PROJECT SUMMARY

Overview:

We propose to develop much faster ways to accurately evaluate the strength of non-bonded interactions, which govern biomolecular structure, the structure of organic crystals, and the binding of ligands by proteins. In previous work, we have produced methods and software to compute intermolecular interactions with of symmetry-adapted perturbation theory (SAPT) based on many-body perturbation theory or coupled-cluster theory. At the most accurate levels, SAPT is of coupled-cluster with perturbative triples [CCSD(T)] quality. Thanks to favorable error cancellation, even the lowest-level treatment, SAPT0, typically provides reliable interaction energies. As a bonus, SAPT provides each energy component separately: electrostatics, induction/polarization, London disperson interactions, and exchange-repulsion. In large, complex systems, even the less expensive SAPT methods are too computationally expensive to be practical. (1) We will speed up SAPT0 dramatically by replacing the rate-determining dispersion step with very fast estimates: (a) semi-empirical damped pairwise-atomic terms, with damping parameters carefully fitted to large training sets (previous parameters from the literature do not appear to perform well), or (b) estimates determined directly from machine learning (ML). (2) We will develop fast ML models of *all four* interaction energy components (electrostatics, dispersion, induction, exchange), based on large training data sets. (3) We will develop advanced physics-based intermolecular potentials, whose many parameters are obtained through ML. (4) We will increase the accuracy of low-order SAPT0, using delta-learning techniques to obtain a ML model of the difference between SAPT0 and "gold standard" CCSD(T).

Intellectual Merit:

Intermolecular interactions pose some particular challenges to machine learning: (1) selection of suitable training and validation data is less obvious, (2) standard descriptors appear non-ideal for intermolecular interactions, (3) we would like to retain the energy components as well as the total interaction energy, leading to a multi-task learning problem. Our proposal is aimed at overcoming these challenges. At the same time, it is not clear whether ML models of intermolecular interaction energies can outcompete high-quality, next-generation force-field model potentials. The biggest problem with the latter is that they can require too many parameters for their extension to general organic molecules to be practical; however, parameterization problem might be neatly solved by using ML to obtain the requisite parameters. We will pursue both pure-ML and physics+ML approaches in an internal competition, using common test sets, which will highlight the strength and weaknesses of both approaches. Finally, we have a unique capability to partition SAPT energies into pairwise-atomic contributions, through our A-SAPT method. Because Behler-Parrinello neural networks model properties on an atomic basis, A-SAPT data is *precisely* what we are attempting to learn. This should make training much easier and require much less data than indirectly training to the *sums* of atomic contributions in van der Waals dimers, and could be a significant innovation in applying ML to intermolecular interactions.

Broader Impacts of the Proposed Work:

Proposed methods, if successful, will be released as open-source software to aid other researchers. Having various levels of accuracy vs computational speed will enable a variety of applications. The faster and more accurate SAPT0 methods will be helpful in understanding longer-range contacts between ligands and proteins. The fastest methods will be suitable for rapid screening of ligand binding or candidate structures of organic crystals. Retaining the ability to compute energy components will provide an ad-

ditional level of insight and interpretability, which has proven very helpful in previous applications of SAPT. The PI will continue to organize summer bootcamps in data science, and will update his popular YouTube videos on quantum chemistry. The project will contribute to the training of postdocs, graduate students, and undergraduates in quantum chemistry methods, data science, machine learning, and software development.