Study of the origin of interstellar molecular complexity through explainable machine learning and statistical models

Marcos Martínez[1,3], Izaskun Jiménez-Serra[1], Raúl Guantes[2] and Jacobo Aguirre[1,3]

[1]Centro de Astrobiología (CAB), CSIC-INTA, Torrejón de Ardoz, Madrid, Spain
[2]Instituto 'Nicolás Cabrera', Facultad de Ciencias, Universidad Autónoma de Madrid, 28049 Cantoblanco, Madrid, Spain
[3]Grupo Interdisciplinar de Sistemas Complejos (GISC), Madrid, Spain

Despite the recent detection of an increasing number of complex organic molecules in outer space [1], the limits of the natural molecular complexity it can harbor are still unknown. NetWorld, a recently developed computational environment based on complex networks in interaction, has shown promising results in the study of the emergence of molecular complexity [2]. This computational framework simulates the evolution of abstract network structures that can join or be divided in different environments, producing an artificial chemistry that, despite its lack of direct connection with real chemistry, is able to reproduce important properties related to the origin of complex molecules with potential prebiotic interest.

In this work we use the NetWorld framework to study the factors that determine network abundance, with the ultimate aim of generalizing them to real molecular abundances in the interstellar medium. NetWorld acts as a virtual laboratory that can be used to generate the large volumes of abundance data required to train machine learning and statistical models, while the limited amount of observational data available is reserved to validate their predictions. In this way, we were able to identify properties of networks based on their structure and on the different pathways that lead to their formation that can be used to predict their abundance with considerable success (Figure 1). Furthermore, we found considerable agreement between the abundance relationships originally found in NetWorld and those observed in real astrochemical environments.

In summary, our results suggest that chemical formation paths or even simple topological properties of the 3D structures of real molecules can be used to obtain baselines for their abundances, helping to determine which molecules are likely to arise abiotically in the interstellar medium.

References

[1] B. A. McGuire, 2018 census of interstellar, circumstellar, extragalactic, protoplanetary disk, and exoplanetary molecules, ApJS 239, 17, 2018.

[2] M. García-Sánchez, I. Jiménez-Serra, F. Puente-Sánchez, and J. Aguirre, The emergence of interstellar molecular complexity explained by interacting networks, Proc. Natl. Acad. Sci. 19, e2119734119, 2022.

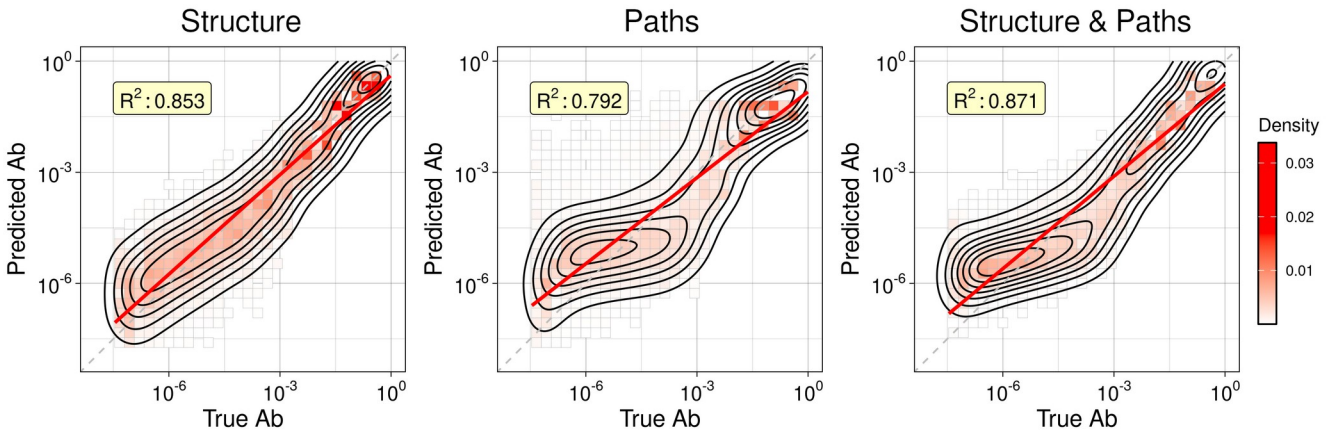

Figure 1: Correlation between true network abundance (True Ab) and model predictions (Predicted Ab) for the best machine learning models obtained in the NetWorld computational framework. This correlation is shown separately for the best-performing models using as input structural measures, path measures, or both types of measures. Regression lines and the joint density of true and predicted abundances are shown in red.