Toy dataset for the molecular recognition problem

Learning-based algorithms for protein interaction prediction are typically trained on datasets extracted from the Protein Data Bank and tested on the Docking Benchmark and CASP/CAPRI datasets. However, these datasets are large and impractical for fast prototyping and, unlike images or natural language, protein structures are not easily interpretable by humans, which makes it hard to analyse the advantages and drawbacks of any new algorithm.

In this paper we propose a simple, interpretable, small-scale dataset that can be used to select algorithms predicting protein-protein interactions. The aim is to provide the equivalent of the MNIST dataset [1] for the molecular recognition problem, formulated here as an energy-based shape recognition problem in two dimensions (2D). We hope this dataset will allow researchers to rapidly prototype new machine learning architectures for molecular recognition problems.

There are two broad categories of data available to infer the details of protein-protein interactions: (1) the structures of protein-protein complexes obtained from X-ray crystallography or electron microscopy and (2) fact-of-interaction data obtained using proteomics methods. In this work we cover both categories using one generating function, parameters of which could be learned independently from these two sources of data. To make the problem more easily tractable and interpretable, we formulate it in two dimensions and use a simple shape-based interaction potential similar to the one proposed by Katchalski-Katzir et al [2].

Additionally we propose baseline models that represent different approaches to the problem. For interaction pose prediction, we provide three baselines: a ResNet model, a brute-force learning model, and an energy-based learning model. For fact-of-interaction prediction, we provide a ResNet model and a brute-force model.

The ResNet baseline attempts to map two shapes either to the correct rotation and translation (for interaction pose prediction) or to the binary variable representing the fact of interaction.

The other two baselines are based on a deep convolutional network that learns protein representations and compute the energy of any conformation of two proteins using the correlation of the representation of the first protein with the translated and rotated representation of the second protein. The brute-force approach trains the representation model by sampling all rotations and translations while the energy-based learning approach uses Langevin dynamics for sampling.

We show that the ResNet model baseline fails to capture the essence of the problem: it overfits both the interaction pose and fact-of-interaction datasets and performs poorly on the test sets. The representation learning approaches, on the other hand, effectively solve the pose prediction problem (**Figure 1**). However, the brute-force approach is impractical in 3D and the energy-based approach does not appear immediately applicable to the fact-of-interaction dataset.

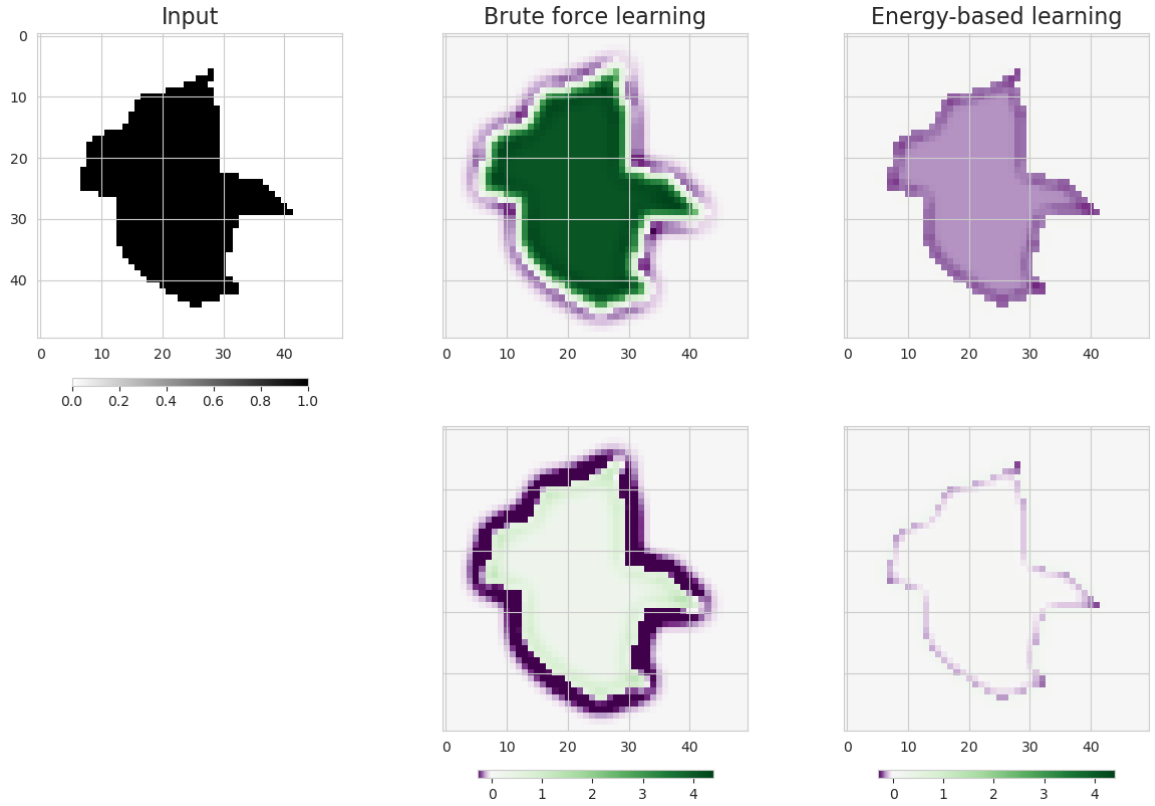

Figure 1. Learned shape representations. From the input image (top left), the brute-force learning algorithm computes the two-feature representation displayed in the middle column and the energy-based learning algorithm computes the representation displayed in the right column. Both representations allow the models to accurately predict the interaction pose of any two shapes with matching interfaces.

[1] LeCun et al, Proceedings of the IEEE, 86 (11), 2278–2324, 1998
[2] Proceedings of the National Academy of Sciences, 89 (6), 2195–2199, 1992