Stanford
SCHOOL OF EARTH, ENERGY & ENVIRONMENTAL SCIENCES
Geological Sciences

# Using Machine Learning Models and Logistic Regression Analyses to Develop a Comprehensive Understanding of Extinction Risk For Marine Animal Phyla Across the Paleozoic


Adarsh S. Ambati(1), Anya Sengupta(1), Theo Chiang(1), Dr. Pedro Monarrez(2), Michael Pimentel-Galvan(2), Dr. Noel Heim(3), Dr. Jonathan Payne(2)

(1) Stanford Earth Young Investigators; (2) Stanford University Department of Geological Sciences; (3) Tufts University Department of Earth and Ocean Sciences


## Abstract


Extinction within the Paleozoic era has been studied in the past, but there still lacks a comprehensive understanding of how extinction risk changed throughout it. Our research project aims to bridge this gap by exploring extinction risk in relation to major Paleozoic phyla and ecological characteristics.

Using R, we analyzed the Stanford Earth Body Size dataset, which includes extensive data (n=8816) on Paleozoic marine animals. In Step 1, regression coefficients were formed, indicating whether being in one of the 6 phyla in each period of the Paleozoic era conferred greater or less extinction risk. In Step 2, the examined ecological characteristics included ocean acidification resilience, feeding patterns, body volume, length, surface area, motility, tiering, circulatory systems, and respiratory organ type. In Step 3, 6 binomial machine learning models were created using the traits from Step 2 to determine whether an individual genus went extinct in a particular period. Our Step 1 results confirm that within these timeframes, while certain phyla have greater extinction risk, extinction risk was not uniform across these groups. Our Step 2 results show certain traits provided advantages and disadvantages for an organism's extinction risk. One interesting pattern was that the only consistently non-significant traits were body length, area, and volume. Likewise with Step 1, extinction risk for each ecological characteristic varied across the Paleozoic. Finally, in Step 3, the results were largely successful. Most of the six models had an accuracy above 80% with the highest being 92% in the Silurian. The areas under the Precision-Recall and the Receiver Operating Characteristic Curves were all in the acceptable (>0.6) range, demonstrating that the model has low false positive/ negative rates and is able to distinguish between what trait indicates extinction or survival for each period.

Our research project identified phyla at risk of extinction in each period of the Paleozoic, determined which natural traits incited greater extinction risk, and demonstrated machine learning models trained on fossil descriptors can predict when an individual genus became extinct. Our results confirmed that extinction risk is not consistently dependent on a singular factor nor is it constant across every period of the Paleozoic era.


## Methodology

The dataset we used included nine biological and ecological traits, and it also includes taxonomic groupings and phyla. After making logistic regression models for each trait, we then made regularized regression models predicting extinction in each period based on these characteristics. Below are the Paleozoic periods that we are analyzing.

**Cambrian**
**Ordovician**
**Silurian**
**Devonian**
**Carboniferous**
**Permian**

All analyses and plots were made using the programming language R. During stages 1 and 2, the following were our categories of analysis.
1. Phyla - Echinodermata, Mollusca, Chordata, Arthropoda, Brachiopoda, Foraminifera (Taxonomic Group)
2. Descriptors - buffering, feeding patterns, motility, oceanic tiering, respiratory organ type, circulatory system type, length, surface area, volume

For Stage 3, we built regularized binomial regression models.

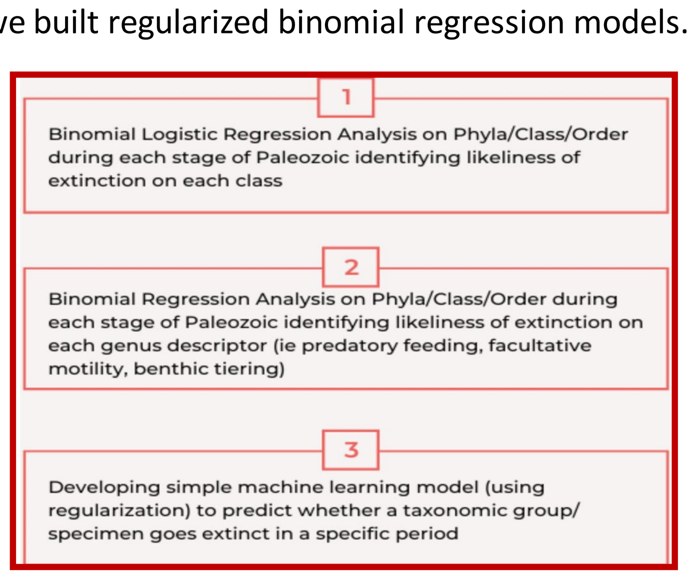
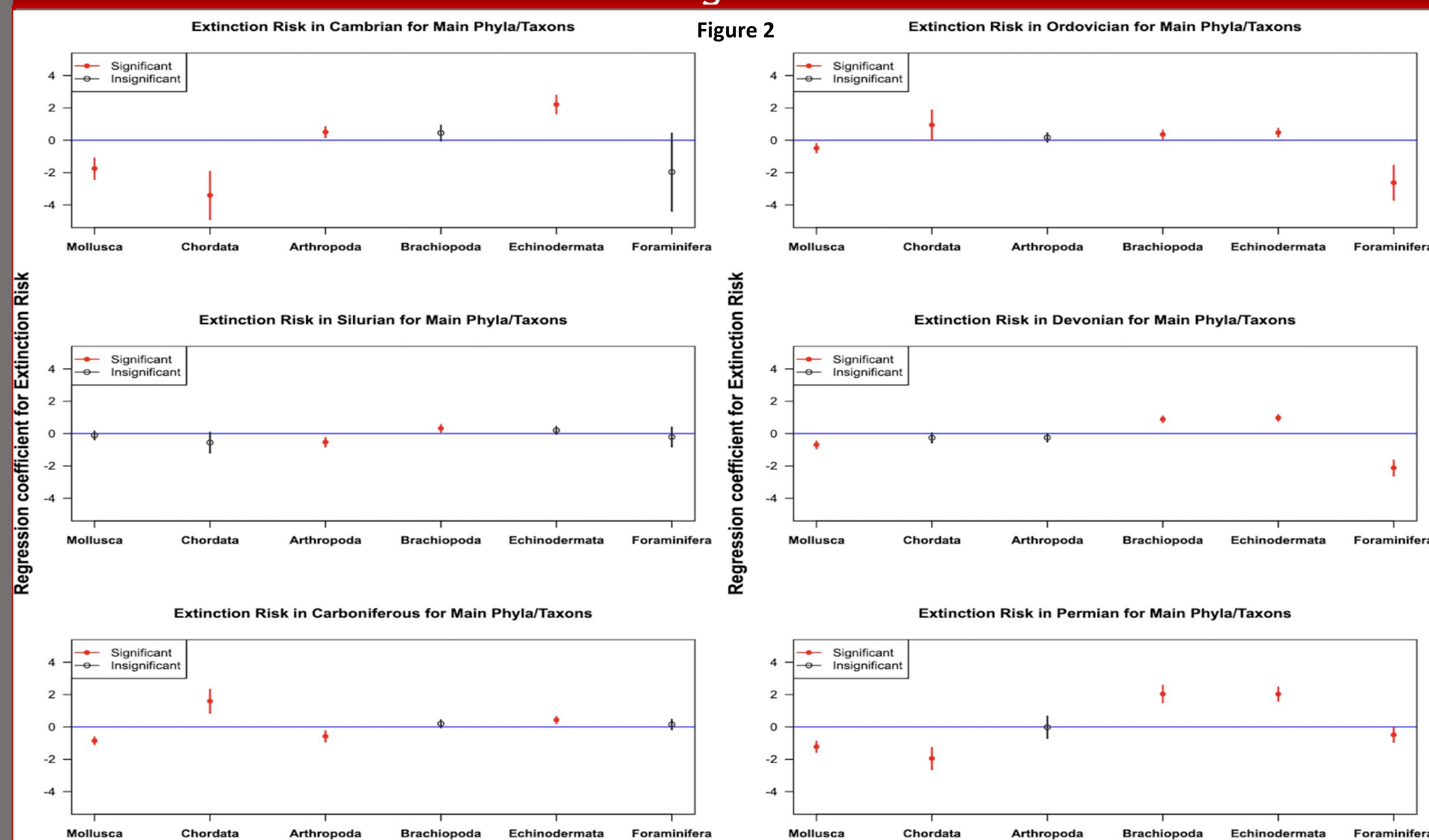



**Figure 1: Summary of Analysis Stages**

## Figures



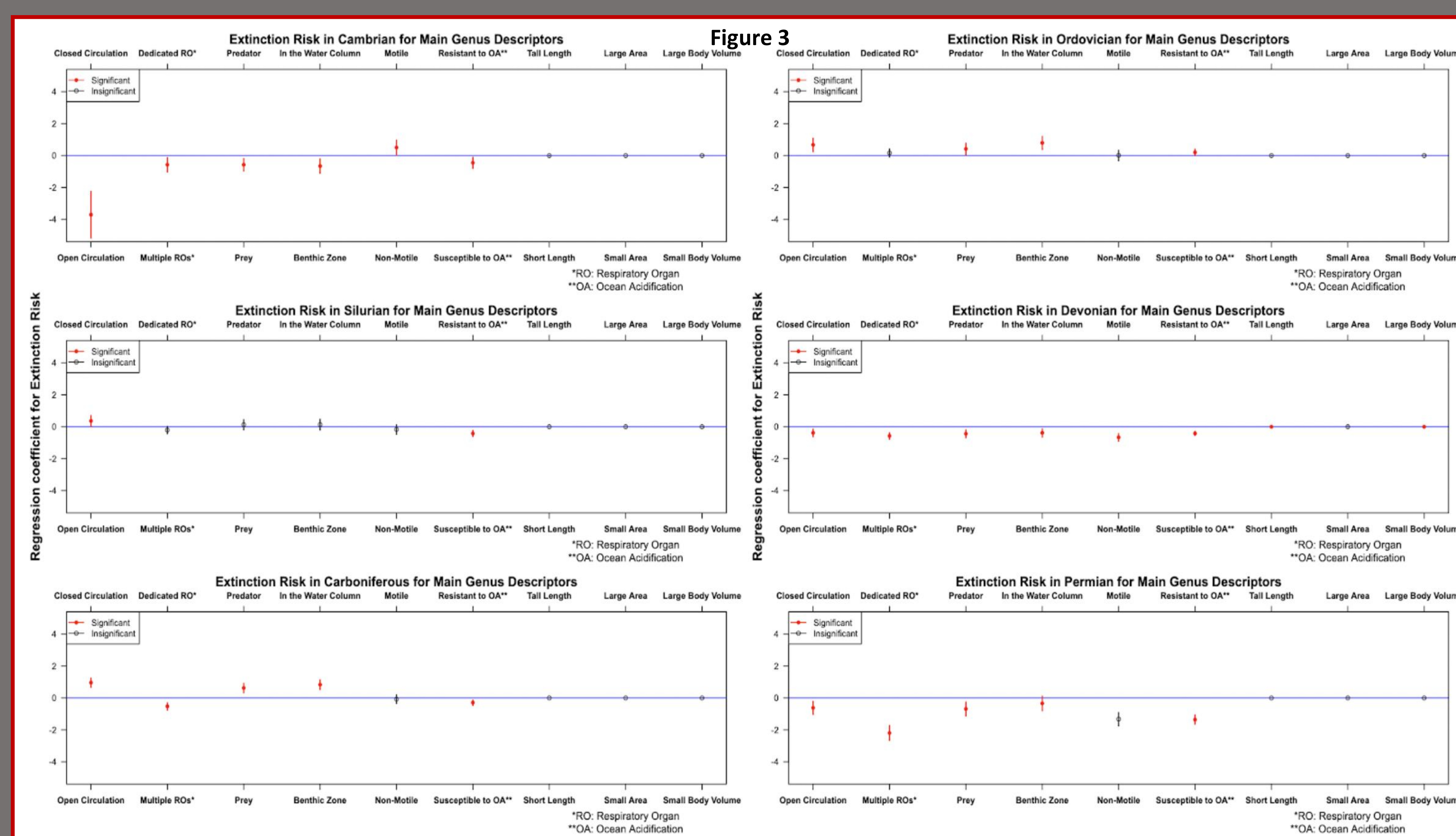




**Figure 4**

| | Cambrian Model Features | Score |
|---|---|---|
| 1 | Circulation System | 3.3422079 |
| 2 | Motility | 2.27727063 |
| 3 | Ocean Acidification Resistance | 1.12284708 |
| 4 | Feeding Patterns | 0.42942449 |
| 5 | Respiratory Organ System | 0.37019192 |
| 6 | Tiering | 0.02890546 |
| 7 | Maximum Length | 0.00479994 |
| 8 | Maximum Area | 4.23E-07 |
| 9 | Calculated Maximum Volume | 2.39E-09 |

| | Ordovician Model Features | Score |
|---|---|---|
| 1 | Motility | 1.06918333 |
| 2 | Ocean Acidification Resistance | 0.85323525 |
| 3 | Tiering | 0.33049509 |
| 4 | Feeding Patterns | 0.32654485 |
| 5 | Respiratory Organ System | 0.15505047 |
| 6 | Circulation System | 0.02294367 |
| 7 | Maximum Length | 0.00124485 |
| 8 | Maximum Area | 4.10E-06 |
| 9 | Calculated Maximum Volume | 1.08E-08 |

| | Silurian Model Features | Score |
|---|---|---|
| 1 | Motility | 1.32779174 |
| 2 | Circulation System | 1.06206068 |
| 3 | Respiratory Organ System | 0.70851463 |
| 4 | Feeding Patterns | 0.6192809 |
| 5 | Ocean Acidification Resistance | 0.59558913 |
| 6 | Tiering | 0.59509338 |
| 7 | Maximum Length | 0.00258455 |
| 8 | Maximum Area | 3.22E-06 |
| 9 | Calculated Maximum Volume | 3.32E-08 |

| | Devonian Model Features | Score |
|---|---|---|
| 1 | Ocean Acidification Resistance | 0.70835892 |
| 2 | Respiratory Organ System | 0.45876009 |
| 3 | Motility | 0.22309167 |
| 4 | Circulation System | 0.06226351 |
| 5 | Maximum Length | 0.00555858 |
| 6 | Maximum Area | 2.00E-05 |
| 7 | Calculated Maximum Volume | 4.32E-08 |
| 8 | Feeding Patterns | 0 |
| 9 | Tiering | 0 |

| | Carboniferous Model Features | Score |
|---|---|---|
| 1 | Circulation System | 2.32544688 |
| 2 | Ocean Acidification Resistance | 0.73146419 |
| 3 | Respiratory Organ System | 0.65025381 |
| 4 | Maximum Length | 0.00189793 |
| 5 | Feeding Patterns | 0 |
| 6 | Tiering | 0 |
| 7 | Motility | 0 |
| 8 | Maximum Area | 0 |
| 9 | Calculated Maximum Volume | 0 |

| | Permian Model Features | Score |
|---|---|---|
| 1 | Circulation System | 2.18390017 |
| 2 | Respiratory Organ System | 1.04694225 |
| 3 | Ocean Acidification Resistance | 0.93551679 |
| 4 | Feeding Patterns | 0.86094481 |
| 5 | Tiering | 0.74849748 |
| 6 | Motility | 0.70993468 |
| 7 | Maximum Length | 0.000241 |
| 8 | Calculated Maximum Volume | 2.56E-08 |
| 9 | Maximum Area | 0 |

**Figure 5**

| Period | Alpha Value | Lambda Value | Accuracy | AUROC | AUPR |
|---|---|---|---|---|---|
| Cambrian | 0 | 0.01 | 0.92 | 0.93 | 0.8 |
| Cambrian | 0.1 | 0.01 | 0.92 | 0.93 | 0.8 |
| Cambrian | 0.2 | 0.01 | 0.92 | 0.93 | 0.8 |
| Cambrian | 0.3 | 0.01 | 0.92 | 0.93 | 0.8 |
| Cambrian | 0.4 | 0.01 | 0.92 | 0.93 | 0.8 |
| Cambrian | 0.5 | 0.01 | 0.92 | 0.93 | 0.8 |
| Cambrian | 0.6 | 0.01 | 0.92 | 0.93 | 0.8 |
| Cambrian | 0.7 | 0.01 | 0.92 | 0.93 | 0.8 |
| Cambrian | 0.8 | 0.01 | 0.92 | 0.93 | 0.8 |
| Cambrian | 0.9 | 0.01 | 0.08 | 0.5 | 0.79 |
| Cambrian | 1 | 0.01 | 0.91 | 0.5 | 0 |

| Period | Alpha Value | Lambda Value | Accuracy | AUROC | AUPR |
|---|---|---|---|---|---|
| Ordovician | 0 | 0.012589254 | 0.83 | 0.67 | 0.73 |
| Ordovician | 0.1 | 0.01 | 0.83 | 0.68 | 0.73 |
| Ordovician | 0.2 | 0.012589254 | 0.83 | 0.67 | 0.73 |
| Ordovician | 0.3 | 0.01 | 0.83 | 0.67 | 0.73 |
| Ordovician | 0.4 | 0.01 | 0.83 | 0.67 | 0.73 |
| Ordovician | 0.5 | 0.01 | 0.83 | 0.66 | 0.73 |
| Ordovician | 0.6 | 0.01 | 0.82 | 0.66 | 0.73 |
| Ordovician | 0.7 | 0.01 | 0.82 | 0.66 | 0.73 |
| Ordovician | 0.8 | 0.01 | 0.82 | 0.66 | 0.74 |
| Ordovician | 0.9 | 0.01 | 0.82 | 0.61 | 0.74 |
| Ordovician | 1 | 0.01 | 0.82 | 0.61 | 0.75 |

| Period | Alpha Value | Lambda Value | Accuracy | AUROC | AUPR |
|---|---|---|---|---|---|
| Silurian | 0 | 0.005416508 | 0.91 | 0.6 | 0.89 |
| Silurian | 0.1 | 0.001219581 | 0.91 | 0.6 | 0.89 |
| Silurian | 0.2 | 0.000578704 | 0.91 | 0.6 | 0.89 |
| Silurian | 0.3 | 0.000578704 | 0.91 | 0.6 | 0.89 |
| Silurian | 0.4 | 0.000578704 | 0.91 | 0.6 | 0.89 |
| Silurian | 0.5 | 0.000578704 | 0.91 | 0.6 | 0.89 |
| Silurian | 0.6 | 0.000578704 | 0.91 | 0.6 | 0.89 |
| Silurian | 0.7 | 0.000578704 | 0.91 | 0.6 | 0.89 |
| Silurian | 0.8 | 0.000578704 | 0.91 | 0.6 | 0.89 |
| Silurian | 0.9 | 0.000578704 | 0.91 | 0.6 | 0.89 |
| Silurian | 1 | 0.000578704 | 0.91 | 0.6 | 0.89 |

| Period | Alpha Value | Lambda Value | Accuracy | AUROC | AUPR |
|---|---|---|---|---|---|
| Devonian | 0 | 0.025118864 | 0.79 | 0.64 | 0.75 |
| Devonian | 0.1 | 0.015848932 | 0.79 | 0.64 | 0.75 |
| Devonian | 0.2 | 0.019952623 | 0.79 | 0.64 | 0.75 |
| Devonian | 0.3 | 0.006309573 | 0.79 | 0.64 | 0.75 |
| Devonian | 0.4 | 0.003981072 | 0.79 | 0.64 | 0.75 |
| Devonian | 0.5 | 0.003981072 | 0.79 | 0.64 | 0.75 |
| Devonian | 0.6 | 0.003162278 | 0.79 | 0.64 | 0.75 |
| Devonian | 0.7 | 0.001584893 | 0.79 | 0.64 | 0.75 |
| Devonian | 0.8 | 0.002511886 | 0.79 | 0.64 | 0.75 |
| Devonian | 0.9 | 0.001995262 | 0.79 | 0.64 | 0.75 |
| Devonian | 1 | 0.003162278 | 0.79 | 0.64 | 0.75 |

| Period | Alpha Value | Lambda Value | Accuracy | AUROC | AUPR |
|---|---|---|---|---|---|
| Carboniferous | 0 | 0.003541732 | 0.83 | 0.6 | 0.77 |
| Carboniferous | 0.1 | 0.002915452 | 0.83 | 0.6 | 0.77 |
| Carboniferous | 0.2 | 0.011383466 | 0.83 | 0.6 | 0.77 |
| Carboniferous | 0.3 | 0.002915452 | 0.83 | 0.6 | 0.77 |
| Carboniferous | 0.4 | 0.002915452 | 0.83 | 0.6 | 0.77 |
| Carboniferous | 0.5 | 0.002915452 | 0.83 | 0.6 | 0.77 |
| Carboniferous | 0.6 | 0.003541732 | 0.83 | 0.6 | 0.77 |
| Carboniferous | 0.7 | 0.002915452 | 0.83 | 0.6 | 0.77 |
| Carboniferous | 0.8 | 0.002915452 | 0.83 | 0.6 | 0.77 |
| Carboniferous | 0.9 | 0.002915452 | 0.83 | 0.6 | 0.77 |
| Carboniferous | 1 | 0.002915452 | 0.83 | 0.6 | 0.77 |

| Period | Alpha Value | Lambda Value | Accuracy | AUROC | AUPR |
|---|---|---|---|---|---|
| Permian | 0 | 0.002915452 | 0.84 | 0.62 | 0.79 |
| Permian | 0.1 | 0.002915452 | 0.84 | 0.62 | 0.79 |
| Permian | 0.2 | 0.002915452 | 0.84 | 0.63 | 0.78 |
| Permian | 0.3 | 0.002915452 | 0.84 | 0.62 | 0.79 |
| Permian | 0.4 | 0.002915452 | 0.84 | 0.56 | 0.82 |
| Permian | 0.5 | 0.002915452 | 0.84 | 0.56 | 0.82 |
| Permian | 0.6 | 0.002915452 | 0.84 | 0.56 | 0.82 |
| Permian | 0.7 | 0.002915452 | 0.84 | 0.56 | 0.82 |
| Permian | 0.8 | 0.002915452 | 0.84 | 0.45 | 0.83 |
| Permian | 0.9 | 0.002915452 | 0.84 | 0.46 | 0.83 |
| Permian | 1 | 0.002915452 | 0.84 | 0.46 | 0.83 |

## Results /Discussion

In Stage One, we conducted logistic regression analyses for each stage of the Paleozoic era for each of the major phyla. The goal was to identify the phyla that have a predilection for extinction during each stage. As evident in Figure 2, we came across some impressive results as 23 of the 36 data points had a significant regression coefficient. 12 data points had a significantly greater extinction risk while 11 were significantly selected for survival. Among various phyla, coefficient values were high in magnitude, but no groups were consistently significant across all periods. However, specifically, Mollusca was generally selected for survival while Echinodermata was generally selected for extinction. For Brachiopoda, you may notice the relatively low coefficients in background periods but a significant extinction risk during the major extinction events in Devonian and Permian. This is in line with the understanding that these extinction events devastated Brachiopoda populations.

In Stage Two, we conducted a logistic regression analysis and binomial test to determine which natural traits incited greater evolutionary selection. The examined factors included ocean acidification resilience (buffering), predatory nature, body volume, length, surface area, motility, tiering, circulatory systems, and respiratory organ type. The largest coefficient value was around -3.9 which demonstrated a high susceptibility of organisms with open circulatory systems for extinction during the Cambrian period. The majority of the data points were significant,30 of 54. Among these, the only consistently insignificant characteristics were factors associated with body size. This shows that body size had little impact on the extinction risk of organisms. Surprisingly, two descriptors out of nine were significant across the board, circulatory systems, and buffering. Although the type of circulation and amount of buffering that was selected for extinction varied across the Paleozoic.

Finally, in Stage Three, we built machine learning models for each period of the Paleozoic using the ecological factors that we tested in Stage Two. In the Cambrian, the model with the highest accuracy was 92%. In chronological order, the remaining periods had a model with the highest accuracies of 83%, 91%, 79%, 83%, and 84%. As evident, the Devonian appeared to have the lowest accuracy. We believe that with increased data and testing out alternative models, this accuracy can increase greatly.

**Major Take-aways:**
- Extinction Risk is not uniform across both geologic history or across taxonomic groups
- Certain traits can act as indicators for higher extinction risk; however, these too vary across geologic history
- These traits can even be used to create relatively accurate predictive models.

**Future Research:**
For future developments, we believe that completing analysis across the rest of geologic history could allow us to identify patterns of how extinction risk changes for each phylum and each trait across every period in Earth's history. This could allow us to know how anoxic conditions affect extinction risk for each phyla/trait or how mass extinctions affect extinction risk for each phyla/trait.

For the machine learning model, we would like to test out Decision Tree or Random Forest Regression Models to predict the exact first and last appearance in geologic history for each genus. Finally, we will look into Building Neural Nets for this type of prediction

## Acknowledgements


Thank you to Stanford post-doctoral scholar Dr. Pedro Monarrez, Stanford Graduate Student Michael Pimentel, and Director of Education outreach, Dr. Jennifer Saltzmann for helping direct and lead the Stanford Earth Biodiversity program. Thanks also to Professor Jonathan Payne for helping make the program possible by hosting it within the Payne Paleobiology Lab.


## Citations


- Bush, A.M., Bambach, R.K., and Daley, G. 2007.Changes in theoretical ecospace utilization in marine fossil assemblages between the mid-Paleozoic and late Cenozoic. Paleobiology 33, 76–97.
- Heim, N. A., Bakshi, S., Buu, L., Chen, S., Heh, S., Jain, A., Noll, C., Patkar, A., Rizk, N., Sundararajan, S., Villante, I., Knope, M. L., Payne, J. L. 2020. Respiratory medium and circulatory anatomy constrain size evolution in marine macrofauna. Paleobiology, 1-16. doi:10.1017/pab.2020.16
- Novack-Gottshall, P. M. (2008). Ecosystem-wide body-size trends in Cambrian–Devonian marine invertebrate lineages. *Paleobiology, 34*(2), 210-228. doi:10.1666/0094-8373(2008)034[0210:ebticm]2.0.co;2
- N. A. Heim, Knope M. L., Schaal. E. K., Wang S. C., Payne, J. L. 2015. Cope's rule in the evolution of marine animals. Science 347, 867-870.
- Payne, J. L., Bush, A. M., Chang, E. T., Heim, N. A., Knope, M. L. and Pruss, S, B. 2016 Extinction intensity, selectivity and their combined macroevolutionary influence in the fossil record. Biol. Lett.1220160202 20160202
- Bush, A. M, Pruss, S. B. 2013. Theoretical ecospace for ecosystem paleobiology: energy, nutrients, biominerals, and macroevolution. In Ecosystem paleobiology and geobiology (eds AM Bush, S Pruss, JL Payne), pp. 1 – 20. New Haven, CT: The Paleontological Society.
- Knoll, A. H., Bambach, R. K., Payne, J. L., Pruss, S., Fischer, W.W. 2007 Paleophysiology and end-Permian mass extinction. Earth Planet. Sci. Lett. 256, 295–313. (doi:10.1016/j.epsl.2007.02.018
- Knope, M., Heim, N., Frishkoff, L. and Payne J. L.. Limited role of functional differentiation in early diversification of animals. Nat Commun 6, 6455 (2015).