SPECIFIC AIMS

While computational techniques are widely used in pharmaceutical drug discovery, current generation technologies (such as docking) are unsuitable for true molecular design. Specifically, these techniques fail to to predict small molecule binding affinities to target and antitarget biomolecules with sufficient accuracy for many applications. While virtual screening techniques can improve enrichment, they lack the accuracy to guide molecular design. A new generation of physical techniques, alchemical free energy calculations, are poised to fill this void by providing quantitative, predictive tools useful in multiple stages of the drug discovery pipeline.

Recent success of alchemical methods in predicting accurate affinities has sparked considerable enthusiasm, but the domain of applicability of these techniques is currently highly limited; broad application and routine use will require further evaluation, refinement, and development. There is a vast gulf between targets within the domain of applicability and those which are outside it, frustrating progress. Bridging this gulf will require focused study of carefully selected systems of intermediate complexity. Without such a bridge, these techniques may encounter the same problems faced by docking and related techniques: routine failure without clear insights into why, and years to decades spent making small methodological modifications without dramatic improvements in predictive power.

We propose a carefully designed, crowdsourcing-based model to drive innovation to bridge this gap. Specifically, we will collect targeted experimental datasets that highlight and focus the community on specific modeling difficulties, fielding blind prediction Challenges to spur new, crowdsourced methodology innovation from the community. This model has been proven to drive progress in a given area, both inside the computational biology community [1–6], and for even more visible projects like Netflix [7] and the XPrizes [8–10]. Here, we design an extension of the Statistical Assessment of Modeling of Proteins and Ligand (SAMPL) series Challenges, which allow a wide variety of different approaches to compete on equal footing on the same data, fostering rapid innovation, quick recognition of the best new technologies, and dissemination of these to the community. The data we generate will allow several Challenges of differing difficulty, between systems tractable with current methodologies and the pharmaceutically-relevant drug targets featured in the NIH-funded D3R effort [11]. Central to this proposal is collecting targeted data and running blind Challenges to drive progress, but the data itself will have an even longer-term impact on the community. Our Aims are:

Aim 1. Collect new physical property datasets to assess accuracy and spur improvements in force fields and modeling of protonation states and tautomers.
We will develop new solution-phase datasets for druglike small molecules. These data can test critical aspects of small molecule modeling (including accounting for interactions and treatment of protonation/tautomeric state) and improve our ability to predict physical properties relevant to drug discovery in new regions of chemical space. We will initially focus on aqueous/nonpolar distribution coefficients and pK_a measurements, advancing to solubilities and membrane permeabilities, while using these data to drive improvements in the modeling of ligand interactions.

Aim 2. Measure affinities of drug-like compounds in supramolecular hosts to challenge quantitative models of binding in systems not plagued by major receptor sampling issues.
We will measure new host-guest binding free energies (using cucurbiturils and deep-cavity cavitands as hosts) to field binding Challenges with varying complexity between physical property prediction and protein-ligand binding. Host guest systems are some of the simplest cases of molecular recognition, and thus these binding data will drive improvements in modeling of simple binding systems with techniques of relevance to drug discovery.

Aim 3. Develop model protein-ligand systems that isolate specific modeling challenges of drug targets.
We will identify suitable biological protein-ligand model systems that isolate individual modeling challenges (selected to push the limits of physical techniques) and develop these for blind Challenges based on new protein-ligand affinity measurements. While the initial year will feature fragment binding to human serum albumin, subsequent Challenge systems will be selected using a novel informatics platform to focus on timely modeling issues.

Aim 4. Crowdsource innovation via community blind Challenges to advance biomolecular design.
The data collected in Aims 1–3 will drive annual SAMPL Challenges, allowing the field to test the latest methods and force fields to assess progress, compare them against one another head-to-head, and perform sensitivity analysis to learn how much different factors (protonation state, tautomer selection, solvent model, force field, sampling method, etc.) affect predictive power. Results will then feed back into improved treatment of these factors for subsequent Challenges, driving cycles of application, learning, and advancement – an approach proven to work in a variety of previous crowdsourcing approaches to science, as well as in prior iterations of SAMPL itself.

Overall, these SAMPL Challenges provide a catalyst to drive the next several generations of computational methodology innovation, development, and assessment for pharmaceutical drug discovery.