Paired residue prediction dependencies in ESM2

During a quick analysis of the ESM2 model for masked token prediction, we noticed that amino acid probability distributions of residues affect each other in a pattern that mirrors a protein's 3D contact map. But less so for the larger model sizes. Our question to you is, why?

Purpose

This notebook details a quick analysis we performed on the ESM2 (Evolutionary Scale Modeling) models [1] for masked token prediction. We stumbled upon a counterintuitive result related to the effect that masking one residue has on the distribution of another.

Before jumping into our results, let's first establish what masked token prediction is, how it works, and our consequent motivation for this analysis.

View the notebook

In the future, we hope to host notebook pubs directly on our publishing platform. Until that's possible, we'll create stubs like this with key metadata like the DOI, author roles, citation information, and an external link to the pub itself.

Contributors (A–Z)

- **Brae M. Bigge**: Supervision

- **Daniel Burns**: Critical Feedback, Formal Analysis

- **Keith Cheveralls**: Critical Feedback, Validation

- **Megan L. Hochstrasser**: Editing

- **Evan Kiefl**: Conceptualization, Formal Analysis, Investigation, Methodology, Software, Visualization, Writing

References

1. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. https://doi.org/10.1126/science.ade2574