Learning on Molecular Graphs

Jay Shen
Department of Physics
University of Chicago
Chicago, IL 60637
jshe@uchicago.edu

Mark Lee
Deparment of Statistics
University of Chicago
Chicago, IL 60637
markyl@uchicago.edu

Abstract

Graphical deep learning has been experiencing a heyday of sorts. Despite being in its relative infancy, the paradigm of graphical neural networks (GNNs) has already had enormous success tackling tasks from weather forecasting to protein folding.

Computational chemistry has perhaps been the field where GNNs have recieved their warmest welcome. The natural interpretation of atoms as vertices and bonds as edges of a molecular graph has resulted in their widespread adoption for application anywhere from molecular dynamics simulation to automated design. The correspondence has gone both ways, however. Because problems such as drug discovery, materials design, and efficient simulation are so important, pure AI research into graphical learning has found a wealth of problems for the tackling.

The science of GNNs today draws upon existing approaches both from deep learning, such as convolutions and attention mechanisms, as well as from the theory of probabilistic graphical models, the inspiration for the message-passing operations crucial to the ways GNNs learn.

Here, we propose to take a step back. What concrete results about molecules can we derive from the lighter-weight, more rigorous methods of probabilistic graphical models? First, how we can represent molecules as data structures amenable for processing? As opposed to a medical diagnosis system, for example, where the set of random variables and their relationships remain for the most part constant, molecules are innumerably unique with vastly differing structures. Second, how can we perform inference on molecules? What conclusions can we draw and what significance might they have for the types of inquiries a chemist might make. Third, how does our approach lend itself to the types of deep learning that are the state of the art today?

With this in mind, we turn to the specifics of this study. As stated, the technical challenges will revolve around useful representation of dynamic molecular data as graphical models. The development of methods for doing this will form the bulk of our technical efforts, and might guide our inquiry so as to be feasible. The success of our approaches will be evaluated against empirical data from experiments, rule-based calculations, et cetera. For example, a property prediction task might be benchmarked by values computed in vitro. The data we need can be sourced from the wealth of both experimental and structure data available open source. For example, the ZINC19, QM9, and GDB13 datasets provide various traunches of molecular structure data, as well as some property measurements.

The timeline for this project is as follows: 3 weeks of research and preparation; 1 week for final implementation and evaluation; 1 week for synthesizing the report.

Preprint. Under review.