Using Molecules for Machine-Learning Chemistry via MoleculeNet

Adarsh Dave
ardave

Jake Parker
jlparker

Sarah Malle
smallepa

1 Project Proposal

Key technology problems in pharmacology, renewable energy, and agriculture could be solved with novel, high-performance materials and chemical compounds. Breakthroughs via physical experiment are often slow-to-arrive, leading to many researchers harnessing high-performance computing for so-called "rational materials design" via high-throughput simulation and "virtual screening" [1]. The corresponding increase in data availability has lead to a surge of machine-learning research.

However, chemistry-related machine-learning presents tricky issues. First, there is the limited availability of "ground-truth" experiment data and such data is also scattered and heterogeneous in format. Second, scientists often want to predict many labels at different scales, e.g. electronic structure properties up to physiological impacts. Third, molecules themselves are variable in dimension, connectivity, and conformers. [2].

MoleculeNet, presented by the Pande lab at Stanford last year, represents a benchmark for machine-learning using molecular inputs. Our goal is to understand and replicate their results, in which they used molecular representations to predict numerous chemical properties [2].

The first task is to recreate these benchmarks from the paper with physical chemistry and quantum mechanical data-sets - these results can be seen in Figs 12, 14, and 15 and Table 3 in Reference 2. We chose this data subset based on the research interests of our team.

By replicating Pande's work, we will learn many real-world machine learning methods , enumerated briefly below as tentative "project plan":

1. Data exploration and processing: how to split chosen data-sets for training, test, and validation (random vs stratified vs structured/grouping by molecular structure)
2. Featurization: ECFP, Coulomb Matrix, Symmetry, Graph Convolutions, Weave
3. Modeling: (Conventional) Logit/KRR, SVMs, RFs/Boosting; (Graph) Convolutional, DA, Weave, Message Passing, Deep Tensor, ANI
4. Parameter tuning: Gaussian process hyperparameter optimization, varying training size / multitasking for model performance [3]

If we can accomplish the above goals this month, we will move on to assess adding additional data-sets from outside the paper. We could expand the molecule data-set with inorganic compounds from crystallographic data-bases and assess performance on the same labeling tasks. We hope to potentially find "positive transfer learning" scenarios.

As noted by the references, we will use DeepChem - a Python library built atop TensorFlow by authors of References 2 and 3 - for most tasks, supplemented by scikit-learn for other machine-learning models. Databases mentioned above are publically available. Learning how to use DeepChem, as well as truly understanding everything Pande's team has done, will take time. We hope to have the project plan accomplished within a month. With faster progress, we will turn to the exploratory second task until project deadline.

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

References

[1] Ceder, G. & Persson, K. (2013) How Supercomputers Will Yield a Golden Age of Materials Science. *Scientific American*, access: `https://www.scientificamerican.com/article/how-supercomputers-will-yield-a-golden-age-of-materials-science/`

[2] Wu, Z. , Ramsundar, B. , Feinberg, E.N. , Gomes, J. , Geniesse, C. , Pappu, A.S. , Leswing, K. & Pande, V. (2017) MoleculeNet: a benchmark for molecular machine learning *Chem. Sci.* **9**(513)

[3] Ramsundar, B, Kearnes, S., Riley, P., Webster, D., Konerding, D. & Pande, V. (2015) Massively Multitask Networks for Drug Discovery *arXiv*, access: `https://arxiv.org/pdf/1502.02072.pdf`