PROJECT SUMMARY / ABSTRACT

This work seeks to advance quantitative methods for biomolecular design, especially for predicting biomolecular interactions, via a focused series of community blind prediction challenges. Physical methods for predicting binding free energies, or “free energy methods”, are poised to dramatically reshape early stage drug discovery, and are already finding applications in pharmaceutical lead optimization. However, performance is unreliable, the domain of applicability is limited, and failures in pharmaceutical applications are often hard to understand and fix. On the other hand, these methods can now typically predict a variety of simple physical properties such as solvation free energies or relative solubilities, though there is still clear room for improvement in accuracy. In recent years, blind prediction challenges have played a key role in driving innovations in prediction of physical properties and binding, especially in the form of the SAMPL series of challenges. Here, we will continue and extend SAMPL prediction challenges to include new physical properties, more complicated host-guest binding data, and application to biomolecular systems. Carefully selected systems and novel experimental data will provide challenges of gradually increasing complexity spanning between systems which are now tractable to those which are marginally out of reach of today’s methods but still slightly simpler than those covered by the Drug Design Data Resource (D3R) series of challenges on existing pharmaceutical data. We will work with D3R to run blind challenges on the data we generate and to ensure it is designed to maximally benefit the field.

In **Aim 1**, we will collect new measurements on partitioning, distribution, and protonation of drug-like compounds, in collaboration with partners in the pharmaceutical industry. In **Aim 2**, we leverage our expertise in host-guest binding to generate new data on host-guest binding in cucubiturils and deep cavity cavitands. And in **Aim 3**, we use high-throughput robotic experiments to generate new protein-ligand binding data of biological relevance. **Aim 4** focuses on using this data to run blind SAMPL challenges, motivating the community to test, understand, and improve these methods. We will also run reference calculations with the latest techniques.

This work will ensure the continued success of SAMPL challenges which have already driven considerable innovation in the field and been the focus of more than 90 different publications (each typically cited 5-50 times) since their inception around 2007, and will play a key role in driving the next several generations of improvements in computational techniques for molecular design. The research proposed here will lead to significant improvements in the predictive power of physical models for drug discovery, molecular design and the prediction of physical properties.