uncertainty age = 100 Ma. (d) SVR modeling of atmospheric O₂ contents using the time series dataset of Keller and Schoene¹. The error bar shows 2 standard deviation (2-SD) uncertainties.

86



87

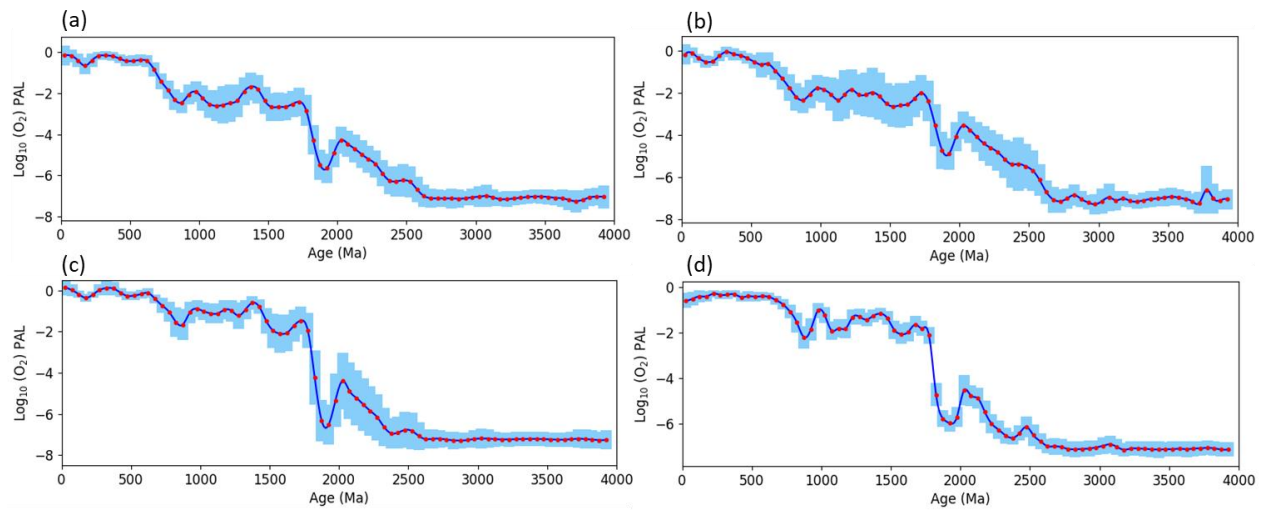
88

Supplementary Figure 11

89 PCA and SVR modeling of Centered Log-Ratio (CLR)³ transformed global mafic composition
 90 data. (a) The first principal component in PCA of global mafic composition data. (b) Atmospheric
 91 O_2 prediction using SVR method with CLR transformed mafic composition data. The error bar
 92 shows 2 standard deviation (2-SD) uncertainties.

93

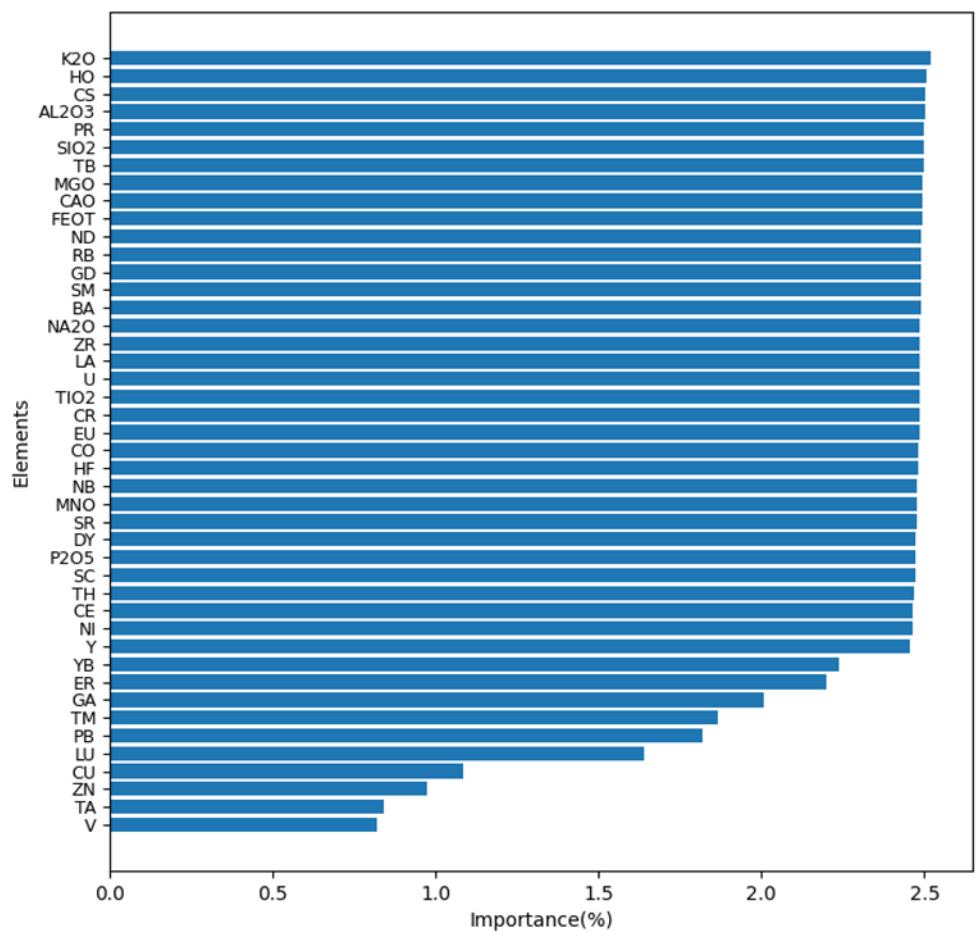
94



Supplementary Figure 12

O_2 modeling curves using different machine learning methods. SVR modeling of atmospheric oxygen level using (a) min-max scaling and (b) log scaling normalization methods. Also shown are reconstruction of atmospheric oxygen content using (c) artificial neural network and (d) random forests algorithms, respectively. The error bar shows 2 standard deviation (2-SD) uncertainties.

106
107



108

109

Supplementary Figure 13

110

111

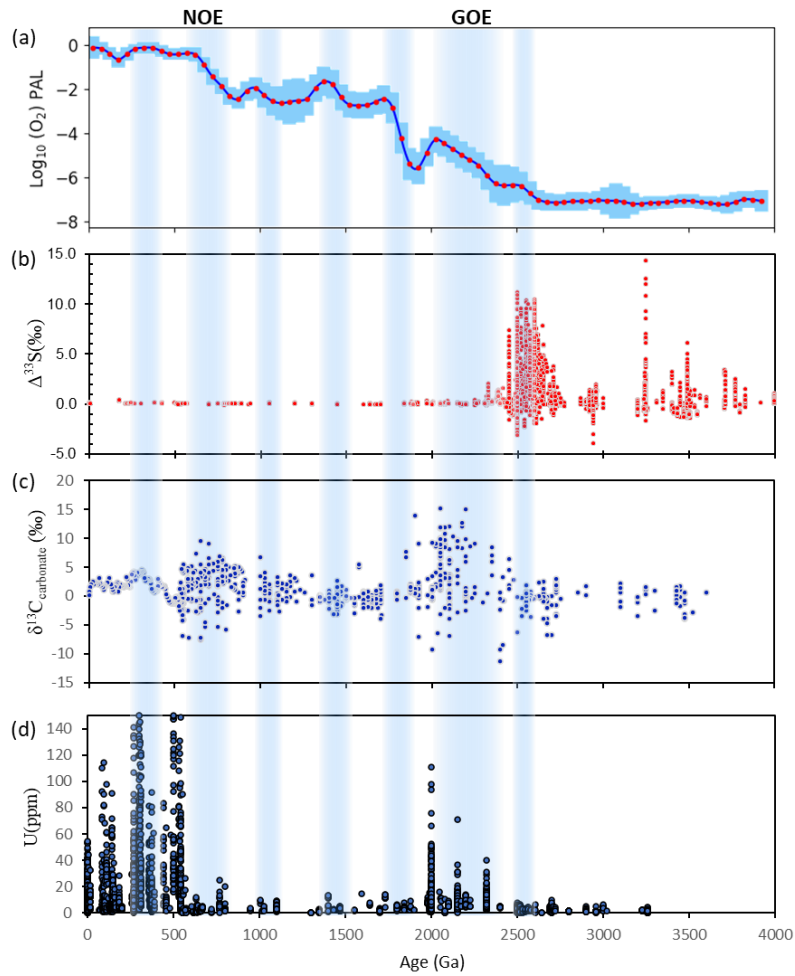
112

113

114

115

Feature importance of all elements in Random Forests modeling. Averaging feature importance of all elements after 10000 stochastic simulations, illustrating how each composition variable influences the final prediction model of atmospheric oxygen levels.



117

118

Supplementary Figure 14

119 Sedimentary-hosted paleo-oxybarometers through time. (a) The emerge curve of atmospheric O_2
 120 level in the present study, the error bar shows 2 standard deviation (2-SD) uncertainties. (b) Pyrite
 121 and sulfate $\Delta^{33}\text{S}$ data⁴. (c) Carbonate $\delta^{13}\text{C}$ data⁵. (d) Uranium concentrations from marine shales⁶.

122

123

Supplementary references

- 1 Keller, C. B. & Schoene, B. Statistical geochemistry reveals disruption in secular lithospheric evolution about 2.5 Gyr ago. *Nature* **485**, 490-493 (2012).
- 2 Doucet, L. S., Gamaleldien, H. & Li, Z.-X. Pitfalls in using the geochronological information from the EarthChem Portal for Precambrian time-series analysis. *Precambrian Research* **369**, 106514 (2022).
- 3 Greenacre, M. Compositional data analysis. *Annual Review of Statistics and its Application* **8**, 271-299 (2021).
- 4 Killingsworth, B. A. *et al.* Constraining the rise of oxygen with oxygen isotopes. *Nature communications* **10**, 1-10 (2019).
- 5 Krissansen-Totton, J., Buick, R. & Catling, D. C. A statistical analysis of the carbon isotope record from the Archean to Phanerozoic and implications for the rise of oxygen. *American Journal of Science* **315**, 275-316 (2015).
- 6 Partin, C. A. *et al.* Large-scale fluctuations in Precambrian atmospheric and oceanic oxygen levels from the record of U in shales. *Earth and Planetary Science Letters* **369**, 284-293 (2013).