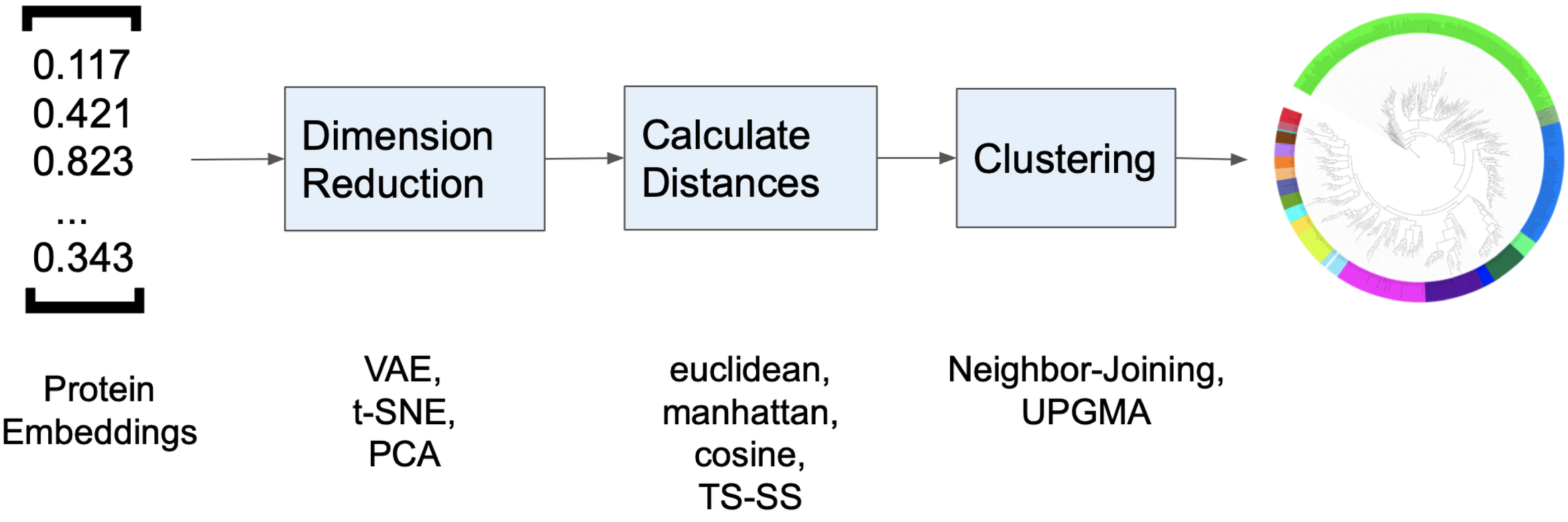
Protein Embeddings to Phylogeny

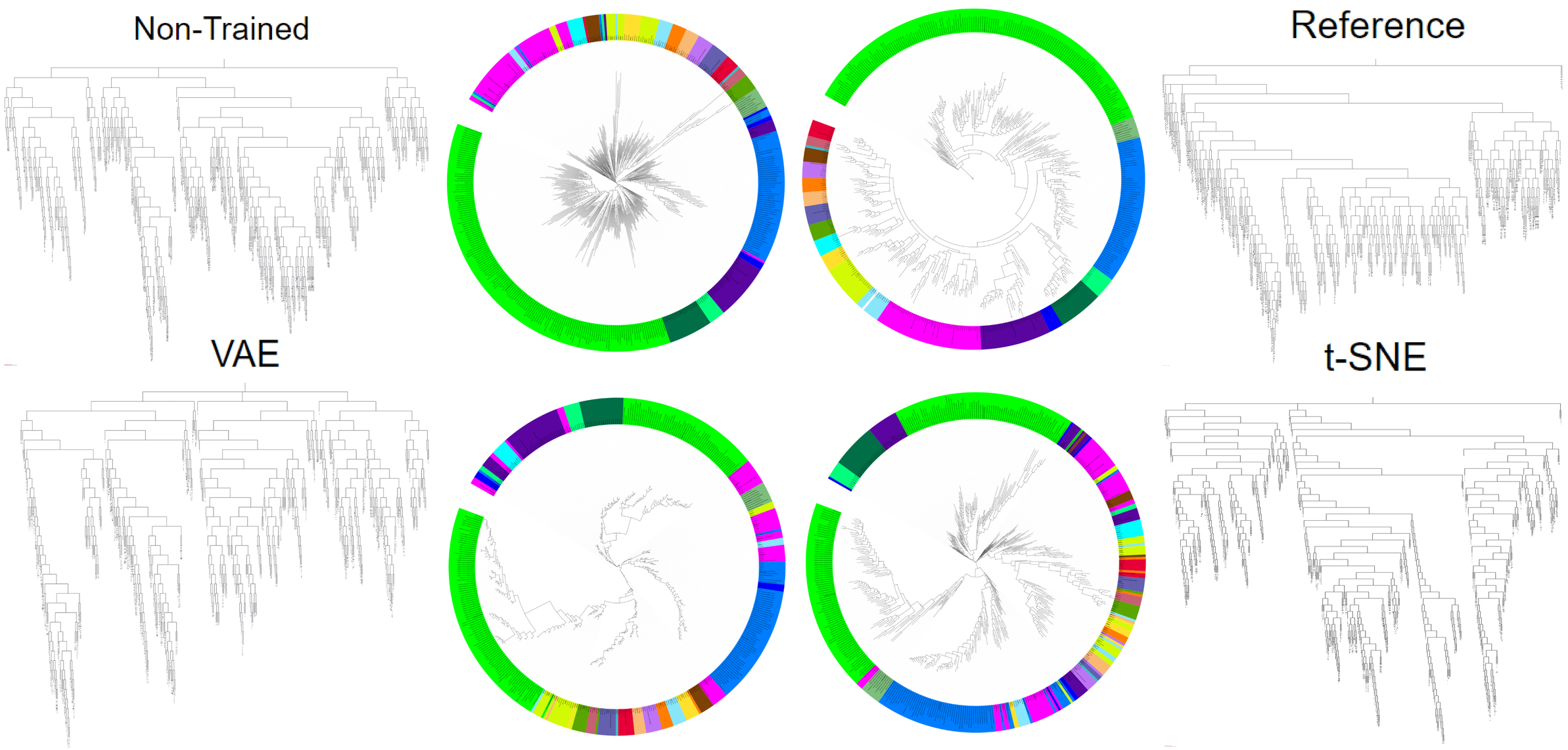
Exploring Evolution in Space

Adel Schmucklermann, Alexander Fastner
Kyra Erckert, Tobias Olenyi, Tobias Senoner, Ivan Koludarov, Burkhard Rost



-Abstract-

We explored the use of protein sequence embeddings generated with ProtT5 and Esm-2 to construct phylogenetic trees, aiming to reduce the time required compared to traditional methods. To filter out the noise and extract more information, we employed PCA and t-SNE, as well as a Variational Auto-Encoder (VAE), to reduce embedding dimensions. We then calculate various distance metrics and cluster them into trees using UPGMA and Neighbor-Joining.



-Results-

Shown are resulting Neighbor-Joining trees (Non-Trained tree calculated directly from distance of embeddings, our reference from MSA program, our VAE, our t-SNE) generated with iTOL.
We conclude that the resulting tree without prior training (non-trained) is the closest to our reference tree.

https://github.com/AlexanderFastner/Protein-Embeddings-to-Phylogenies