The word from on high...

> ...the development of new mathematical approaches for an integrated computational and experimental pipeline for systematic improvment of biomolecular energy functions and protein design methodology...
>
> ...The design process consists of choosing a new protein structure or function (binding, catalysis, etc), computing sequences predicted to have as their lowest energy state the desired structure and/or function, and then manufacturing and testing the designs...
>
> ...Third, in the design computations, how can the energy of the designed systems most effectively be minimized so that the feedback is on actual ground states or near ground states...

TO ADDRESS: search and diversity

1 Protein Design Methodology

1.1 Fitness

The goal of protein design is to produce a sequence of amino acids that will fold into a protein that performs a desired function. The protein designer defines a function $Fitness(S, X)$ mapping a sequence S along with a three dimensional structure X to a scalar quantity assessing fitness for a desired purpose. We attempt to optimize $Fitness(S, X)$ subject to the constraint that sequence S must actually fold into structure X. In order to define reasonable fitness functions and assess the folding of S, we employ a physical model of protein energetics to define an energy function $Energy(S, X)$ mapping a sequence-structure pair to an estimate of the free energy of the system. Desired fitness is often closely related to energy. In the simple case of optimizing for stability, $Energy(S, X)$ may be used directly as a fitness function. To design a protein P that will bind a target T, an appropriate fitness function could be the energy delta upon binding: $Fitness(S, X) = Energy(S, XP) + Energy(XT) - Energy(S, X)$, where S is the sequence of P and X is a combined structure of P docked to T. In addition to optimizing fitness for purpose, we must also consider that sequence S may not fold into the desired structure X. Thus, we must optimize $Fitness(S, X)$ subject to the constraint that X is the minimum of $Energy(S, X)$ given S.

1.2 Independent Variables

Protein design is technically challenging due to both scale and a mixed continuous/discrete character of the involved problems. Given an alphabet of the twenty natural amino-acids, the number of possible sequences S is 20^N. The space of possible structures X is comparably large but consisting of continuous degrees of freedom, comprising several types with different properties: amino acid backbone (Φ^{3N}), amino acid side-chain ($\chi^{\approx 2N}$), and rigid-body ($\Re^6$). Φ^{3N} arise from the homogeneous backbone portion of amino acid residues containing three rotate-able bonds common across most amino acids. $\chi^{\approx 2N}$ arise from the heterogeneous portion of amino acid residues, the contain up to four rotatable chi-angles per residue with diverse structural characteristics. $\Re^6$ arise from interactions between kinematically disconnected bodies that could be protein chains, small molecules, nanoparticles, surfaces, et cetera. Φ^{3N} and $\Re^6$ are independent of sequence, while $\chi^{\approx 2N}$ are only well-defined for a particular sequence. A typical calculation designing a small 60 amino acid protein to bind a target involves considering a 180 dimensional space of Φ^{3N} combined with a six dimensional rigid body space, each with 10^78 possible sequences, each with a roughly 120 dimensional spaces $\chi^{\approx 2N}$. The scope of this problem space is daunting (REF Levinthal paradox), but allowable configurations are highly constrained. Φ^{3N} are stereotyped by the formation of alpha helix and beta sheet secondary structures, and the tendency for adjacent sections of the protein chain to form locally optimal substructures. Global sampling of Φ^{3N} is preformed in coordinated fashion using short structural fragments harvested from structures of natural proteins (REF fragments). $\chi^{\approx 2N}$ are stereotyped by the detailed structure of their respective amino acid side chains, which occur in natural proteins within discrete energy wells known as rotamers (REF dunbrack). Global optimization of CHI space can thus be incorporated into discrete sequence optimization by expanding the amino acid alphabet to include multiple letters per amino acid, one per rotameric energy well. Unlike Φ^{3N} and CHI, $\Re^6$ are not stereotyped in the same manner as Φ^{3N} and $\chi^{\approx 2N}$ both because the extent of each dimension is higher, and because the rigid body relationships they capture are directly intermediated by diverse side chain interactions. Thus, $\Re^6$ contribute disproportionately more complexity to the protein design problem than a naive accounting of dimensionality would indicate.

1.3 Phases of Search/Optimization

Current optimization methodology for protein design makes extensive use of monte carlo search and optimization in various forms along with continuous minimization. Each candidate sequence-structure pair is generated in three phases.

Phase 1, Φ^{3N} and $\Re^6$ are optimized using a smoothed, sequence independent version of the fitness function. For de novo design of protein folds we focus on Φ^{3N}, using monte carlo sampling of local structural fragments is used. For binder/enzyme design, Φ^{3N} are set from predefined scaffold proteins, and $\Re^6$ can be handled with a variety of docking technologies (section 1.4)

Phase 2, atomic detail is added by combinatorially optimizing sequence and $\chi^{\approx 2N}$ using monte carlo or in conjunction with a fitness function (section 1.1), given Φ^{3N} and $\Re^6$ positions from Phase 1.

Phase 3, local continuous minimization of all continuous Φ^{3N}, $\Re^6$, and $\chi^{\approx 2N}$ is performed using [LBFGS or similar techniques that can handle high dimensional problems, taking advantage of gradient information available from our fitness functions and physical model, ? REF?].

Phases 2 and 3 may be cycled several times in a single trajectory. As cycles progress, the fitness function is adjusted by relaxing non-physical fitness constraints such that the final round is typically pure energy minimization. This process is repeated over many independent trials, producing a large number of minimal energy sequence-structure pairs predicted to be fit for desired purpose. When feasible, the most promising of these are subjected to a more computationally intensive structure prediction technique called forward folding (REF?) to better assess whether the designs are indeed globally minimum energy given their sequence.

1.4 Proposed Work for Design Phase 1

We propose to dramatically improve Phase 1 of our design process through the use of a seamless multi-scale model we call a rotamer interaction field (RIF). This model employs hierarchy of bounding functions which estimate the best possible fitness achievable within an ensemble of Φ^{3N}

$\Re^6$ position of a given radius. The bounding functions use spatial indexing techniques (REF jorge paper) to rapidly look up precomputed locally optimal solutions to small subproblems of the full sequence/CHI optimization. The product of these sub-solutions is not generally a feasible solution, but provides a bound on the maximum achievable fitness. In preliminary work, we have applied these bounds to binder design using a nesting hierarchical decomposition of $\Re^6$ space to perform a branch and bound search called RIFdock. We are in the process of conducting an extensive computational and experimental benchmark of RIFdock. By computational metrics, designs produced with RIFdock are significantly better than those produced with three other previous methods. RIFdock has also produced working binders to IL17 with far better binding affinity than has ever been achieved with purely computational methods (Franziska et. al., unpublished). We propose to collaborate with the computer science department to improve RIF through the use of tighter, more sophisticated bounds, such as those possible using cost function networks (REF THOMAS PAPER), and to expand the RIFdock search technique to Φ^{3N} as well as $\Re^6$.

1.5 Proposed Work for Design Phase 2

We propose to dramatically improve Phase 2 of our design process by expanding the class of fitness functions that can be efficiently optimized using combinatorial optimization techniques from computer science (REFs SAT, CFN, CST, MIP etc). Such techniques have been explored for protein design by our group and others (REFs thomas, ALF), but have not replaced the monte-carlo search used in rosetta because rigorously optimizing the fitness function is not worth the extra computational cost; designs must be exact energy minima for their sequence, but fitness must only be good enough for the desired purpose. However, the current monte carlo optimization technique is limited to objective functions containing single and pairwise terms. This has proven sufficient to design hydrophobic protein cores and interfaces, but cannot produce fully satisfied polar networks like those observed in natural proteins. Such networks increase the specificity and solubility of designed proteins, greatly increasing the likelihood both that they are globally minimum energy for their sequence, and that they will function in the designed way. We have developed a very successful technique called HBNet (REF scotts paper) that produces native-like polar networks using an extremely computationally demanding brute force search. In collaboration with members of the UW computer science dept, we have recently demonstrated that polar networks can be designed orders of magnitude faster than brute force using any of a variety of classes of combinatorial solvers including partial weighted MaxSAT (REF), mixed integer programming (REF), and cost function networks (REF). Cost function networks are particularly promising, as they are also the leading method for exact solution of the general sequence optimization problem in Phase 2. We propose to further explore the use of these solvers to design HBNets, and to integrate these solvers into Phase 2 of our design process so that HBNet design, and other complex constraints, can be incorporated directly into fitness functions used in protein design.

The above described advances in methodology will enhance blah blah.

2 Quantitative and Interpretable Machine Learning

The computation and characterization of complex and multi-scale biophysics (CMSBP) phenomenon is at the forefront of modern mathematical and scientific exploration. Neuronal networks of the brain, large-scale biochemical kinetics, cell network dynamics, and gene regulatory networks, for instance, remain some of the most difficult and challenging problems for accurate quantification and future state prediction. There are promising indicators that these high-dimensional, multi-scale problems may be tractable with emerging mathematical techniques. These techniques aim to exploit the dominant low-rank structures (features) that are typically exhibited. Indeed, much of the success of reduced order models (ROMs) rests on the ability to identify and take advantage of patterns and features in high-dimensional data. These low-dimensional patterns are often identified using dimensionality reduction techniques [1] such as the proper orthogonal decomposition (POD) [2, 3, 4] or more recently via dynamic mode decomposition (DMD) [5, 6, 7, 8]. In addition to advances in dimensionality reduction, key developments in optimization, compression, and the geometry of sparse vectors in high-dimensional spaces are providing powerful new techniques to obtain approximate solutions to NP-hard, combinatorially difficult problems in scaleable convex optimization architectures [9]. For example, compressed sensing [10, 11, 12] provides convex algorithms to solve the combinatorial sparse signal reconstruction problem with high probability. Ideas from compressed sensing have been used to determine the optimal sensing locations for categorical decisions based on high-dimensional data [13], including in fluid dynamics [14, 15]. Recently, compressed sensing, sparsity-promoting algorithms such as the lasso regression [16, 17, 18], and machine learning have been increasingly applied to characterize and control dynamical systems [19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31]. These techniques have been effective in modeling high dimensional fluid systems using POD [32] and DMD [33, 34, 35, 36]. Information criteria [37, 38, 39] has also been leveraged for the sparse identification of nonlinear dynamics [40], as in [41, 42]. For CMSBP, these advances are insufficient for providing guidelines for building accurate ROMs. Indeed, if the various multi-scale, spatio-temporal features of the biological data are not separated, then an artificially high-rank for the dynamics is exhibited. In contrast, if time scales are separated and their individual low-rank embeddings constructed, then the machine learning and sparse sampling can once again be combined to great effect. The multi-resolution dynamic mode decomposition (mrDMD) [43] is an ideal tool for separating multiscale data, thus allowing us to more readily capitalize on the recent innovations in ROMs and model inference techniques. Thus key advances in a number of fields are fundamentally changing our approach to the acquisition, analysis, and model building from data acquired in multi-scale systems. Specific to this proposal are (i) advances in DMD which can be used to disentangle and exploit spatio-temporal patterns in high-dimensional data, and (ii) model identification methods which can be used to construct optimal, data-driven models at different temporal scales through DMD, Koopman theory and/or sparse identification of nonlinear dynamics [44]. Importantly, the DMD architecture allows for a mathematical strategy for connecting and/or inferring different spatio-temporal scales. Reduced order models (ROMs) provide a transformative paradigm for simulating high-dimensional systems of equations which are often derived from discretizing partial differential equations (PDEs). Specifically, the goal of ROMs is to replace simulations of an n -dimensional (n 1) system of differential equations bfu = F (**u** ,) **with an r -dimensional surrogate model bfa = f (a ,) where r n and the parameter vector represents an explicit parametric dependence. The reduction in dimension is accomplished by projecting to an optimalsubspace , the proper orthogonal decomposition (POD) modes, which is computed from a singular value decomposition of snapshots of the high-dimensional system u . Despite the many successes of the ROM** method, critical challenges remain. In this proposal, we address two of these open challenges: (i) a data-driven approach for discovering the governing equations $F(u, \ )$ when time-series measurements of the system alone are available, and (ii) an efficient method for building ROMs that accounts for the parametric dependence of the governing equations on (t). By integrating sparse sampling methods with recent regression techniques for the discovery of dynamical systems, an online and non-intrusive method is demonstrated for the discovery of parametric systems which is capable rapid model updates without recourse to the full high-dimensional system. Indeed, low-rank ROMs are identified and updated directly from time series data while concurrently inferring the governing spatio-temporal equations responsible for generating the data. The traditional ROM architecture assumes that the underlying evolution dynamics $u = F(u, \ )$ is known. However, in a majority of applications across the biological sciences, the governing equations are unknown or only partially known. Or alternatively, observed low-rank dynamics occurs due to microscale interactions which generate emergent macroscale properties. Thus there exists a variety of white-, gray- and black-box modeling strategies for dealing with full, partial or no knowledge of the governing system respectively. Emerging data-driven strategies are giving rise to new mathematical architectures that can learn white-box models from time series measurements of a given system, i.e. the correct governing model can be discovered. These data-driven methods, which are rooted in sparse regression, have been proven to be successful for discovering a variety of differential and partial differential equations. Further, the discovered models can be cross-validated using well established information criteria methods from model selection. In the innovations to be developed here, we discover the governing low-dimensional ROM directly from data, thus bypassing any recourse to the high-dimensional system. Moreover, the method infers the governing high-dimensional PDE, thus providing critical insight into the spatio-temporal evolution dynamics. Specifically, we consider a data-driven method whereby low-rank ROMs can be directly constructed from data alone using SINDy and PDE-FIND. The methods can be enacted in an online and non-intrusive fashion using randomized (sparse) sampling. The method is ideal for parametric systems, requiring rapid model updates without recourse to the full high-dimensional system. Moreover, the method discovers the high-dimensional, white-box PDE model responsible for generating the spatio-temporal data. Parametric dependencies of the underlying PDE compromise not only the PDE and ROM discovery process, but also quickly invalidates a given ROM. For such systems, the low-rank embedding subspace requires frequent updating since the ROM model is based upon accurately projecting into this subspace through inner products. Recent efforts have attempted to address parametric dependencies by the construction of libraries of POD modes which are valid across a wide range of parameters. Thus POD modes and ROM models can be quickly switched out as appropriate in an *online manner. In the method proposed here, a new ROM can be quickly computed without recourse to an* offline training stage (POD libraries) or recourse to the original high-dimensional system. Moreover, our method can manage parametric dependencies which are

beyond the ability of current methods to handle. Figure demonstrates three prototypical parametric dependencies: (i) a PDE model bfu = F(u) whose parameters change at fixed points in time, (b) a PDE model u = F(u, (t)) that depends continuously on the parameter (t), and (c) a system where the underlying PDE changes in time from u = F(u) to u = G(u). Although methods for handling the first case have been developed, (ii) and (iii) remain exceptionally challenging. ROM discovery provides a principledapproach to efficiently handling all three parametric cases in an ***online fashion without recourse to the high-dimensional system.***