Confidence of taxonomic identification: the first step of biostratigraphy

Yukun Shi*†[1] and Xiangdong Wang[1]

[1]School of Earth Sciences and Engineering and Frontiers Science Center for Critical Earth Material Cycling, Nanjing University – China

Abstract

Fossils are the most important indices for stratigraphy, but the issue of fossil identification confidence or consistency has been discussed and debated for decades (i.e. Gradstein et al., 1985; Polly and Head, 2004). With the highly developed database technology in the last decade, millions of fossil data of varying quality have been accumulated through formal or informal databases (i.e. The Paleobiology Database, OneStratigraphy, World Register of Marine Species). After evaluation, these data have been used for high-resolution stratigraphic correlation (Fan et al., 2020; Deng et al., 2021). However, when tens of thousands of fossil data are used in one quantitative analysis, how accurate, confident, or consistent of their identification could we achieve? For millions of fossil data in the databases, how many could be re-examined and evaluated, and in what way?

Here, an AI-based machine learning approach for fossil image identification is proposed to test the influence of the training set consistency. Three different training sets, one with the original labels from various identifiers and the other two with revised labels from two different experts, are used for independent training with the same deep learning model and then perform identification for the same test set. The consistency among the training set, as well as the machine identification results, could provide arguments for the taxonomic identification confidence issue and the above-mentioned questions.

The work is supported by the Natural Science Foundation of China (Grant 42293280).

References

• Deng, Y. Y., Fan, J. X., Zhang, S.H., Fang, X., Chen, Z.Y., Shi, Y.K., Wang, H. W., Wang, X. B., Yang, J., Hou, X.D., Wang, Y., Zhang, Y.D., Chen, Q., Yang, A. H., Fan, R., Dong, S.C., Xu, H.Q., Shen, S.Z., 2021. Timing and patterns of the Great Ordovician Biodiversification Event and Late Ordovician mass extinction. Earth-Science Reviews, 220: 103743.

• Fan, J., Shen, S., Erwin, D. H., Sadler, P. M., MacLeod, N., Cheng, Q., Hou, X., Yang, J., Wang, X., Wang, Y., Zhang, H., Chen, X., Li, G., Zhang, Y., Shi, Y., Yuan, D., Chen, Q., Zhang, L., Li, C. and Zhao, Y. 2020. A high-resolution summary of Cambrian to Early Triassic marine invertebrate biodiversity. Science 67(6475):272-277.

*Speaker

†Corresponding author: ykshi@nju.edu.cn

• Gradstein, F. M., Agterberg, F. P., Brower, J. C. and Schwarzacher, W., 1985. Quantitative stratigraphy. Dordrecht: Reidel. pp.24-25.

• Polly, P. D. and Head, J. J., 2004. Maximum-likelihood identification of fossils: taxonomic identification of Quaternary marmots (Rodentia, Mammalia) and identification of vertebral position in the pipesnake Cylindrophis (Serpentes, Reptilia). Morphometrics: applications in biology and paleontology, pp.197-221.

Keywords: Fossil identification, Machine learning, Quantitative stratigraphy, Database