Research Statement

My long-standing scientific interest is to understand the structure and dynamics of matter at the atomic/molecular level ("the jigglings and wigglings of atoms" from the wonderful Feynman *Lectures on Physics*), and their relationship to the functions and properties of our physical and biological macro-world (the spatial-temporal collective behavior). The knowledge domains span from X-ray macromolecular crystallography, molecular mechanics and dynamics, quantum chemistry to statistical mechanics and non-equilibrium and equilibrium thermodynamics. Since the underlying complexity makes most of the research problems analytically intractable, computational approaches have to be adopted (the so-called third science of computation/simulation after theory and experiment).

Technically, I would like to start with the study of molecular recognition because molecular recognition between biomolecules like proteins, DNAs, RNAs and small ligands is of central importance in biological and pharmacological processes like signal transduction, DNA transcription, enzymatic reaction and drug action. Whether two molecules bind to each other or not and how strong that binding is depend on the binding free energy difference. Therefore, predicting absolute and relative binding free energies of molecular associations is of great scientific and practical value. In fact, free energy calculation remains one of the most challenging issues in current computational sciences, even though great progress has been made and creative approaches have been developed over the past few decades.

In the following sections, I narrow my strategic interests into operational research plans in the areas of drug design, novel free energy methods and X-ray crystallography.

Structure-based/Computer-aided drug design

Structure-based drug design is based on a firm understanding of molecular recognition between binding site groups and interacting molecules and is a strategy that has become an integral part of modern drug discovery. Due to the recent volume and pace at which the 3-D structures of protein targets and their co-crystals have been made available, coupled with advances in computational tools, structure-based drug design has become a tool for lead generation as well as for optimization. I am interested in new computational methodologies including flexible, faster docking techniques like our continually-evolving AutoDock, virtual screening like our hierachical staged-screening and target/structure focused library design based on "click chemistry".

1) Inhibitor design against ATIC.

   This is an on-going collaboration with the Art Olson, Ian Wilson and Dale Boger groups at the Scripps Research Institute. ATIC (AICAR transformylase/IMP cyclohydrolase) catalyzes the last two steps in the *de novo* purine biosynthetic pathway to generate IMP. Both enzyme activities are potential inhibition targets for antineoplastic drug design because rapidly dividing cancer cells rely mainly on the *de*

novo synthesis of nucleotides rather than the more economical salvage pathway due to the requirement of significant amount of purines to sustain rapid growth. Besides their anti-cancer effect, the ATIC inhibitors also have anti-inflammatory and anti-diabetic functions due to the ensuing accumulation of *in vivo* AICAR substrate.

a) Inhibitor design against AICAR transformylase.

The folate-dependent AICAR Tfase adds the formyl group onto the 5-amino position of aminoimidazole carboxamide ribonucleotide (AICAR) to form the stable intermediate 5-formyl aminoimidazole carboxamide ribonucleotide (FAICAR). We have developed a new hierachical virtual screening protocol to search novel non-folate leads against AICAR Tfase by using AutoDock. The first pass is used to discover the appropriate scaffolds through focusing on compound structural diversity. At this stage, utilization or construction of a structurally diverse library is the key. The second pass is to uncover the best analogues of the compounds discovered in the first pass by combining docking with similarity searching. It is highly successful in our case as nineteen inhibitors from the NCI-3D database have been discovered.

Our future work on AICAR Tfase would focus on: i) Combinatorial library design focused on the privileged scaffolds and selection of the more potent and selective inhibitors through detailed docking and free energy simulation from the constructed focus library. ii) Fragment-tethering inhibitor design. A fragment library docking screening against AICAR Tfase active site has been done. The linker design will be the next step.

b) Inhibitor design against IMP cyclohydrolase.

IMP cyclohydrolase catalyzes the last step of the *de novo* purine biosynthetic pathway by converting FAICAR into IMP. A binding free energy analysis via MM-PBSA with AMBER molecular dynamics has been done to an IMP cyclohydrolase/nucleoside transition-state analog complex. The simulated binding free energy closely matches the experimental value. The analysis reveals the detailed contributions to the binding from the van der waals and electrostatic interactions between enzyme and TS analog, their desolvation and the entropic gain through the enzyme internal degrees of freedom.

Our future work on IMP cyclohydrolase would focus on: i) A hierachical virtual screening to discover novel IMP cyclohydrolase inhibitors. ii) A hybrid QM/MM (quantum mechanical/molecular mechanical) molecular dynamics to simulate the free energy profiles along the reaction coordinates to study the enzymatic reaction mechanism and to help design transition-state analogs in addition to the current sulfonyl analogs.

2) Activator design against AMPK/GK.

This is a new docking/screening project I would like to explore. Diabetes mellitus is characterized by elevated levels of blood glucose and fatty acids, and reduced insulin sensitivity. The disease now affects 3-5% of the population in Westernized countries, and the treatment of diabetic complications accounts for almost 10% of health care costs in both Western Europe and North America. The incidence of Type II diabetes, already at near epidemic proportions, is projected to double by 2025.

We are going to focus on the design of the allosteric activators of two very important anti-diabetic targets AMPK and GK. Allosteric modulators could offer several advantages over orthosteric inhibitors, including greater selectivity and saturability of their effect.

a) Activator design against AMPK.

AMPK (AMP-activated protein kinase), the master metabolic regulator, is a metabolic-stress-sensing protein kinase that regulates metabolism in response to energy demand and supply by directly phosphorylating rate-limiting enzymes in metabolic pathways as well as controlling gene expression. Activation of AMPK has many beneficial health effects, like promoting glucose uptake and fatty acid oxidation, reducing glucose production and lipid synthesis, increasing insulin sensitivity and protecting pancreatic β-cells, to name just a few.
AMPK is a heterotrimer containing α-, β- and γ-subunits, each of which has at least two isoforms. The α-subunit is the kinase catalytic unit; the β-subunit, a glycogen binding domain; the γ-subunit the AMP-binding allosteric domain which contain four highly conserved CBS (named after cystathionine β synthase) domains.
Our purpose is to design isoform-selective activators of AMPK. Before doing virtual screening and library design, we will focus our attention to determine the three dimensional structure of AMPK either via X-ray crystallography or homology modeling/protein-protein docking/molecular dynamics computational techniques.

b) Activator design against GK.

Glucokinase (GK) plays a key role in whole-body glucose homeostasis by catalyzing the phosphorylation of glucose in cells that express this enzyme, such as pancreatic β cells and hepatocytes. Experiments show that glucokinase activators (GKAs) augment both hepatic glucose metabolism and glucose-induced insulin secretion from isolated rodent pancreatic islets. In Type II diabetes rodent models, GKAs lower blood glucose levels and increase hepatic glucose uptake.
Our purpose is to design potent and selective GKAs by either discovering novel leads or improving the existing GKAs, focusing on the allosteric site of the active GK structures.

3) Collaborations are welcome.

I like to collaborate with other theoretical/computational and experimental groups to do interesting research on existing or new projects collectively, to understand things from different angles and to integrate our knowledge together like I have been doing for years.

Theoretical/Computational approaches to free energy calculation

The following is the areas that I think new, novel free energy methods could emerge.

1) "Fast-growth" methods based on the Jarzynski equality.

The remarkable and elegant Jarzynski equality, $e^{-\Delta F/kT} = < e^{-W/kT} >$, relates the equilibrium free energy difference to the probability distribution of non-equilibrium work values. In theory, the identity stands no matter how far away the system is perturbed from equilibrium, so a "fast-growth" phase-switching methodology can be developed. In practice, the switching has to be slow and work distribution has to be close to Gaussian because the average of the exponential work is dominated by the trajectories corresponding to small work values that arise relatively rarely. The flip side of this means great research opportunities to develop new computational methodologies exist. As for configuration sampling, modified molecular dynamics appears to be suitable, steered MD, for example, is one of the choices. New sampling methods based on both Monte Carlo and molecular dynamics may be highly desired. Overall, exciting new research avenues are waiting to be made in this pristine area.

2) "Phase-mapping" methods based on transformation construction.

Traditional free energy perturbation converges poorly because in most situations there is little overlap in configuration space between the initial and final state ensembles. One way to increase convergence is to sample from non-canonical distribution plus re-weighting scheme such as umbrella sampling and weighted histogram analysis method. A newer way is to carefully construct an invertible transformation under which the initial state ensemble gets mapped to a new ensemble which overlaps significantly with the final state ensemble, thus speeding up the convergence. Such a transformation can be linear spatial transformation, non-linear metric scaling, Fourier space mapping based on collective normal mode moves, etc. Indeed, creative ideas and implementations can be realized here. Finally, combining extended sampling and ensemble transformation can be used to further speed up the convergence.

3) "End-point evaluation" methods based on MM-PB(GB)SA.

Since the free energy is a state variable, computational efficiency can be enhanced by evaluating only the initial and final thermodynamic states. MM-PB(GB)SA combines molecular mechanics evaluation and solvation evaluation through Poisson-Boltzmann or generalized Born models. This methodology is still a work in progress despite widespread usage because a) The MM part requires very accurate force fields, so developing more sophisticated FF is needed; b) Even though electrostatic solvation

free energy calculation via PB is more accurate than GB, development of better GB model(s) is always needed because GB is much faster than PB and free energy decomposition analysis can be done with GB but not PB. One way to improve GB is to expand the solvation atom types and to parameterize them with LGA (Lamarckian genetic algorithm) against existing experimental values. c) Better entropy calculation methods are needed (current method is based on the assumption of the decoupling of translation, rotation and internal entropies and normal mode approximation with regard to internal vibrations). Even within the framework of normal mode calculation, proper solvation consideration is surely needed.

X-ray macromolecular crystallography

X-ray macromolecular crystallography is a very powerful experimental technique to unlock the three-dimensional atomic arrangements of biomolecules. Over the past decade or so, I have been fortunate to experience and grow with the burst of technological breakthroughs of modern synchrotron radiation, freezing techniques, CCD detectors, anomalous scattering signal collection in hardware, and Bayesian data filtering, substructure solution, multi-resolution density modification, molecular dynamics coupled refinement in computational methodologies. I would like to continue to utilize this technology to solve interesting structures and to facilitate my theoretical/computational endeavor, and even like to develop computational methods to relate crystallographic B-factor values to biomolecular dynamic ensemble, at least for the very high-resolution structures.