Active probabilistic modeling of biophysical simulations and experiments
Josh Fass — ACE Specific Aims — April 7, 2016

Human protein kinases are challenging drug targets. To design specific inhibitors of protein kinases, we would like to be able to predict the binding affinities of arbitrary small molecules to arbitrary human kinases. Doing this rigorously entails predicting free energies of binding, as well as identifying appropriate experimental conditions for expressing all human kinases, so that predicted affinities can be measured. A recurring, rate-limiting sub-task in these efforts is to **estimate integrals from expensive samples.** I aim to develop sample-efficient methods for three specific instances of this task, using shared formalism: (1) Estimate the **posterior distribution** of plausible molecular mechanics force field parameters, given observed physical properties; (2) Estimate the distribution of suitable conditions for **cell-free protein expression**, using the fewest experiments possible; (3) Estimate **equilibrium conformational distributions** of all protein kinases, making efficient use of computational effort.

I propose to adopt the framework of model-based "probabilistic integration," which poses integration as an inference problem: iteratively collect samples, build a probabilistic model of the integrand surface from those samples, and use the model to select maximally informative locations to collect more samples. This approach provides finite-sample estimates of the error, and guides the proposal of new samples to reduce this expected error. Algorithms constructed in this framework can also be extremely sample-efficient, achieving exponential rates of posterior contraction. However, many practical challenges remain.

Aim 1: Active estimation of parameters in molecular mechanics models
Atomistic simulations rely on accurate potential energy functions. These "force fields" depend on numerous free parameters. Typically these are chosen on the basis of chemical intuition, or optimized to predict physical properties in agreement with a set of experimental observations. Instead of producing a point-estimate of these parameters, I aim to estimate the full posterior distribution of plausible parameters for a given model. However, drawing samples from this posterior is extremely expensive. A workable Bayesian inference strategy would be valuable to: (1) quantify predictive uncertainty, (2) avoid "overfitting" to a single set of experimental data, (3) flexibly incorporate new experimental evidence, and eventually (4) perform model selection. I aim to accelerate Bayesian inference in this context by applying model-based integration. To do so, I will extend existing methods to the "likelihood-free" setting— since we can only approximate the likelihood by simulation.

Aim 2: Batch exploration of cell-free expression conditions
To perform high-throughput kinome-wide ligand-binding experiments, we need to be able to express many related proteins cheaply. Many human kinases are not easily expressible in *E. coli*, motivating interest in highly parallel cell-free expression systems. However, the range of effective reaction conditions for cell-free expression (including buffer and reactant concentrations, temperature, and incubation time) is unknown. Experiments cost time and money, so we would like to systematically find high-yield reaction conditions using the fewest possible experiments. This problem can also be cast in a probabilistic integration framework, by viewing the experimental yield function (that maps reaction conditions to expression levels) as an integrand we can noisily observe. I will extend existing methods to accommodate this large-batch design setting.

Aim 3: Adaptive modeling of kinase conformational distributions
To identify cryptic binding sites and compute binding free energies, we need to draw samples efficiently from the equilibrium conformational distribution of a solvated macromolecule. Thermostatted molecular dynamics (MD) is frequently used for this purpose. However, the MD propagator can be extremely slow-mixing, motivating interest in "adaptive sampling." Many adaptive sampling methods have been proposed, using uncertainties from a previously estimated Markov State Model (MSM) transition matrix to select initial conditions for subsequent rounds of MD. MSMs approximate a system's conformational kinetics by partitioning the system's configuration space and estimating transition rates between the partitions. Although MSM-based adaptive sampling methods appear to improve sampling efficiency in practice, they pose further challenges, and their convergence properties are unknown. It may be possible to overcome limitations of these methods by recasting them in the framework of probabilistic integration, for which strong convergence results have recently been proven. To bridge the gap between existing adaptive sampling methods and probabilistic integration, I aim to construct an extension of MSMs that represents uncertainty in the state decomposition as well.