Analysis of Biological and Chemical Systems Using Information Theoretic Approximations

by

Bracken Matheny King

B.S. Biomedical Engineering
Washington University (2004)

Submitted to the Department of Biological Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2010

Author ....
Department of Biological Engineering
April 30, 2010

Certified by..
Bruce Tidor
Professor of Biological Engineering and Computer Science
Thesis Supervisor

Accepted by..
Darrell J. Irvine
Associate Professor of Materials Science & Biological Engineering
Chairman, Department Committee on Graduate Theses

Thesis Committee

Accepted by . . . .

Douglas A. Lauffenburger

Ford Professor of Bioengineering, Chemical Engineering, and Biology

Chairman of Thesis Committee

Accepted by . . . .

Bruce Tidor

Professor of Biological Engineering and Computer Science

Thesis Supervisor

Accepted by . . . .

K. Dane Wittrup

Carbon P. Dubbs Professor of Chemical Engineering and Bioengineering

Thesis Committee Member

Analysis of Biological and Chemical Systems Using Information Theoretic Approximations

by

Bracken Matheny King

Submitted to the Department of Biological Engineering
on April 30, 2010, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy

Abstract

The identification and quantification of high-dimensional relationships is a major challenge in the analysis of both biological and chemical systems. To address this challenge, a variety of experimental and computational tools have been developed to generate multivariate samples from these systems. Information theory provides a general framework for the analysis of such data, but for many applications, the large sample sizes needed to reliably compute high-dimensional information theoretic statistics are not available. In this thesis we develop, validate, and apply a novel framework for approximating high-dimensional information theoretic statistics using associated terms of arbitrarily low order. For a variety of synthetic, biological, and chemical systems, we find that these low-order approximations provide good estimates of higher-order multivariate relationships, while dramatically reducing the number of samples needed to reach convergence. We apply the framework to the analysis of multiple biological systems, including a phospho-proteomic data set in which we identify a subset of phospho-peptides that is maximally informative of cellular response (migration and proliferation) across multiple conditions (varying EGF or heregulin stimulation, and HER2 expression). This subset is shown to produce statistical models with superior performance to those built with subsets of similar size. We also employ the framework to extract configurational entropies from molecular dynamics simulations of a series of small molecules, demonstrating improved convergence relative to existing methods. As these disparate applications highlight, our framework enables the use of general information theoretic phrasings even in systems where data quantities preclude direct estimation of the high-order statistics. Furthermore, because the framework provides a hierarchy of approximations of increasing order, as data collection and analysis techniques improve, the method extends to generate more accurate results, while maintaining the same underlying theory.

Thesis Supervisor: Bruce Tidor
Title: Professor of Biological Engineering and Computer Science

Acknowledgments

I would first like to thank Bruce for his mentoring over the years, particularly in setting strong standards for the phrasing, execution, and analysis of principled scientific research. I am extremely grateful for his encouragement in pursing the information theoretic work at the core of this thesis, while constantly examining ways to incorporate it with relevant biological and chemical questions.

Bruce has also collected an outstanding group of lab members, all of whom deserve my thanks. In particular, Jared Toettcher provided invaluable help during the formal derivation of MIST, and throughout the course of this thesis. The minimum spanning tree formulation was first suggested by Brian Joughin and Mala Radhakrishnan. Shaun Lippow and Michael Altman were excellent mentors during my first years in lab, and were key contributors to the lab software. Nate Silver and Jay Bardhan played crucial roles in adapting MIST to the analysis of configurational entropies. Josh Apgar and Dave Huggins also contributed considerably to much of my work.

The members of my thesis committee, Doug Lauffenburger and Dane Wittrup, have played provided crucial insight and perspective in my time at MIT. The information theoretic approximations at the core of this thesis were first developed in conjunction with Paul Kopesky as part of a course taught by Doug and Pete Dedon.

I have also been fortunate to have a number of excellent collaborators. Ben Cosgrove, Kristen Naegle, Maya Hasan, Ericka Noonan, Langdon Martin and their advisors, Doug Lauffenburger, Forest White, Linda Griffith, Leona Samson, and Barbara Imperiali were instrumental in applications of MIST to a variety of experimental and computational problems. Collaborations with DuPont and my internship at Merck also provided an important complement to my thesis research.

Finally, my family deserve more thanks than can possibly fit on the rest of this page for their love and support before, during, and (presumably) after my graduate studies. My appreciation for science, research, and critical thinking was instilled at as young an age as possible by my dad. My wife, Rachel, has also been a constant source of encouragement and love, particularly over the last few months of my thesis.

Contents

Chapter 1

Introduction

For much of scientific history, the examination of biological systems was largely focused on the isolation and direct characterization of individual molecular species or interactions between small sets of species. This mode of investigation proved quite successful, and over time, complex networks of interactions were seen to develop by integrating the results of many, many experiments of relatively small scope. Over the past few decades, remarkable technological advances have provided the ability to interrogate these same biological systems on a near global scale for many classes of molecular species, including mRNA, proteins, and metabolites [34, 85]. The data collected from these studies have also highlighted the multivariate nature of biological systems [43, 38].

Given this increasingly popular systems-level view of biology, a growing number of data sets have been, and continue to be, collected to identify and characterize the networks dictated by the molecular interactions. These data sets track species of interest in the context of multiple stimulation conditions, and/or cellular states. Data representing various cellular responses, such as the level of migration, proliferation, differentiation, or apoptosis may also be measured under the same set of experimental conditions. In general, these data sets represent the response of large numbers of signals across a relatively modest number of experimental conditions. From these data, the goal is to identify and quantify the multivariate relationships between the measured species, and to understand how these potentially complex relationships lead to cellular responses. While mechanistic understanding of such interactions is the ultimate goal, much of the analysis to date has focused on statistical modeling, as it is well suited for the types of data that can be collected on a system-wide scale [38, 46, 79]. Despite a variety of important advances, the development of techniques for the analysis of such data remains an active area of research.

The characterization of the statistical relationships between large numbers of variables is also of interest in the analysis of chemical systems. In particular, a variety of important thermodynamic properties, including the molecular configurational entropy, can be phrased in terms of the multivariate couplings between the degrees of freedom of the system [44]. A variety of different approaches have been pursued to quantify these relationships in computational models. Methods based upon enumerating and characterizing minima in the energy landscape have proved particularly successful in calculating ensemble properties such as free energy and configurational entropy [9]. For larger systems, however, these detailed methods are generally infeasible, and alternative approaches have been pursued which analyze snapshots of the system as generated from simulation methods such as molecular dynamics (MD) or Monte Carlo search [44, 40, 36]. Within these latter phrasings, the parallels between the biological systems described above and these chemical systems become more clear. In both cases, one is presented with a finite set of multivariate samples from which one tries to extract and characterize the relevant multivariate couplings represented within the system.

Information theory as a general framework for quantifying multivariate relationships

A variety of techniques exist for addressing the type of questions posed above. Of particular interest to this thesis are methods within the field of information theory, which provides a general framework for quantifying statistical couplings. As originally formulated by Shannon, information theory draws a parallel between the variance of a random variable — as measured by the information entropy — and the information contained by the variable [74]. The rationale for relating variance to information can be seen in the context of a hypothetical experiment performed to identify the value of a variable. For a variable that is able to adopt any of a number of values (e.g. a protein whose concentration varies over a wide range, or a molecular degree of freedom that permits occupancy of a number of torsional configurations), an experiment to determine the exact value of the variable generates a large amount of information (e.g., the concentration of the protein is exactly 100 nM, or a molecular torsion is restricted to 180°). In contrast, performing the same experiment to determine the value of a variable with limited variance does not provide much new information. By computing and combining these information entropies across multiple dimensions, information theory also provides a general framework for quantifying statistical relationships between variables. Of particular note is the mutual information which represents the loss of information entropy of one variable when the value of an associated variable is known [74, 16].

For biological and chemical systems, information theory is attractive due to its ability to identify any arbitrary statistical dependency between variables, unlike variance-based methods which are limited to linear representations of such dependencies. Additionally, as can be derived from the conservation of information, many information theoretic statistics, including mutual information, are invariant to reversible transformations. Information theory can also handle categorical or continuous data, as well as mixed systems, and naturally extends to relationships between an arbitrary number of variables [16, 57].

Despite these advantages, the application of information theory can be challenging given the relatively large number of sample points needed to generate converged estimates of the statistics, particularly those involving high-dimensional relationships. As a result, most applications of information theory to the type of data generated from biological systems have focused on first- or second-order information theoretic statistics [79, 22, 60]. Even in the context of chemical systems, where sample sizes tend to be dramatically larger, direct application of high-dimensional information theory may still be infeasible due to the exponential growth in sample size requirements as a function of system size [44, 35]. As such, the limited application of high-dimensional information theory in biological and chemical applications appears to be a practical one (i.e., the statistics are poorly converged given the available data sizes), as opposed to a theoretical one.

As mentioned above, a variety of analyses employing information theory in the examination of biological and chemical systems have been performed. For the most part, these applications have focused on the information content of single variables, or the shared information between pairs. For example, pairwise mutual information has been used in the context of gene selection [22], clustering [79], network inference [60], sensitivity analysis [55], identification of residue couplings from multiple sequence alignments [26, 31, 49], and a host of other applications. Furthermore, some higher-dimensional phrasings have been proposed for feature selection [22, 68], chemical library design [48], and the calculation of configurational entropies [44, 35], but for most cases, the small quantities of available data have limited the application of information theory to a its full extent.

In this thesis we present a systematic framework to enable the use of high-dimensional information theoretic problem phrasings, even when a limited number of data samples are available. We accomplish this by developing a principled approximation to high-dimensional information theoretic statistics that are constructed using associated statistics of arbitrarily low dimension. The idea that low-order statistics could be used to represent the multivariate behavior of biological and chemical systems is rooted in the observation that these systems often consist of modest numbers of species interacting with each other directly, resulting in a relatively sparse number of direct high-order relationships. Biological and chemical systems seem to build up complex relationships, not through simultaneous coupling of large sets of variables, but by stringing together small sets of interconnected ones.

Through our approximation framework, we enable a variety of high-dimensional information-theoretic phrasings that can elegantly represent key questions in the analysis of multivariate data. For example, a commonly addressed task in the context of biological data is that of feature selection, in which one aims to identify subsets of variables that maximally explain some output of interest. In early applications, such sets were identified by individually ranking each variable by its relationship with the output [34]. Later work found that sets chosen in such a way tend to include largely redundant information, and that superior feature sets could be identified by simultaneously weighing the "relevance" and "redundancy" of the selected features [22]. In the context of information theory, feature selection can be simply phrased as identifying the subset of species that together have maximal mutual information with the output. This phrasing appropriately weighs the relevance and redundancy of the constituent species against each other in a principled manner. Similar high-dimensional phrasings exist for such tasks as representative subset selection, clustering, experimental design, and network inference. In all of these cases, pairwise phrasings have primarily been pursued, due to the poor convergence of the high-dimensional statistics. In this thesis and in ongoing work, we demonstrate that the general high dimensional phrasings, when addressed through our approximations, show comparable performance to state of the art pairwise methods developed for specific applications, while providing a framework for incorporating increasingly high-order information as data collection methods improve.

The structure of this thesis

In the work presented here, we start, in Chapter 2, by developing and characterizing our approximation framework. The approximation is developed in the context of an expansion of the full information entropy as a function of increasingly high-order terms, enabling direct inspection of the assumptions made when utilizing the approximations. We also demonstrate that the approximation provides a guaranteed upper bound to the full entropy when the lower order terms are known exactly, and that the approximation error decreases monotonically as the approximation order is increased. We then validate and examine the approximation framework in the context of synthetic systems where the exact statistics are known analytically, as well as in application to mRNA expression data extracted from multiple tumor tissues.

In Chapter 3, we extend the information theoretic framework to the analysis of a phospho-proteomic signaling data set. This system represents a common structure of biological data in which the number of signals (68 phospho-peptides, each measured at four separate time points) dramatically exceeds the number of experimental conditions (6 total conditions). Using our framework, we identify a subset of 9 phospho-peptides that are shown to provide significantly improved modeling performance in comparison to other selection methods. We also employ a variety of high-dimensional phrasings to examine the relationships between relevant groups of signals, such as the four time points representing each phospho-peptide. In many cases, the relationships identified by our high-dimensional analyses are consistent with known biology, and with previous analysis in the same data set.

Finally, in Chapter 4, we extend our approximation framework to the calculation of molecular configurational entropies from molecular dynamics simulation data. We compare the performance of our framework against an existing approximation method that represents a similar but distinct expansion and truncation of the full entropy. In the context of simulations of linear alkanes, we observe that while our approximation shows slightly worse agreement with well established methods, it demonstrates considerably faster convergence. As such, we identify sampling regimes in which our approximation provides superior agreement with established methods. We also investigate a series of idealized rotameric systems in which the low-order information terms can be determined exactly. In these systems, we consistently observe low errors with our framework, whereas the comparison method demonstrates erratic behavior. Additionally, we highlight bounding and monotonicity guarantees maintained by our framework that may prove important in future applications.

As discussed above, biological networks and molecular systems share a similar structure that provides both challenges and opportunities for their analysis. For both types of systems, many relevant properties involve the multivariate interaction of large numbers of molecular species (in biological networks) or degrees of freedom (in molecular systems). Extracting these key properties directly from data drawn from the multivariate distributions representing the systems can be unreliable, given the so called "Curse of Dimensionality" which suggests that the number of samples needed to describe multivariate relationships scales exponentially with the size of the system. In potential mitigation of these challenges is the observation that while large multivariate interactions exist, they may often be decomposable into core relationships involving relatively few species. For biological networks, the vast majority of behavior is mediated through successive pairwise interactions (binding, catalysis, etc), due at least in part to the vanishingly small likelihood of simultaneous three-body interactions. In chemical systems, many inter-atomic forces can be well approximated as being pairwise-additive, and these forces tend to drop off rapidly with distance, resulting in a similarly decomposable structure.

In this thesis, we have taken advantage of this structure of biological and chemical systems to enable the application of general information theoretic phrasings, even when direct estimation of the high-order statistics is infeasible due to sample sizes. In so doing, we provide a principled, general framework for approximating high-dimensional statistics across a wide range of sampling regimes. Additionally, this framework carries guaranteed bounding properties, as well as monotonic decrease in approximation error with increasingly level of theory. As such, in addition to providing useful approximations for the type of data that is currently being collected, the framework naturally extends to provide increasing accuracy as data collection and analysis methods improve while maintaining a consistent underlying theory.

Chapter 2

MIST: Maximum information spanning trees for dimension reduction of biological data sets[1]

2.1 Introduction

As the size and dimension of biological data sets have grown, a variety of data-mining and machine-learning techniques has been employed as analytical tools. Among these are techniques aimed at a class of problems generally known as dimension reduction problems [34, 79, 38]. Dimension reduction techniques can improve the interpretability of data, either by representing high-dimensional data in a reduced space for direct inspection, or by highlighting important features of data sets that warrant more detailed investigation. For many biological applications, notably the analysis of high-dimensional signaling data, principal component analysis (PCA) and partial least squares (PLS) decomposition are increasingly popular dimension reduction techniques [38, 46]. Whereas these techniques reduce the number of variables in a system by including only statistically important linear combinations of the full set of variables, the related techniques of representative subset selection (RSS) and feature

[1] This chapter has been previously published as: King BM and Tidor B. *Bioinformatics* 25(9):1165–1172, 2009.

selection (FS) instead aim to identify subsets of variables that are statistically important. These techniques can be used as preprocessing steps prior to application of machine learning methods such as classification [22], and have also been applied in chemical library design [48] and biomarker discovery [54].

While many tools reduce dimensionality to maintain variance (variance-based techniques), recent directions have led to information theoretic phrasings [22, 79]. Compared to variance-based methods, information theory has notable advantages. Information theoretic statistics can capture all relationships among a set of variables, whereas variance-based methods may miss nonlinear relationships. Additionally many information theoretic values are invariant to reversible transformations, limiting the need for such common (and somewhat ad hoc) methods as mean-centering, variance-scaling, and log-transforming. Finally, information theory provides a framework for treating both continuous and categorical data, in contrast to variance-based methods, which are unsuitable for categorical data [57, 16]. This common framework can be especially important when incorporating categorical data, such as the classification of a type of cancer, into the analysis of a continuous data set, such as mRNA expression microarrays.

A variety of dimension reduction problems has already been phrased using high-dimensional information theoretic statistics [48, 79, 69]. Notably, the maximum-dependency criterion (maximizing the MI between the feature set and the output) has been proposed for feature selection [69]. While the high-dimensional phrasing is theoretically more correct, difficulties in estimating high-dimensional statistics with finite sample sizes have resulted in poor performance when compared to techniques using only lower-order statistics [69]. That is, methods that are better in principle perform worse in practice due to their need for larger sample sizes. While some low-order methods have been shown to be related to the high-dimensional phrasing [69], they have generally been developed for a specific application, and their utility in other problems is unclear. To our knowledge, there is no available method for systematically replacing high-order metrics with associated low-order ones. Such a method would enable utilization of the general high-dimensional phrasing but avoid the sampling issues that plague direct applications.

In this chapter we present a general framework for approximating high-dimensional information theoretic statistics using associated statistics of arbitrarily low order. Due to a relationship to the minimum spanning tree over a graph representation of the system, we refer to these approximations as Maximum Information Spanning Trees (MIST). The framework is demonstrated on synthetic data and a series of microarray data sets relevant to cancer classification, and the performance is compared to other approaches.

2.2 Theory

Information theory is a framework for describing relationships of random variables [74]. The two most heavily used concepts from information theory with regard to dimension reduction are the concepts of information entropy and mutual information. The entropy of a random variable, $H(x)$, quantifies the uncertainty or randomness of that variable and is a function of its probability distribution, $p(x)$, also called the Probability Mass Function (PMF)

$$H(x) = -\sum_{i=1}^{b} p(x_i) \log \left[p\left(x_i\right)\right], \tag{2.1}$$

where the summation is over all b bins representing the states of x. To describe the relationship between two random variables x and y, one can consider the conditional entropy of x given that y is known, $H(x|y)$. If x and y are related in some way, knowledge of y may reduce the uncertainty in x, thus reducing the entropy. Conditioning can never increase the entropy of a variable, so $H(x) \geq H(x|y)$. The difference between the entropy and the conditional entropy of a variable is a measure of the amount of information shared between the two variables. This difference is defined as the mutual information (MI), $I(x;y)$, and is symmetric

$$I(x;y) = H(x) - H(x|y) = H(y) - H(y|x) = I(y;x). \tag{2.2}$$

All of these concepts are similarly defined for vectors x and y, where they are functions of the associated higher-order probability distributions [57, 16].

MIST Entropy Approximation Framework

The goal is to find an approximation H_n^k to the joint entropy of n variables using entropies of order no greater than some $k < n$,

$$H_n^k \left(H_1 \ldots H_k\right) \approx H_n \left(x_1 \ldots x_n\right), \tag{2.3}$$

where H_i denotes a true entropy of order i and H_i^j denotes a j^{th}-order approximation to an entropy of order i. To arrive at such an approximation, we begin with an exact expansion of the joint entropy of n variables [16]

$$H_n \left(x_1 \ldots x_n\right) = \sum_{i=1}^{n} H_i \left(x_i | x_1 \ldots x_{i-1}\right). \tag{2.4}$$

Note that Equation 2.4 produces the same LHS information entropy H_n for all permutations of the indices of the x_i and that the RHS is a series of terms of increasingly higher order. We collect the first k terms on the RHS and identify this as the k^{th}-order information entropy of the first k variables, giving

$$H_n \left(x_1 \ldots x_n\right) = H_k \left(x_1 \ldots x_k\right) + \sum_{i=k+1}^{n} H_i \left(x_i | x_1 \ldots x_{i-1}\right). \tag{2.5}$$

We replace each term in the summation by its k^{th}-order approximation. Because conditioning cannot increase the entropy, each approximation term is an upper bound on the term it replaced,

$$H_n \left(x_1 \ldots x_n\right) \leq H_k \left(x_1 \ldots x_k\right) + \sum_{i=k+1}^{n} H_i \left(x_i | x_1 \ldots x_{k-1}\right) = H_n^k. \tag{2.6}$$

All the terms in this sum are k^{th}-order, providing an approximation, H_n^k, which is formally an upper bound. Note that for $k = n$ this expression returns to the exact expansion from Equation 2.4.

Because the indexing of the variables is arbitrary, there are a combinatorial number of approximations consistent with Equation 2.6, all of which are upper bounds to the true joint entropy. There are actually two levels of arbitrary indexing, one being which variables make up the first k and the second being the selection of $k-1$ variables used to bound each term beyond the first on the RHS of Equation 2.6. The best of these approximations is therefore the one that generates the minimum H_n^k, as this will provide the tightest bound consistent with this framework. To complete the approximation, we therefore desire a method for choosing the indexing that produces the best of these bounds.

For low dimensional problems one can enumerate the space of consistent approximations and use the smallest one. To provide a general solution, we first separate out elements that are independent of the indexing. Each conditional entropy term can be divided into an entropy and a MI component, as shown in Equation 2.2.

$$H_n^k = H_k\left(x_1 \ldots x_k\right) + \sum_{i=k+1}^{n} \left[H_1(x_i) - I_k(x_i; x_1 \ldots x_{k-1})\right]. \tag{2.7}$$

Because all individual self entropy terms will ultimately be included in the summation, they are not affected by the indexing, whereas the MI terms do depend on the indexing. For $k = 2$, we arrive at a compact expression of the best second-order approximation within this framework that depends only upon the indexing of the pairwise MI terms,

$$H_n^2 = \sum_{i=1}^{n} H_1(x_i) - \max_{\vec{j}} \sum_{i=2}^{n} I_2(x_i; x_{j_i \in [1,i-1]}). \tag{2.8}$$

The goal is to select the ordering of the indices, i, and the conditioning terms, j, to minimize the expression. The selection of i and j has no effect on the left-hand sum, so it can be ignored during the optimization. We are then left with $n-1$ second-order terms to consider. To phrase the optimization of indices over these terms, consider a graph where the nodes are the variables and the edges are all possible pairwise MI terms. The result is a fully connected graph of n nodes from which we choose $n-1$ edges to maximize the sum of the edge weights. The choice of edges is constrained such that every node must have at least one edge. Because only $n - 1$ edges are chosen, this also constrains the graph to be acyclic.

By negating the edge weights and adding a sufficiently large constant to ensure positivity, the problem is equivalent to the Minimum Spanning Tree (MST) from graph theory. A variety of algorithms has been developed to find the optimal solution, including Prim's algorithm [14], a greedy scheme in which the smallest allowed edge is chosen during each iteration. Using this algorithm, we define a method for efficiently finding the best second-order approximation consistent with Equation 2.8. The computational complexity of Prim's algorithm for a fully connected graph, and thus of our method, is $O(N^2)$. For the higher-order approximations, we apply the greedy algorithm to select the best k^{th}-order approximation consistent with Equation 2.6. Although it is not guaranteed to be optimal, in small test systems where enumeration is possible, the greedy scheme resulted in bounds nearly as tight. Note that the MST phrasing, as used here, is merely an optimization method for finding the best approximation consistent with the mathematical framework, and is not necessarily an inherently meaningful representation.

Bias-Estimation and Propagation

The bias associated with computing the MIST approximation can be estimated by propagating the bias associated with estimating each of the low-order terms. For clarity we focus on the second-order approximation (MIST_2) although the method can be easily extended for arbitrarily high approximation order. The error model we use takes advantage of two properties of entropy estimation: (1) higher entropy variables are more difficult to estimate (have higher errors), and (2) entropy estimates are negatively biased (direct estimates are generally underestimates) [67]. While neither of these properties is guaranteed for any single estimate, they are true on average. We also assume that the estimation errors associated with the first-order entropies are negligible with respect to the errors in the higher-order terms.

We first consider the bias associated with estimating a single second-order entropy. For any pair of variables with fixed self entropies, nonzero MI between them will reduce the joint entropy of the pair. Because higher entropy variables have higher estimation bias, the highest possible bias comes when the variables are independent. By forcibly decoupling any pair of variables (by shuffling their order with respect to each other), we compute an estimate that is greater than or equal to the true bias,

$$\begin{aligned} H(x,y) - \left\langle \overline{H(x,y)} \right\rangle &\leq H_{\text{ind}}(x,y) - \left\langle \overline{H_{\text{ind}}(x,y)} \right\rangle \\ &\lesssim \overline{H(x)} + \overline{H(y)} - \left\langle \overline{H_{\text{ind}}(x,y)} \right\rangle \end{aligned} \tag{2.9}$$

where the angled brackets indicate averages over repeated samples and the overbars indicate entropy estimates. All quantities on the RHS are directly computable, and by repeating the shuffling procedure, the average estimation bias can be estimated or confidence limits can be established quantifying the likelihood of the true estimation error being greater than the computed value.

With a reasonable estimate of the bias associated with computing each second-order entropy, we need to propagate the bias through the MIST approximation. We start by rewriting Equation 2.8 assuming that the indexing i, j has been determined using the MST approach as described above, and by expanding the MI term into the corresponding difference of entropies

$$\begin{aligned} H_n^2 &= \sum_{i=1}^{n} H_1(x_i) - \sum_{i=2}^{n} \left[H_1(x_i) + H_1(x_j) - H_2(x_i, x_j) \right] \\ &= H_1(x_1) - \sum_{i=2}^{n} \left[H_1(x_j) - H_2(x_i, x_j) \right]. \end{aligned} \tag{2.10}$$

Because we assume the bias in estimating first-order entropies to be small with respect to the bias in higher-order terms, the propagated bias in this expression is dominated by the errors in approximating the $n-1$ second-order entropies. Because all of these terms are negatively biased, we expect that overall propagated error to be negatively biased as well; i.e., the computed H_n^2 is expected to be an underestimate of the approximation assuming no estimation errors in the low-order terms. Consequently, by summing the second-order bias approximated by Equation 2.9, we arrive at an expected bias for the full approximation:

$$H_n^2 - \left\langle \overline{H_n^2} \right\rangle \lesssim \sum_{i=2}^{n} \left[\overline{H(x_i)} + \overline{H(x_j)} - \left\langle \overline{H_{\text{ind}}(x_i, x_j)} \right\rangle \right]. \qquad (2.11)$$

As with Equation 2.9, repeated shuffling allows one to estimate the expected bias and to compute confidence limits on the calculation.

2.3 Methods

Direct Entropy Estimation

While the framework developed here is equally applicable to continuous phrasings of information theory, all variables in this work were treated as discrete. For continuous data, variables were discretized into three equiprobable bins unless otherwise stated. Similar results were achieved using different binning protocols and numbers of bins. For discrete data no pre-processing was performed. Entropies of arbitrary order were computed from data by approximating the PMF by the frequencies and using the resulting PMF estimate in Equation 2.1. The MI's were then computed from the estimated entropies according to Equation 2.2.

Bias Estimation

Bias estimates were computed as described in Section 2.2. The bias of all pairs of variables was first estimated using Equation 2.9 by shuffling the ordering of samples for each pair and recomputing the entropy directly. This procedure was repeated until the bias estimate computed from two halves of the shuffling samples agreed within 0.01 nats. The pairs' biases were then used to approximate the bias of each high-order approximation according to Equation 2.11. The terms included in the summation were chosen according to the MIST method prior to any error analysis. Two cases were examined for computing the term in angled brackets. Either the converged mean value was used to compute the expected bias, or 100 samples were drawn and the maximum error from this set was used for each term in the sum, resulting in a $p = 0.01$ confidence limit that the true value of the entropy approximation lies below this max-error value.

Validation Framework

To evaluate the approximation, we developed a framework for generating relational models with analytically determinable entropies from which we could draw sample data. These networks consisted of 5–11 discrete nodes connected by randomly placed unidirectional influence edges. All nodes initially had an unnormalized uniform probability of 1 for each state. If node A influenced node B with weight w, then B was favored to adopt the same state as A by adding w to the unnormalized probability of that state in B. For higher-dimensional influences, the states of all parents where summed and remapped to the support of the child, and the corresponding state in the child was favored by adding the influence weight to that state. Influences including 1–4 parents were included, with 4–19 influences of each order, depending on the number of nodes in the system. Influence weights ranged from 1–10 and all variables had 3 bins. For each system, the joint entropy of all combinations of nodes was computed analytically and 10,000 samples were drawn from each network.

Feature Selection and Classification Error

For the feature selection task, an incremental method was used in which features were added one at a time to the set of already chosen features either at random or in order to maximize the score of the new feature set according to: (1) maximum dependency using direct estimation, (2) maximum dependency using MIST of order two (MIST$_2$), or (3) a second-order approximation proposed elsewhere specifically for feature selection know as minimum-redundancy-maximum-relevance (mRMR) [22]. All feature selection methods were evaluated by training on 75% of the samples and testing on the remaining 25%. This procedure was repeated 200 times and the mean behavior is reported. The data were discretized and the features chosen using only the training data. The frequency of each gene across the 200 trials was also recorded, and the Bonferroni-adjusted p-value for each gene occurring this many times was computed compared to a null model in which features are chosen at random. The subset of features was then used to train support vector machine (SVM) using a linear kernel, linear discriminant analysis (LDA), 3-nearest-neighbor (3NN), or 5-nearest-neighbor (5NN) classifiers [33, and references therein]. Additional SVM kernels (polynomials of order 2 and 3, Gaussian Radial Basis Function, and Multilayer Perceptron) where also examined; while these kernels generally resulted in better fits to the training sets, they performed worse than the linear kernel in cross-validation. To compute the correlation between the metric scores and classification error, 100 subsets each of 1–15 features were chosen at random and the cross-validation classification error was computed. Additionally, the MI of each feature set was computed using all samples according to MIST_2, mRMR, and direct estimation.

Data Sets

Gene expression data sets relating to the classification of four cancer types were used for the feature selection task. Samples from prostate [78], breast [83], leukemia [34], and colon [2] were analzyed. Additional information on the data sets is available in Table A.1.

2.4 Results

2.4.1 Direct Validation

To validate the method, we examined the performance of the MIST approximation in systems with analytically computable entropies. For real-world applications the entropies of the true distribution are estimated from limited data sets, and the corresponding numerical experiments were performed here. To serve this function, we developed a framework to generate networks with a variable number of nodes, interactions, orders of interaction, discrete states, and weights of influence between nodes.

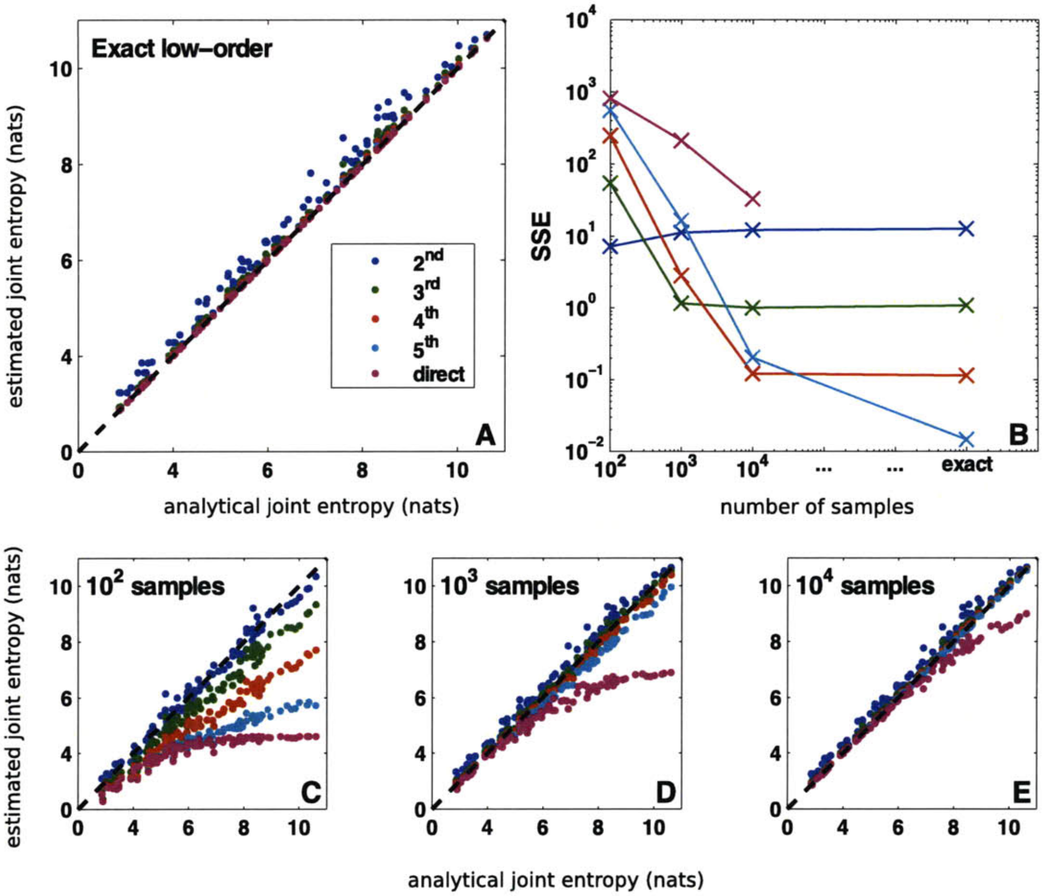


Figure 2-1: **Direct validation of MIST entropy approximation.** To evaluate the MIST framework, we simulated 100 randomly generated networks with analytically computable joint entropies and applied the metrics using a range of sample sizes. When the analytical entropies are known exactly (A), the higher-order approximations performing increasingly well. When the entropies are estimated from a finite sample, however (C–E), the approximations provide the best estimates, with the higher-order approximations performing better as more data become available. This behavior is quantified by computing the sum-of-squared error of each metric as a function of the sampling regime (B). The best approximation to use depends upon the amount of data available, but for all cases examined with finite sample size, the approximations outperform direct estimation and the second-order approximation provides a good estimate.

For each of these networks, all of the joint entropies were analytically determined for comparison to the approximations (see Methods).

Using this framework we randomly generated 100 networks containing between five and eleven variables each with widely varied topologies, and we sampled 10,000 points from the joint distribution. For each network, we then computed the joint entropy of all variables in the network either (1) analytically, (2) directly from the data, (3) using the the second- through fifth-order MIST approximations with analytical low-order entropies up to and including k, or (4) using MIST after estimating the low-order entropies from the sampled data. Additionally, half of the nodes in each network were randomly chosen and the MI between the chosen set and the unchosen set was computed according to all the metrics. The results for entropy and MI approximation are shown in Figures 2-1 and A-1, respectively.

The scatter-plots show the relationship between each of the MIST approximations and the analytical value. As guaranteed by the theory, when the exact low-order entropies are known (panels A), all joint entropy approximations are greater than or equal to the true joint entropy, and the higher-order approximations are increasingly accurate. While there are no guarantees for the behavior of the MI approximation, all approximations tend to underestimate the true MI and the higher-order approximations generally perform better. In some cases the lower-order approximations are able to fully represent the network, resulting in perfect accuracy and in all cases the MIST approximations tend to be fairly accurate.

For biological applications, the exact low-order terms are not available and must instead be estimated from a finite sample of the underlying distribution (panels C–D). Because estimating high-order joint entropies requires larger sample sizes than estimating low-order entropies, the relative performance of the approximations is crucially tied to the number of samples available. In the least sampled case shown here (100 points, panels C), the second-order approximation (MIST$_2$) yielded more accurate results than any of the other methods for computing entropy, while the second- and third-order approximations performed about equally well for MI. As more samples were used to estimate the low-order terms, the higher-order approximations began

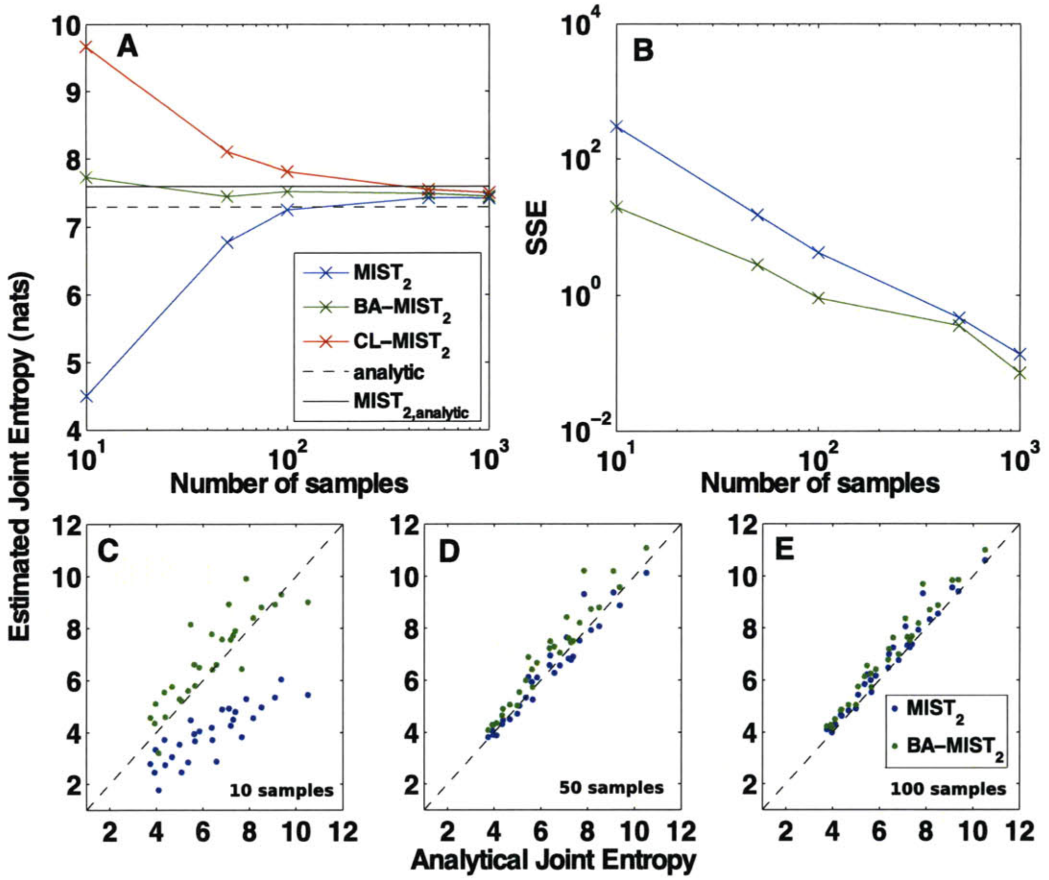


Figure 2-2: **Bias-adjusting for the MIST entropy approximation.** Networks were generated and simulated as in Figure 2-1. The joint entropy of each network was computed by the second-order MIST approximation with (BA-MIST$_2$) or without (MIST$_2$) bias adjusting. (A) The performance of both metrics as well as a $p = 0.01$ confidence limit for MIST (CL-MIST$_2$) approach the analytical MIST$_2$ with increasing samples. (B) The sum-squared-error (SSE) for estimating the analytical MIST$_2$ is shown to decrease as a function of sample size. (C–D) MIST$_2$ and BA-MIST$_2$ were computed using 10, 50, or 100 samples and are plotted against the analytical MIST$_2$.

to outperform the lower-order ones. This trend is quantified in the upper-right plots (B), which show the sum-of-squared error (SSE) for each approximation tested. For all sample-sizes tested here, direct estimation performed the worst, demonstrating the impracticality of estimating high-order information theoretic terms directly. Furthermore as can be seen in panels C–E, the MIST$_2$ approximation is quite accurate for all sample sizes. When more data are available, the higher-order approximations can provide even better accuracy than MIST$_2$, but MIST$_2$ itself appears to be a good metric for all sample sizes tested.

We also examined the behavior of our bias approximation framework in the same systems for MIST_2. For each pair of variables, we computed the converged bias and the maximum observed error over 100 shuffling iterations. For each MIST-approximated joint entropy we propagated both error sets through to determine a bias-adjusted entropy (BA-MIST_2) and $p = 0.01$ confidence limit. We then compared these values to the analytically determined ones in different sampling regimes (Figure 2-2).

In these systems, the bias-adjusted entropy proved to be a significantly better estimator of the MIST approximation than the unadjusted estimator. This result is not necessarily expected, as the bias was computed using the different, but related, system in which all variables were forcibly decoupled. That the bias-adjusted values are not strictly greater than the approximation using analytically determined values is likely a result of the approximations made in the analysis: namely, neglecting the errors in first-order terms and adjusting from a single observed value, rather than a mean from repeated samplings. As expected, the bias decreases as more samples are used, resulting in the bias-adjusted and unadjusted approximations converging for higher sampling regimes. Because the BA-MIST is always greater than MIST without bias-adjusting, and the MIST approximation itself is an upper bound to the true entropy, for higher sampling regimes, bias-adjusting actually results in poorer performance with respect to the analytical answer. While the bias is likely to be small in these cases, this result suggests that while BA-MIST is likely more accurate for low-sampling regimes, when more data is available, MIST without bias-adjusting may have lower error with respect to the true joint entropy.

The confidence limit also shows the expected behavior. While it is not as good an estimator as the bias-adjusted metric, it does provide an upper bound to the approximation computed with analytical entropies within the resolution of the estimation techniques. As such, this metric can provide a guide towards the convergence of the MIST approximation techniques and may lend some insight into the selection of the appropriate order of approximation.

2.4.2 Biological Application

To further characterize the MIST approximation and to evaluate performance in tasks relevant to the interpretation of biological data, we employed MIST in the task of feature selection, which has been previously phrased using information theory [69]. Feature selection is the task of choosing a subset of available features for use in some learning task, such as classification; the information theoretic phrasing seeks the feature subset with maximal MI with the classification. A well studied example is that of selecting a subset of gene expression levels to use when building classifiers to discriminate among cancer types [22, 32, 25].To explore the performance of the MIST approximation in this task, we analyzed four gene expression data sets (which varied both in the number of samples and the number of genes) that had previously been used to classify cancer type in prostate [78], breast [83], leukemia [34], and colon [2].

The rationale behind using MI to choose gene subsets comes from the relationship between MI and classification error [65]. To evaluate the relationship between MIST_2 and the true relationships in these biological data sets, we therefore computed the cross-validated classification error using 100 randomly chosen subsets including 1–15 genes and a range of classifiers. We also computed the MI of the same feature sets with the class variable according to MIST_2 and direct estimation, as well as an existing incremental feature selection metric that has been shown to be an approximation of high-dimensional MI known as minimum-redundancy-maximum-relevance (mRMR) [69]. The Pearson correlation coefficient between the SVM cross-validation classification error and the MI metrics for each set size is shown in Figure 2-3. Results using 3NN, 5NN, or LDA classification error showed similar trends, as did those using the fit error rather than the cross-validation error (data not shown). The SVM classifier was chosen due to its superior performance across the four data sets.

For all four systems, all three metrics have a strong negative correlation coefficient for the feature sets of size one, indicating that high MI corresponds to low classification error, as expected. For larger numbers of features, however, while the MIST_2 approximation maintains reasonable negative correlation for all sizes and data

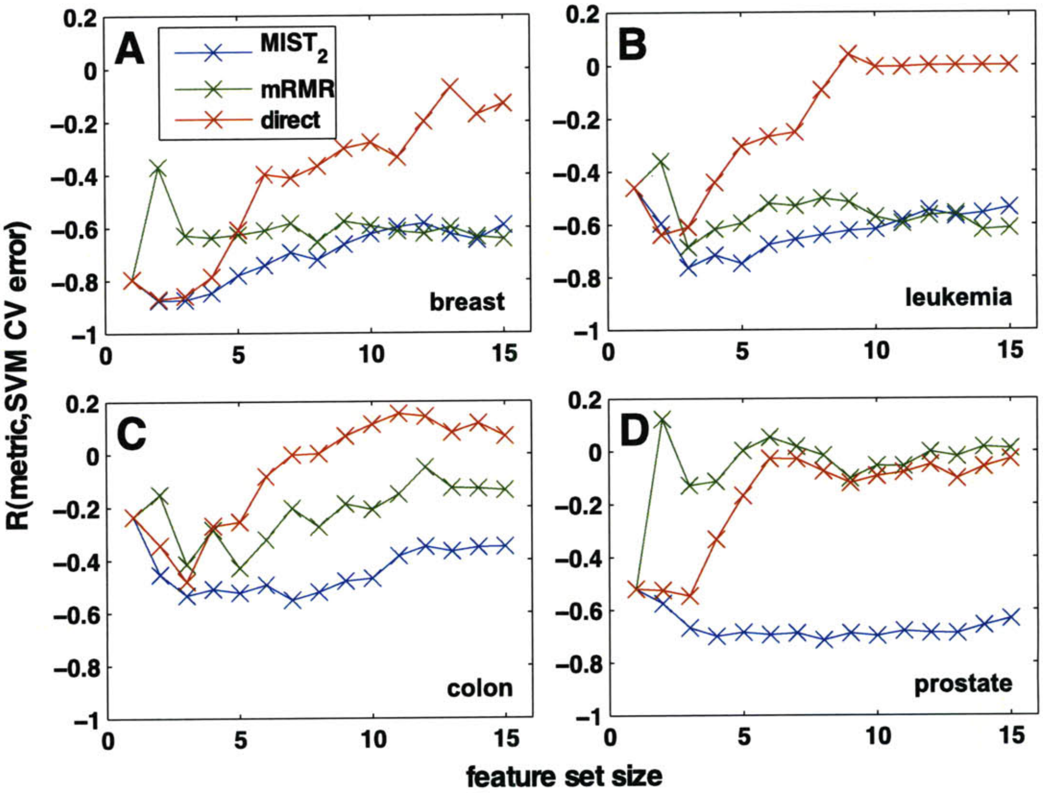


Figure 2-3: **Correlation of MI metrics with classification error.** The classification error of randomly chosen subsets of 1–15 genes was computed through cross-validation with an SVM based classifier. The same sets were then scored by MIST$_2$, MI computed with direct estimation, and mRMR. The Pearson correlation between each metric and the error was computed for gene expression data sets collected in (A) breast, (B) leukemia, (C) colon, and (D) prostate tissue. For all cases, MIST$_2$ shows strong negative correlation with CV error, meaning high MI is associated with low error. While correlated in some cases, both mRMR and direct estimation show poor correlation for some set sizes and data sets

sets, the direct estimation has virtually no correlation with classification error for sets larger than five. For breast (A) and leukemia (B), MIST_2 and mRMR are relatively close though MIST_2 generally exhibits slightly better correlation. For colon (C) and prostate (D), however, MIST_2 exhibits significantly better correlation for larger feature sets. The correlation across sets of different size was also computed and is shown in Figure A-4. While correlation between different sizes is not necessary for standard FS phrasings, the strong negative correlation of MIST_2, even across sets of varied size is further evidence that the approximation reflects the underlying relationships of the system.

In practice, for feature selection the MI metric would be used to select a single subset of features that is expected to have low classification error. In this task, correlation across all sets is not necessary as long as the top ranked set is a good one. To evaluate the utility of MIST in this application, we included it, as well as direct estimation and mRMR, in an incremental feature selection task to choose subsets of genes with which to build a classifier for each of the four tissue types. For each data set, 75% of the samples were used to select the best set of size 1–15 (or 1–10 for direct estimation) according to each metric in an incremental fashion. SVM classifiers were then trained on the same 75% and used to predict the class of the remaining 25% of the samples. This procedure was repeated 200 times to determine the average cross-validation error of the feature selection/classification methods. The performance of randomly chosen feature sets was also computed and in all cases was significantly worse than all tested methods (Figure A-2). Parallel studies were performed using 3NN, 5NN, and LDA classifiers (Figure A-3), as well as ones in which features were preselected using the full data set rather than only 75% (data not shown). Leave one out cross-validation schemes were also examined (data not shown). While the results in all cases showed similar trends, the SVM classifier consistently outperformed the other classifiers and the 75% cross-validation scheme seemed to be the most stringent test. The mean SVM classification errors are shown in Figure 2-4.

For all cases, the MIST_2 feature sets showed lower classification errors relative to direct estimation and mRMR when choosing a small number of features (2–5). This is consistent with the better correlation with the classification error for MIST$_2$ shown in Figure 2-3. For the breast data, this improvement was maintained for feature sets of all sizes. For the other three systems, however, both direct estimation and mRMR generated sets with lower classification errors for sets including more than 5–7 genes. This result is particularly surprising given that this is the regime in which MIST showed improved correlation with classification error relative to the other metrics. Regardless, while MIST appears to select superior subsets of size 2–5, this behavior does not generally appear to extend to large set sizes and deserves further study.

In the above validation scheme, many different feature sets were chosen using different subsets of sample data so as to characterize the expected performance of the metric for predictive tasks. In application however, the features would be selected using all the samples available for training. We therefore incrementally selected the set of 10 most informative genes according to MIST$_2$ for each of the data sets. An ordered list of these genes along with references demonstrating the relevance to cancer biology or cancer diagnosis for a subset of the genes can be found in Table A.2. All of the selected feature sets contained genes that have been either statistically or functionally related to cancer. Many of the genes have also been identified in other computational studies. The most informative gene for all four datasets had previously been identified in multiple studies. For the highly studied leukemia and colon datasets, nearly all of the genes have been identified in some study, though not always in the top 10 ranked genes. Notably, three of the genes identified in the breast dataset (NM_003981, AI918032, and AF055033) consistently appeared in the globally optimal feature sets of size 2–7 in [11].

We also evaluated the robustness of the chosen genes by observing how often they were chosen in the 200 CV trials. The p-value for having at least this frequency for each of the chosen genes is shown in Table A.2. While some of the globally chosen genes are not robustly re-selected, the majority (32/40) of the genes appear in the 200 trials more often than expected at random (Bonferroni-corrected p-value ≤ 0.01), particularly for the breast (8/10) and prostate (10/10) datasets which have larger sample sizes.

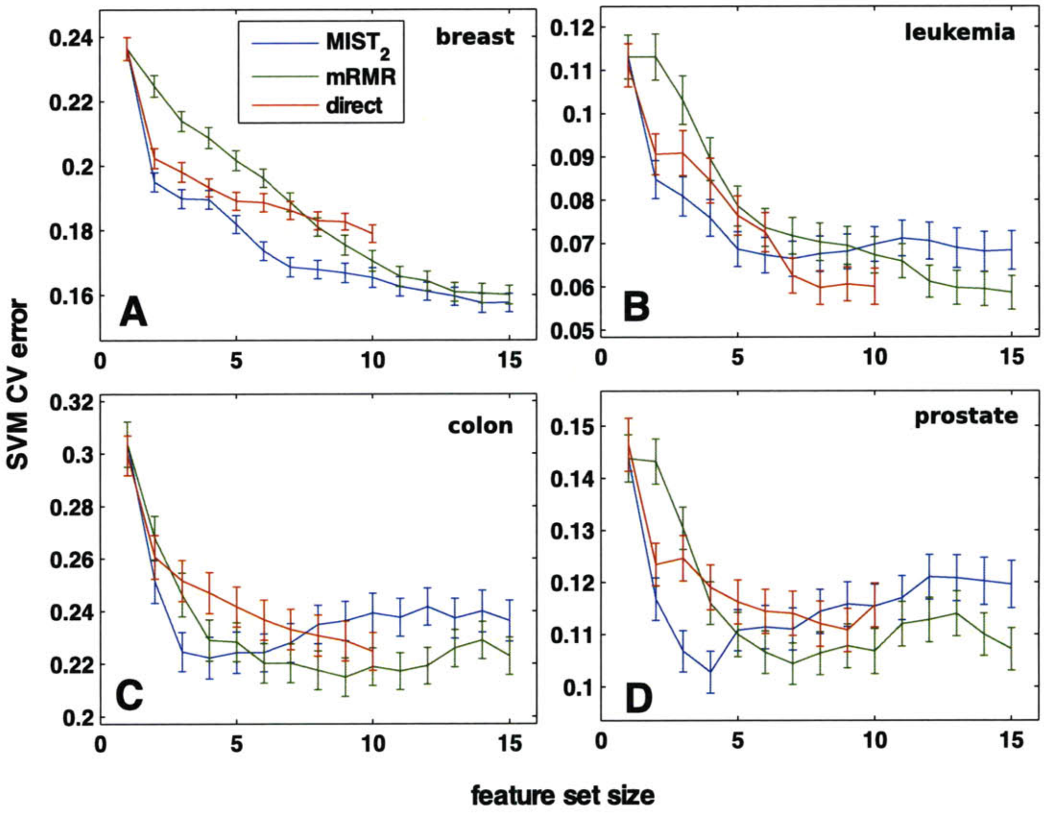


Figure 2-4: **Gene subset selection for cancer classification.** Subsets of gene expression levels were chosen incrementally to maximize the information with the cancer class according to MIST$_2$, direct estimation of MI, or mRMR and scored by the cross-validation error of an SVM classifier. For all data sets, 75% of the data was separated and used to select features and train the model; the classifier was then used to classify the remaining 25%. The mean classification error and standard error of the mean for 200 training/testing partitionings are reported. Genes were selected for data sets relating to (A) breast, (B) leukemia, (C) colon, and (D) prostate cancer. The performance of randomly chosen feature sets along side these methods can be seen in Figure 2-4 .

2.5 Discussion

Here, a novel framework for approximating high-order information theoretic statistics with associated statistics of arbitrarily low order has been developed and validated. Due to the generality of information theory, the MIST approximation should allow the use of high-dimensional information theoretic phrasings for a variety of problems, even in cases when data quantities are limited. Information theoretic phrasings exist for such tasks as feature selection (shown here), representative subset selection [48], clustering [79], network inference [52], and other applications where relationships of multiple variables are important. Though high-dimensional phrasings are theoretically correct, difficulties in estimating these terms has led to low-order approximations having better performance. While these approximations have been applied to many problems, task-specific metrics were usually developed that are not generally usable across multiple applications. By instead developing a principled approximation to joint entropy and MI, we propose a general method for application to many problems.

In regards to the feature selection task shown here, while MIST$_2$ correlates well with the classification error and generates low-error sets when picking a small number of genes, the overall behavior for choosing larger sets could still likely be improved. For incremental feature selection, MIST and mRMR are similar with the primary difference being that MIST selects a subset of MI terms to consider, whereas mRMR averages all gene-gene terms to compute the redundancy. While both have been shown to relate to the maximum dependency criterion, MIST represents a more general framework for extension to different problems phrasings. In contrast, mRMR has been well calibrated for feature selection, and some features of mRMR may be useful in improving the performance of MIST in feature selection. In particular, preliminary work on incorporating weighting factors to influence the relative importance of the relevance and redundancy suggests that such a scheme may result in a better feature selection method. Additionally, while the current work has focussed on incremental feature selection, the generality of MIST and the good correlation with classification error suggest that global search methods using MIST could be feasible. In it's current form, MIST provides a well principled framework without any ad hoc parameterization that performs comparably to current feature selection methods. Furthermore, MIST can be generalized and ported to other problem phrasings and take advantage of larger data-quantities when they become available.

One natural extension of the MIST approximation is feature selection with multiple outputs. Typical FS phrasings focus on a single output variable, resulting in most FS methods not being directly applicable to multiple-output scenarios. Instead, separate subsets may be chosen for each output and combined subsequently, or multiple outputs can be combined into a single variable. With high-dimensional statistics, rephrasing the maximum dependency criterion for multiple outputs is trivial, by replacing the single output variable with the set of all outputs of interest (i.e. find the set that maximizes MI between the gene set and the output set). In cases where different feature sets can be used for each output, such as preprocessing before machine learning, multiple output feature selection may not be appropriate as a single consensus set will not represent each output as well as the individually chosen sets. In other cases, however, a fixed number of features may be needed to describe multiple outputs and a single optimization for this task could be valuable. Considering the relationships between multiple outputs could be particularly important if the outputs are closely related. For example, in the case of FS for cancer classification, one might consider tumor progression measurements at multiple time points. Alternatively, defining a compact set of features that can classify multiple disease states could be valuable in more efficient diagnostic tools. Designing experiments that are richly informative of a particular set of output variables might also benefit from such methods. In general, having metrics that support multiple outputs allows phrasing FS problems that better reflect questions of interest.

The ability to maintain the general information theoretic phrasing also allows the results between different tasks and experiments to be compared. Information theory is able to treat data from different experimental modalities within the same framework, enabling one to quantitatively compare the information content of different data types without significant preprocessing. Information theory also allows the treatment of categorical and continuous data, and can consider nonlinear relationships, unlike variance-based techniques. While these benefits of information theory have long been understood, the inability to estimate information theoretic terms has often precluded their use in biological systems. By reducing the data requirements for computing high-order entropies, MIST enables the use of information theoretic statistics even when few samples are available, as is often true in biological systems.

Although we have used only the second-order MIST approximation here, the framework provides a range of approximations of higher order, allowing increased accuracy when sufficient quantities of data are available. As high-throughput data collection continues to improve, the framework extends to incorporate third- and fourth-order relationships. Even as larger quantities of data become available, MIST is likely to be useful, as in our synthetic system, even with 10^4 samples, all orders of approximation tested outperformed direct estimation. In Figure 2-1 we have shown how one might select an approximation order based on the sample size. For applications where the analytical solutions are unknown, however, it is unclear how to choose the best approximation order. Additional work is required to fully enable such a method. Despite this, it is encouraging that the second-order approximation performs well both on synthetic and microarray data, even though high-order relationships are known to exist.

While the MIST framework arises from a mathematical approximation, it can alternatively be thought of as a method to infer a relational model of low-order interactions. This model is then used to estimate the high-order statistics of interest. Currently this model is used only for the approximation, however, the good agreement between the approximation and the analytical entropies suggests that the inferred model captures many of the relevant relationships. The generation of relational models for biomarker discover has been previously proposed [84], and network inference tools have been proposed that use pairwise MI as the primary metric [52, 63]. There is reason to believe, therefore, that the relational models inferred may be meaningful, as they reasonably represent the system's statistical relationships.

2.6 Conclusion

Here we have presented a novel method for approximating high-dimensional information theoretic statistics with significantly improved performance when data quantities are limited, as is often true when dealing with biological data. While we have demonstrated the utility of this approximation in feature selection, the generality of information theory should enable application in a number of different learning tasks, including representative subset selection, clustering, and network inference. While previous low-dimensional information theoretic phrasings exist for these problems, they have generally been developed on a problem-by-problem basis, and are thus not directly portable between tasks. By instead focusing on ways to approximate the information theoretic statistics directly, we can take advantage of general information theoretic phrasings in a variety of problems. In addition, our MIST approximation naturally allows for incorporating arbitrarily high-order information as sample sizes increase, providing a consistent framework as the collection of biological data continues to increase in scale.

Chapter 3

Analysis of a high-dimensional phospho-proteomic data set using information theory

3.1 Introduction

A fundamental goal of systems biology is to understand and quantify the multivariate relationships between various molecular species in the cell. Towards this end, increasingly high-throughput experimental techniques have enabled the tracking of concentrations of mRNA, proteins, protein modification states, and other molecular species on a near-global scale [34, 85]. Furthermore, the identification of multivariate relationships requires multiple samples across varied conditions to highlight the correlated changes in these species. Despite increasing work towards performing measurements across multiple samples, these data sets tend to have significantly more species than samples.

As data collection methods have continued to improve, analytical methods have also been developed to deal with the challenges presented by such data. In particular, the large number of species relative to the number of samples generally leads to vastly under-determined systems for many traditional methods. Various dimension-reduction methods have been used to address this issue by reducing the effective number of species under consideration [34, 38, 79]. Principal Component Analysis (PCA) and Partial Least Squares Regression (PLSR) methods have been particularly successful at finding small sets of basis vectors that can explain the variance in the data set itself, or covariance between the data and output variable of interest, respectively [38, 46, 43].

In addition to the variance-based methods such as PCA and PLSR, information theoretic methods have been increasingly used to analyze multivariate biological data [15, 45, 22, 79]. While these methods generally require larger sample sizes than the corresponding variance-based methods, information-based metrics have attractive properties including the handling of both categorical and continuous data, invariance to reversible transformations (such as variance scaling and log-transforming), and the capturing of any statistical dependency, not just the linear relationships captured by variance-based methods [16]. While information theory can in principle be used to directly query high-dimensional relationships, most applications have focused on pairwise relationships between variables due to the limited number of samples available. As discussed in Chapter 2, we have recently pursed directions focused on enabling more general phrasings of information theory by approximating the high-dimensional statistics using only low-order information [45]. These methods allow one to directly phrase high-dimensional questions even when data sizes are limited.

One area where high-dimensional information theoretic phrasings have been applied to the analysis of biological data relates to the machine learning task of feature selection [15, 45, 22]. The goal of feature selection is to identify a compact set of variables that represents the information contained in the full data set, often as a filtering step prior to building a model. Such sets are useful both to improve the interpretability of these models, and to avoid problems related to high dimensionality and overfitting. Typically, feature selection problems are phrased as supervised learning tasks, where features are chosen to be maximally informative of an output variable or response. The maximum depenceny criterion (maximizing the mutual information between the feature set and the outputs), is a common information theoretic phrasing for finding such a set, and has been shown to provide good feature sets both using our approximation framework [45], and using other methods [22]. In this chapter, we again use the maximum dependency phrasing to identify informative subsets. We have also, in previous work, explored unsupervised feature selection phrasings (also known as representative subset selection), in which subsets that maximally represent the full data set are chosen by maximizing the joint entropy of the chosen set [15].

Among the data that have been successfully modeled using PLSR are quantitative mass spectrometry studies monitoring the phosphorylation state of tyrosines in response to various stimulating conditions [85, 46]. Due to the many tyrosines involved in signaling downstream of the epidermal growth factor receptor (EGFR) family of receptor tyrosine kinases, these data have been largely collected in the presence of activating ligands such as epidermal growth factor (EGF) and heregulin (HRG). One such study investigated the tyrosine-phosphorylation pattern at multiple time points in response to stimulation with EGF or HRG, in the background of human mammary epithelial cells with typical (~20,000 copies per cell) or elevated (~600,000 copies per cell) levels of human epidermal growth factor receptor 2 (HER2) [46, 85]. The levels of proliferation and migration were separately examined under the same conditions as well as in unstimulated cells, enabling the mapping of relationships between phospho-tyrosine mediated signaling and the cellular responses, under the various stimulation and HER2 overexpression conditions. A PLSR model built using data from the 62 phospho-peptides in the data set was shown to exhibit excellent fit and cross-validation in describing the proliferation and migration response data across the 6 conditions (no stimulation, EGF, HRG; with typical or elevated HER2). These PLSR models have been analyzed in detail to identify the relevant signals and linear combinations of signals that mediate the responses. Furthermore, a "network gauge" including 9 of the original 62 phospho-sites, across 6 different proteins, was identified by examining weights in the PLSR models, and was shown to exhibit similar fit and cross-validation properties to the models built using the full data set, the suggestion being that these 9 signals alone were predictive of cellular behavior [46].

The data described above represents a common scenario in systems biology in which the number of signals (62 sites at 4 time-points each) dramatically exceeds the number of experimental conditions (6 total conditions). While it is clear that many if not all of the measured signals are involved in mediating the migratory and proliferatory response to the ligands, it is a challenge to understand these potentially complex multivariate relationships given the relatively small number of conditions. PLSR addresses this challenge by finding a compact set of basis vectors that can describe the relationships between the large number of signals and the outputs. The weights of the various species in the PLSR model can then be examined to gain insight about the multivariate relationships present in the data [38, 46].

In this chapter, we evaluate and employ an alternative method of investigating the multivariate relationships present in such data centered around information theory and feature selection. The study relies on our previously established method for approximating high-dimensional information theoretic statistics by combining univariate and pairwise information terms, as described in Chapter 2 and [45]. In contrast to principal component based methods, this analysis enables one to directly query the information content of arbitrary sets of species, as well as the information shared between species and outputs. We therefore believe that information theory, in conjunction with existing methods such as PLSR, can serve as a valuable tool in interpreting a variety of systems biology data sets.

Because information theory does not provide an integrated predictive modeling framework, as PLSR does, we first validate the information theoretic approximations in the context of PLSR models. In particular, we show that subsets of the signals in the data set that are chosen to have maximal information about the migration and proliferation responses generate PLSR models with improved fit and cross-validation performance with respect to other sets of the same size, including the previously determined network gauge. We also show how one can group signals (e.g., multiple time points of the same species) to generate a more intuitive set of signal-signal or signal-response relationships. While we only show a few examples applications here, the generality of information theory allows a variety of simple phrasings to query multivariate data sets. In combination with the MIST approximation framework, such analysis is possible even when a relatively small number of experimental conditions are available.

3.2 Methods

Data preparation

The phospho MS data was preprocessed as described previously, with the exception of using all 68 phospho-peptides captured in the data, as opposed to the subset of 62 used for the original work [46, 85]. All species trajectories were normalized to the 5 minute time point of the parental cell line in response to stimulation with 100 ng/ml EGF. The zero minute time point (pre-stimulation) was used as the constant value for all time points in the serum-free conditions, and the integral over the remaining three time-points was appended to all signals to serve as a metric of total activation. As a result, each of the 68 phospho-peptides was represented as a series of 4 variables (5 min, 15 min, 30 min, and integrated) across 6 different conditions (serum free, +EGF, +HRG, in parental or 24H cells). Each output (migration or proliferation) was represented as a single vector of length 6, corresponding to the conditions.

Calculation of pairwise mutual information

The mutual information between all pairs of signals was first computed according to

$$I(x;y) = H(x) + H(y) - H(x,y) \tag{3.1}$$

where H is defined as

$$H(\boldsymbol{x}) = -\int \rho(\boldsymbol{x}) \log \rho(\boldsymbol{x}) d\boldsymbol{x} \tag{3.2}$$

and ρ is the estimated probability density over all dimensions of $\boldsymbol{x}$. Probability densities were estimated using Parzen windowing [47] with a Gaussian kernel with the covariance matrix set to be equal to the sample covariance matrix scaled by $1/\log(N)$, and truncated at 2 standard deviations in each dimension. For each MI calculation, a fixed window size was used for the one- and two-dimensional entropies. The integrals in the entropy were computed using the QUAD command in MATLAB release 2008b (The Mathworks Inc., Natick, MA) with a tolerance of 10^{-6}.

Approximation of high-order terms using MIST

All calculations of information terms containing more than two variables were computed using the Maximum Information Spanning Trees (MIST) method with an approximation order of two, as described in Chapter 2 and [45]. Briefly, to approximate the joint entropy of N variables, the entropy of all variables and the MI between all $\binom{N}{2}$ pairs of variables is first computed as described above. The 2nd-order approximation to the N^{th}-order entropy is then computed as:

$$H_N^2 = \sum_{i=1}^{N} H_1(x_i) - \max_{j} \sum_{i=2}^{N} I_2(x_i; x_{j_i \in [1,i-1]}). \tag{3.3}$$

The maximization was performed using Prim's algorithm [14] to generate the minimum spanning tree over the fully connected graph represented by the negated MI matrix. For the 2nd-order approximation, this algorithm guarantees the optimal solution compatible with the MIST framework.

Complex high-dimensional statistics

In addition to the high-dimensional entropy terms described above, a variety of composite high-dimensional statistics was used for much of the analysis. In all cases, these statistics were first converted into forms containing only joint entropy terms, and then computed using the joint entropies approximated with MIST as described above. The decompositions used are shown in Table 3.1.

Table 3.1: Decomposition of complex terms into joint entropy formulations

Name	Symbol	Decomposition
Conditional entropy	$H(\boldsymbol{x}\|\boldsymbol{y})$	$H(\boldsymbol{x},\boldsymbol{y}) - H(\boldsymbol{y})$
Mutual information	$I(\boldsymbol{x};\boldsymbol{y})$	$H(\boldsymbol{x}) + H(\boldsymbol{y}) - H(\boldsymbol{x},\boldsymbol{y})$
Conditional MI	$I(\boldsymbol{x};\boldsymbol{y}\|\boldsymbol{z})$	$H(\boldsymbol{x},\boldsymbol{z}) + H(\boldsymbol{y},\boldsymbol{z}) - H(\boldsymbol{x},\boldsymbol{y},\boldsymbol{z}) - H(\boldsymbol{z})$

Partial Least Squares Regression (PLSR) modeling

All PLSR models were built using the PLSREGRESS function in MATLAB release 2008b (The Mathworks Inc., Natick, MA), which implements the SIMPLS algorithm. For each model mapping N of the signals onto a single output (migration or proliferation), the signals were represented as a $6 \times N$ matrix indicating the value of the signal at each of the conditions, and the output was represented as a 6×1 vector indicating the value of the output for each condition. Up to two principal components were included in the model. For cross-validation statistics, each of the 6 conditions was omitted and a new model was regenerated using the 5 remaining conditions. The trained model was then used to assign a prediction of the output for the omitted condition. All variables were variance-scaled with respect to the training set prior to learning the model. Models were evaluated according to two metrics: (1) the fit of the model, i.e., the Pearson correlation between the model output and the true output (R^2); and (2) the predictive power of the cross-validation models, i.e. the Pearson correlation between the predictions of output in the omitted conditions and the true output (Q^2).

Feature selection methods

We examined a variety of schemes for choosing subsets of signals to be maximally informative of the output. In all cases, selecting a signal meant including all 4 measurements associated with the phospho-peptide (3 time points, and the integrated signal). In addition to the previously determined network gauge [46], we examined two classes of selection schemes: ranking and incremental. Ranking schemes involved sorting the 68 signals according to some metric and taking the top members of the list. Incremental methods were able to explicitly consider signals that had already been selected, enabling them to provide complementary signals, as opposed to just individually informative signals. The ranking metrics that we evaluated were:

- **Rank 1 – MI of time-point set with both outputs**: The MI of each of the 68 signals with both outputs together was computed ($I(\boldsymbol{s}; \{m, p\})$, where $\boldsymbol{s}$ is a vector representing the 4 measurements associated with the signal, m is the migration response, and p is the proliferation response). This MI was assigned as the score for each signal.

- **Rank 2 – Max MI of individual with both outputs**: The MI of each of the 4 measurements for each of the 68 signals with both outputs together was computed and the maximum over the 4 measurements was assigned as the score of each signal ($\max_i I(s_i; \{m, p\})$, where s_i is one of the four measurements associated with a signal).

- **Rank 3 – Max MI of individual with either output**: The MI of each of the 4 measurements for each of the 68 signals with each output individually was computed and the maximum over the 8 values was assigned as the score for each signal ($\max_{i,j} I(s_i; o_j)$, where s_i is one of the four measurements associated with a signal, and o_j is one of the two outputs).

- **Rank 4 – Max R^2 of individual with either output**: The Pearson correlation between each of the 4 measurements for each of the 68 signals with each output individually was computed and the maximum over the 8 values was assigned as the score for each signal ($\max_{i,j} R^2(s_i; o_j)$).

Incremental schemes involved adding one of the 68 signals at a time to maximize a scoring function. The incremental schemes are thus able to consider relationships between the already chosen members in order to choose a more effective set as opposed to assigning a single score to each signal. We evaluated two incremental schemes:

- **MIST opt – Time-point sets**: For each step, each candidate signal was evaluated for inclusion by scoring the MI of the new full set (including all time points of the candidate signal) with both outputs. The signal maximizing this value at each step was chosen.

- **FSi ind – Individual measurements**: Each of the $68 \cdot 4 = 272$ individual measurements were considered in each selection step, generating an incremental selection of all measurements. Signals were assigned a score corresponding to the best rank for any of the 4 associated measurements, and the top ranking signals were chosen.

In addition to the schemes shown above, we also estimated the background performance of feature sets of nine signals by randomly choosing 1000 sets of nine signals. All subsequent performance metrics were computed for these 1000 sets in addition to the 7 rationally chosen sets and the network gauge.

3.3 Results

3.3.1 Quantifying relationships among phospho-peptides and with responses

Although the focus of the analysis in this chapter is on high-dimensional information relationships (e.g., the total mutual information between a set of signals and all outputs), the method that we use only requires direct calculation of first- and second-order terms (i.e., the entropy of each measurement and the mutual information between each pair of measurements). We then employ the MIST framework to approximate all higher-dimensional terms of interest. As such, all of the analysis relies upon the pairwise mutual information (MI) matrix shown in Figure 3-1. The MI matrix shows the relationships between all pairs of measurements, where high MI values indicate a strong statistical dependency between the pair. For example, each of the four variables associated with the 3 phospho-peptides in SHC have high information with each other, as one might expect given that these sites are known to be activated in concert downstream of EGFR [72].

While the MI matrix can be informative on its own, it can also be difficult to interpret for the data set of focus because each phospho-peptide is represented by four separate variables (the 3 time points and the integrated signal). To quantify the statistical dependencies between phospho-peptides across all four variables, we can use a high-order information term. In this case, we compute the 8-dimensional mutual information between the four variables corresponding to signal s_i and the four

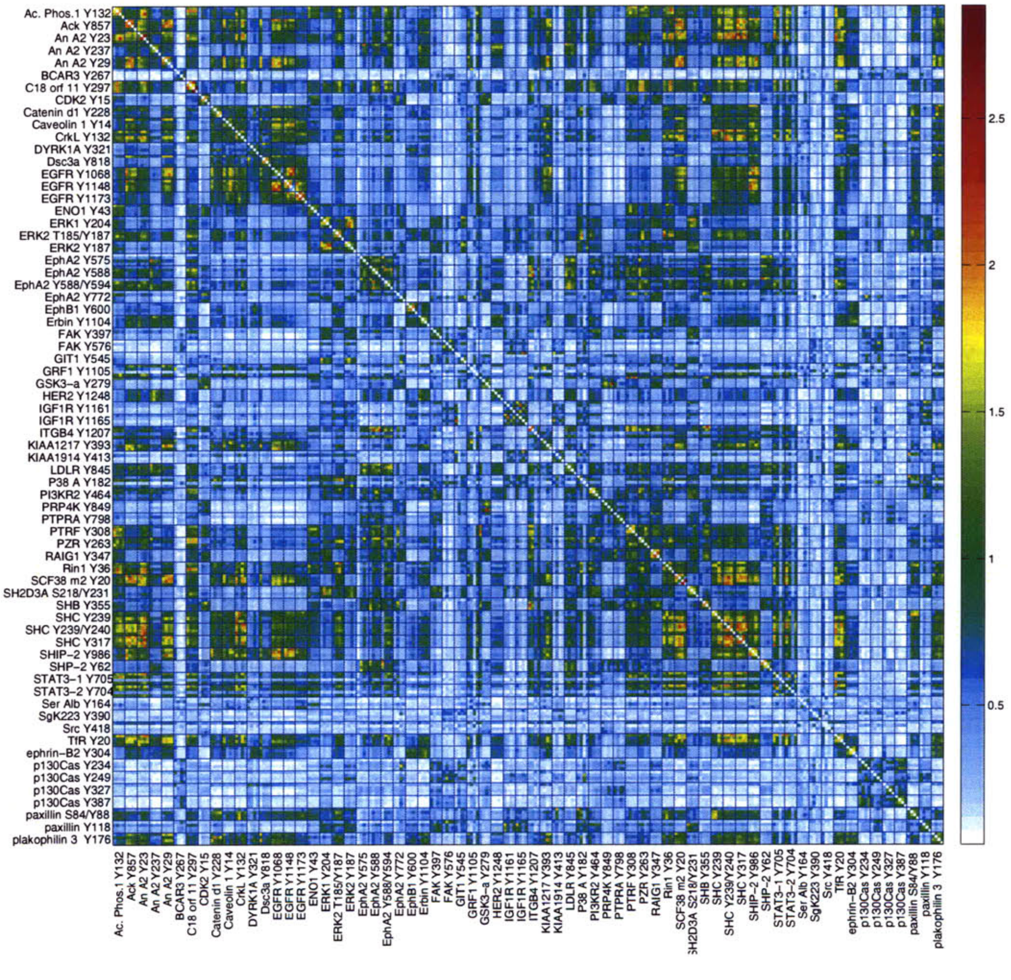


Figure 3-1: **Full MI matrix of signals**: The mutual information between each pair of variables in the data set was computed using Parzen windowing. High mutual information indicates strong statistical dependency across the six experimental conditions. Each phospho-peptide is represented by four variables, ordered from left to right or top to bottom as: 5-, 10-, 30-minute time point, integrated signal.

variables corresponding to signal s_j using the 2nd-order MIST approximation to the corresponding 4th- and 8th-order entropy terms as

$$I_8(\boldsymbol{s_i};\boldsymbol{s_j}) = H_4(\boldsymbol{s_i}) + H_4(\boldsymbol{s_j}) - H_8(\boldsymbol{s_i},\boldsymbol{s_j}) \approx H_4^2(\boldsymbol{s_i}) + H_4^2(\boldsymbol{s_j}) - H_8^2(\boldsymbol{s_i},\boldsymbol{s_j}) \quad (3.4)$$

where the subscripts indicate the dimensionality of the term and the superscripts, where present, represent the order of approximation. In this way, we now have a single value that represents the relationships between all of the measurements associated with one signal and all of the measurements associated with a second signal. The grouped MI matrix for all phospho-peptides is shown in Figure 3-2. Notably, generating this same matrix directly, without the MIST approximation framework, is difficult due to convergence issues as well as computational limitations. For the data shown here, in which only 6 samples are available, traditional histogramming methods are unsuitable for estimating an 8-dimensional entropy, as the vast majority of bins contain no samples, and no single bin contains more than one sample. We employed Parzen Windowing in order to combat these issues for the low-order entropy calculations, but such methods proved computationally intractable for the 8-dimensional terms.

In contrast to the full MI matrix in Figure 3-1, the grouped MI matrix enables easy inspection of the statistical relationships between pairs of phospho-peptides in the data set. A list of the strongest MIs is shown in Table 3.2. Two of the top five interactions, including the strongest overall, are between different phosphorylation sites on the same protein (EGFR Y1148–Y1068 and IGF1R Y1165–Y1161), and a third is shared between sites on isoforms 1 and 2 of STAT3. SHC and CrkL are known to transiently interact [11] and the Ack–EGFR relationship is consistent with the identification of Ack as an early transducer of EGF stimulation [30]. As such, all of the top interactions from the grouped MI matrix are consistent with known biology.

The results of performing the same analysis to identify the strongest relationships between individual peptide-time points using the full MI matrix from Figure 3-1 can

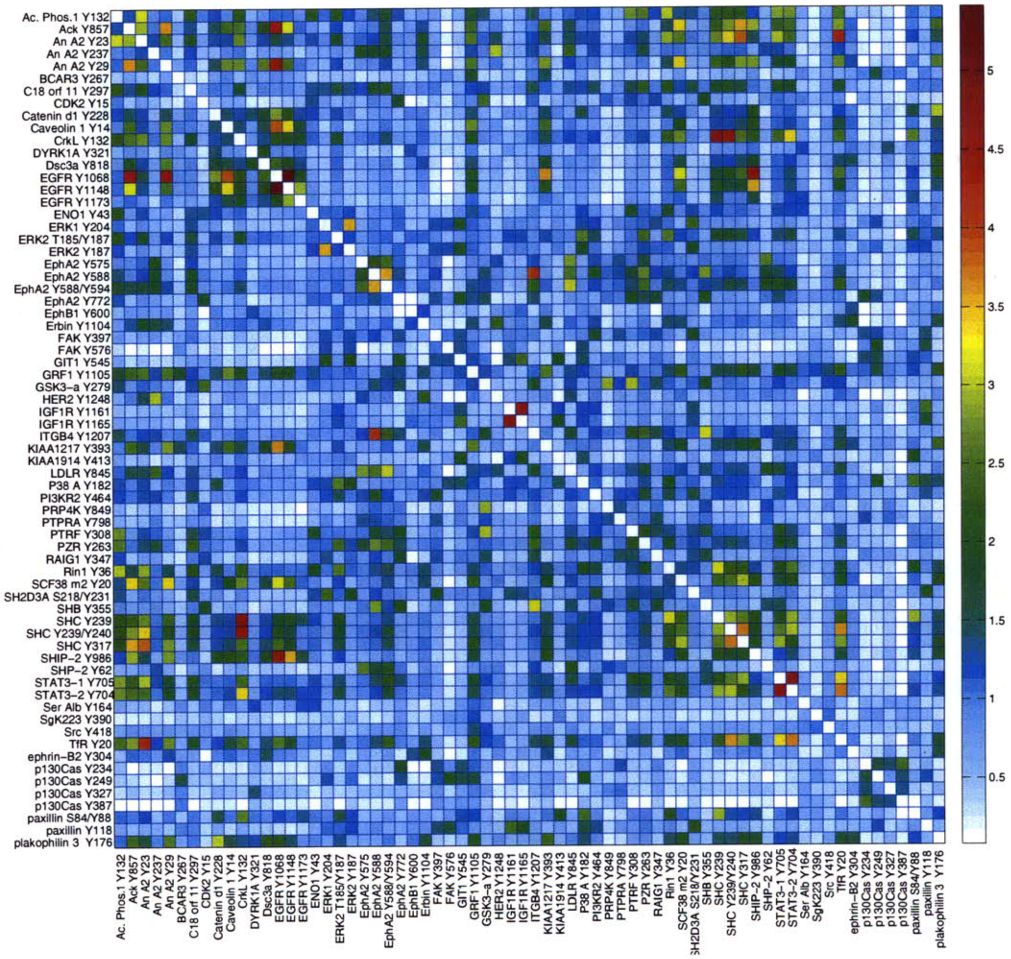


Figure 3-2: **Grouped MI matrix of signals**: For each pair of phospho-peptides, we computed the mutual information between the four measures associated with each signal, as approximated by MIST. This 8-dimensional term represents the statistical dependency between all variables used to represent each pair, enabling comparison of relationships between phospho-peptides as opposed to between individual time points.

Table 3.2: Top scoring pairs of phospho-peptides by grouped MI

MI	phospho-peptide 1	phospho-peptide 2
5.42	EGFR Y1148	EGFR Y1068
4.61	SHC Y239	CrkL Y132
4.58	STAT3-2 Y704	STAT3-1 Y705
4.46	IGF1R Y1165	IGF1R Y1161
4.43	EGFR Y1068	Ack Y857

be seen in Table 3.3. While inspection of these relationships is somewhat more complicated due to the time component, time points from many of the same relationships seen in the grouped MI matrix rank highly in the full matrix. Four of the five strongest pairs, including the three strongest overall, are between different time points of the same phospho-peptide. As with the grouped MI matrix, two sites on EGFR (Y1148 and Y1068) show high information with each other at the 30 minute time point. The five minute timepoint of CrkL Y132 showed high information with the integral of the SHC Y239/Y240 signal, similar to the relationship shown in the grouped MI matrix, though with doubly phosphorylated SHC rather than singly phoshphorylated on Y239. Two of the remaining top relationships in the full MI matrix are between SHC Y317 and Annexin A2 Y23. Annexin A2 has been identified as a mediator of migration [76], and although its association with SHC is not well established, SHC is also known to be associated with migration through EGFR signaling [74]. The fact that the SHC–Annexin A2 relationship is not among the strongest relationships in the grouped MI matrix, despite showing two individual time point pairs with high MI, highlights the difficulties of interpretting the full MI matrix directly without a rigorous framework for appropriately weighting all 16 pairwise relationships between the timepoints associated with each phospho-peptide.

In addition to quantifying relationships between pairs of phospho-peptides, we computed the mutual information between each phospho-peptide profile and the two measured outputs of the system. For each phospho-peptide, we computed the MI between its four variables and: (a) the migratory response, (b), the proliferatory response, or (c) both responses together. As before, because these terms require 5th- and 6th-order terms, we employ our MIST approximation framework. The result is a

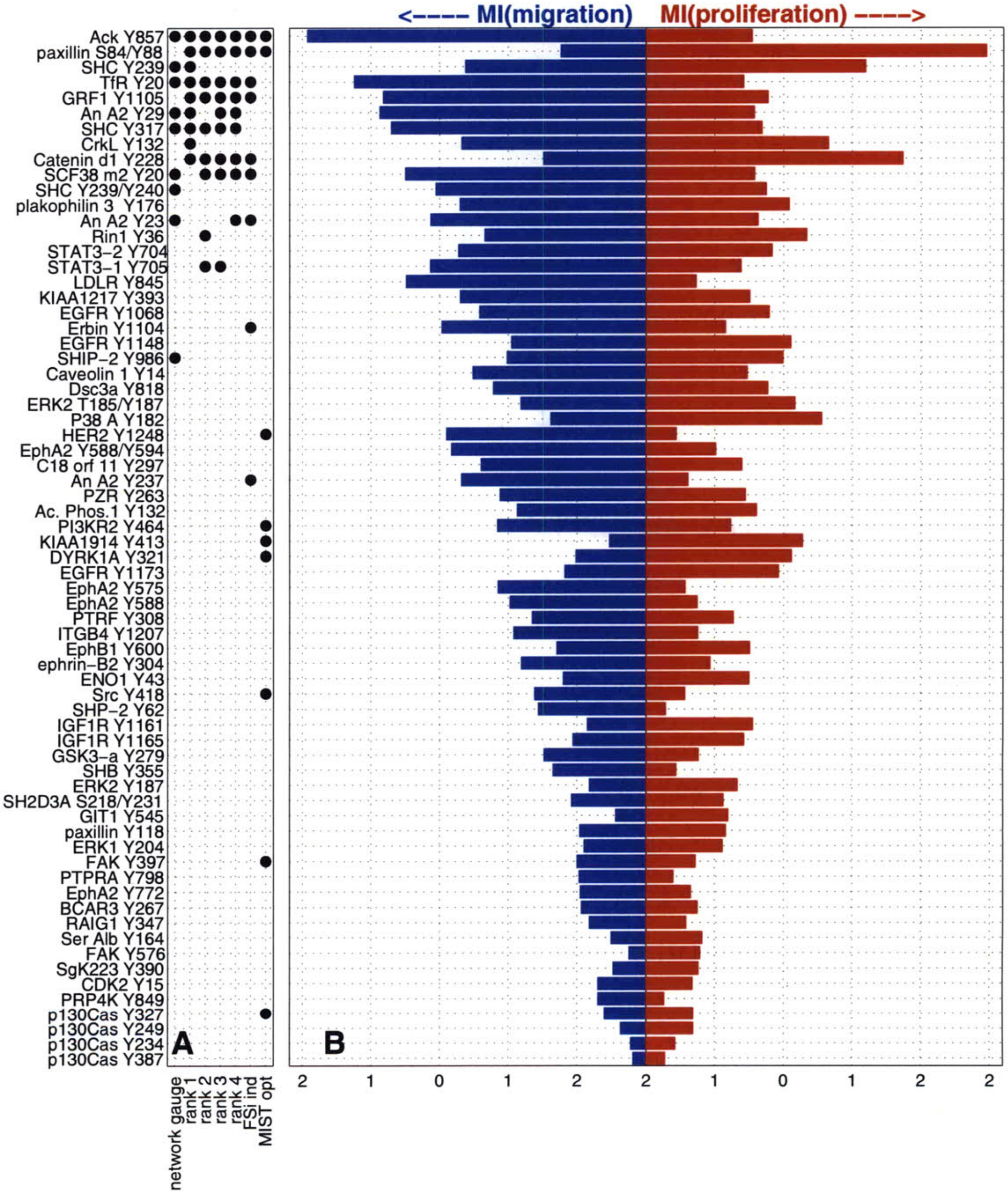


Figure 3-3: **MI of each signal with the outputs**: For each signal, we computed the MI between all four measures and each output individually or both inputs together using the MIST approximation. The species are shown as ranked by MI with both outputs, with the blue bars in panel (B) showing the MI about migration, and the red bars showing the MI about proliferation. The black dots in panel (A) indicate the feature sets chosen by the metrics summarized in Table 3.4 and described in detail in Section 3.2.

Table 3.3: Top scoring pairs of phospho-peptide time points by full MI

MI	phospho-peptide 1 - time	phospho-peptide 2 - time
2.89	C18 orf 11 Y297 - 30 min	C18 orf 11 Y297 - int
2.45	An A2 Y23 - 10 min	An A2 Y23 - 30 min
2.34	SCF38 m2 Y20 - 30 min	SCF38 m2 Y20 - int
2.26	SHC Y317 - 10 min	An A2 Y23 - 30 min
2.25	C18 orf 11 Y297 - 5 min	C18 orf 11 Y297 - 30 min
2.23	SHC Y317 - 5 min	An A2 Y23 - 10 min
2.23	EGFR Y1148 - 30 min	EGFR Y1068 - 30 min
2.22	CrkL Y132 - 5 min	SHC Y239/Y240 - int
2.22	SHB Y355 - 30 min	ITGB4 Y1207 - 5 min
2.22	EphA2 Y588 - 10 min	ITGB4 Y1207 - 10 min

measure of the information that each peptide has about the outputs. The results can be seen in Figure 3-3(B) in which the phospho-sites have been sorted by the MI with both outputs together, and the MI with each of the outputs individually is plotted. For the most part, the signals seem to be particularly informative of either migration (blue bars) or proliferation (red bars), but not both. A few highly informative sites, however, show equivalent information with both outputs.

The signals exhibiting the most information with the migratory response have largely been previously implicated as relevant to migration. The protein Ack is known to regulate cell spreading in HeLa cells [14]. Annexin A2, glucocorticoid receptor DNA binding factor (GRF1), and SHC are also known mediators of the migratory response. While not as intuitive of a signal, TfR has been proposed as an indicator of EGFR activation in this data set through association with endocytosis [47]. The highest ranking signals for proliferation are somewhat surprising. Other than SHC Y239, which has been found to be an early responder to EGFR signaling, the most informative sites for proliferation (paxillin S84/Y88, Catenin d1 Y228, and CrkL Y132) are more strongly associated with the canonical migratory network. One possibility is that these migratory signals may still be reflective of the proliferatory response, even if they are not direct mediators of it. Interestingly, SHC Y239, the third most informative phospho-peptide overall, was computed to have similar MI with both outputs. This is consistent with SHC's role in multiple pathways downstream of EGFR [74].

3.3.2 Choosing maximally informative subsets

We next asked whether our high-dimensional approximation framework could successfully identify informative subsets of signals in the data set. Analyses in multiple biological systems have demonstrated that small subsets of species are often sufficient to describe phenotypic variation [35, 39, 47], and in many cases, limiting the number of species has improved the performance of statistical models [23, 12]. In the context of the current work, previous analysis of the phospho-proteomic data examined in this chapter identified a "network gauge" consisting of 9 of the 68 phospho-sites. PLSR models constructed using only these nine sites had similar fit and cross-validation performance compared to models built using of all the signals.

Our reasons for selecting informative subsets according to our metrics were twofold. First, identifying compact subsets of the signal data that can still capture the response data can be valuable in interpreting the important dimensions of the data set, as well as in developing future studies in which interrogation of the full spectrum of phospho-peptides may not be feasible. Sets of measurements that are highly informative of particular outputs present hypotheses for the signaling mechanisms governing cellular response, and the statistical properties of the sets may be instructive as to desirable properties for applications such as biomarker identification. Secondly, by demonstrating that subsets chosen according to MIST perform well according to previously established metrics (e.g., that the subsets yield models with similar characteristics to models built with all the data), we can validate our information theory framework in the context of phospho-proteomic data. While we have previously validated the performance of MIST in the context of mRNA expression data extracted from tumor samples [46] and in hepatotoxicity response data in multiple cell systems [16], it is important to validate the approximation in applications to new systems.

In order to allow comparison to the previously proposed network gauge, we sought to identify a set of nine phospho-sites with maximum information about the two outputs. As with the previous work, all 4 measures of each site (3 time points and integrated signal) were included if the site was chosen. While the network gauge sites were spread over only six proteins, we did not constrain the number of proteins chosen, as this restriction was not explicitly imposed in the original selection of the network gauge. In the information theory phrasing, we sought to identify the 9 peptides whose 36 total measures had maximal mutual information with the two outputs, as computed by the 2^{nd}-order MIST approximation:

$$\underset{\boldsymbol{s_i}}{\operatorname{argmax}}\left[I^2_{36}(\boldsymbol{s_i};\{m,p\})\right];|\boldsymbol{s_i}|=9 \tag{3.5}$$

such that $\boldsymbol{s_i}$ can take as a value any set of 9 of the 68 potential sites. Due to the size of the search space, enumeration to find the globally optimal set was not feasible. Instead, we employed a greedy selection strategy in which signals were added one at a time to maximize the MI of the newly formed set at each step (i.e., the most informative signal is added first, then the signal that results in the maximum MI when combined with the first signal is, and so forth until 9 signals were chosen). In Chapter 2, we employed a similar scheme to select mRNA expression levels with maximal information for cancer classification [46]. While this scheme does not guarantee the global maximum, it does represent a local maximum of the objective function stated in Equation 3.5. Because this set represents an optimum according to the MIST approximation, we refer to it as MIST opt in the figures and text.

Table 3.4: Feature selection schemes

FS Scheme	**type**	**metric**
Rank 1	rank	$I(\mathit{signal};\mathit{outs})$
Rank 2	rank	$\max\limits_{i} I(\mathit{timepoint}_i;\mathit{outs})$
Rank 3	rank	$\max\limits_{i,j} I(\mathit{timepoint}_i;\mathit{out}_j)$
Rank 4	rank	$\max\limits_{i} R^2(\mathit{timepoint}_i;\mathit{outs})$
FSi ind	incr	$I(\mathit{timepoints};\mathit{outs})$
MIST opt	incr	$I(\mathit{signals};\mathit{outs})$
Network gauge, see [47] for details		

As a point of comparison, we also examined a variety of related selection schemes, in addition to the network gauge and MIST opt sets. While these schemes do not use the exact information theoretic phrasing described in Equation 3.5, they all attempt to identify informative sets in some way. Four of the selection schemes involved ranking all of the signals according to some metric and choosing the top nine signals by rank. The ranking metrics are summarized in Table 3.4 and are described in more detail in Section 3.2. We also employed an additional incremental selection scheme in which each of time points was able to be selected individually. Each signal was then scored by the best rank achieved by any of its four associated measures. All together, we examined the two incremental schemes (in which complementary sets of features were chosen) and four ranking schemes (in which each feature was individually scored and the top scoring individual features were chosen) as well as the previously proposed network gauge.

The results of applying the selection schemes can be seen in Figure 3-3(A), where the black dots indicate the nine selected signals for each scheme. Despite the fact that only one of the ranking metrics used the combined MI with both outputs as its ranking metric, all four ranking schemes tend to select signals near the top of the global MI list; none of the ranking metrics chose a signal outside of the top 25% most informative. Additionally, although the network gauge was chosen based on the weightings in a series of PLSR models, and not using any information theoretic metrics, the consitutent signals also tend to occupy the top of the most informative list. Eight of the nine signals rank in the top 20% by MI, including three of the top four (Ack Y857, SHC Y239, TfR Y20). Only SHIP-2 Y986 appeared lower in the list (ranked 22/68). In contrast, the optimal MIST set (MIST opt) chose signals spread throughout the spectrum of total MI. While this set did include the two most informative signals (Ack Y857 and paxillin S84/Y88), all other signals fell outside the top 35% of the most individually informative list. The other incremental scheme (FSi Ind) showed intermediate behavior, selecting many individually informative signals (6 of the top 10), but also including some lower information signals, such as AnA2 Y237, ranked 30/68.

3.3.3 Evaluating the feature sets by PLSR modeling

In order to evaluate the subsets chosen by our various selection schemes, we built statistical models of the output data using each subset as an input. This method of validation was previously used to show that the 9-signal network gauge had similar fit and predictive power to a model built with all the signals [47]. In that work, PLSR was used to generate the models, which was a natural choice given that the network gauge was originally chosen based on the analysis of PLSR models. For our case, information theory does not provide an integrated predictive modeling framework with which to validate the sets. As such, we choose to build PLSR models with each of the sets, in order to evaluate the set choices, and to enable comparison with previous work. To this end, we built one- and two-component PLSR models separately for each set of nine signals and each of the two outputs. These models were then scored by their fit to the data (R^2) and by a cross-validation metric (Q^2). Additional details are available in Section 3.2.

The performance of models generated using each of the feature sets can be seen in Figure 3-4. Panels A and B show the model fits for the one- and two-component models respectively, and panels C and D show the cross-validation scores. For all plots, the x- and y-axes show performance for separately modeling the migration and proliferation response, respectively. If one views ability to model the two outputs as equally important, the dashed lines represent contours of equivalent overall performance, with better-performing models lying higher and further to the right. In addition to the 7 selection schemes described above (colored $\times$'s), a histogram of the performance of 1000 models built with randomly selected sets of nine signals are shown (gray heatmap). For the most part, the selection schemes tend to fall on similar contours to each other, trading off poor performance in one dimension for success in the other. Surprisingly, all of the ranking schemes fall within the randomly selected distribution, indicating that although they were selected to be informative, the signals were not enriched in their ability to model the outputs. In contrast, the optimal MIST set (black $\times$'s) generates consistently better models than random. The other incremental method also shows good performance in most cases, particularly in the cross-validation metric of the two-component model. While the network gauge (blue $\times$'s) generates good PLSR models of migration, it generally does so at the expense of accuracy in proliferation. In summary, the set chosen to maximize the

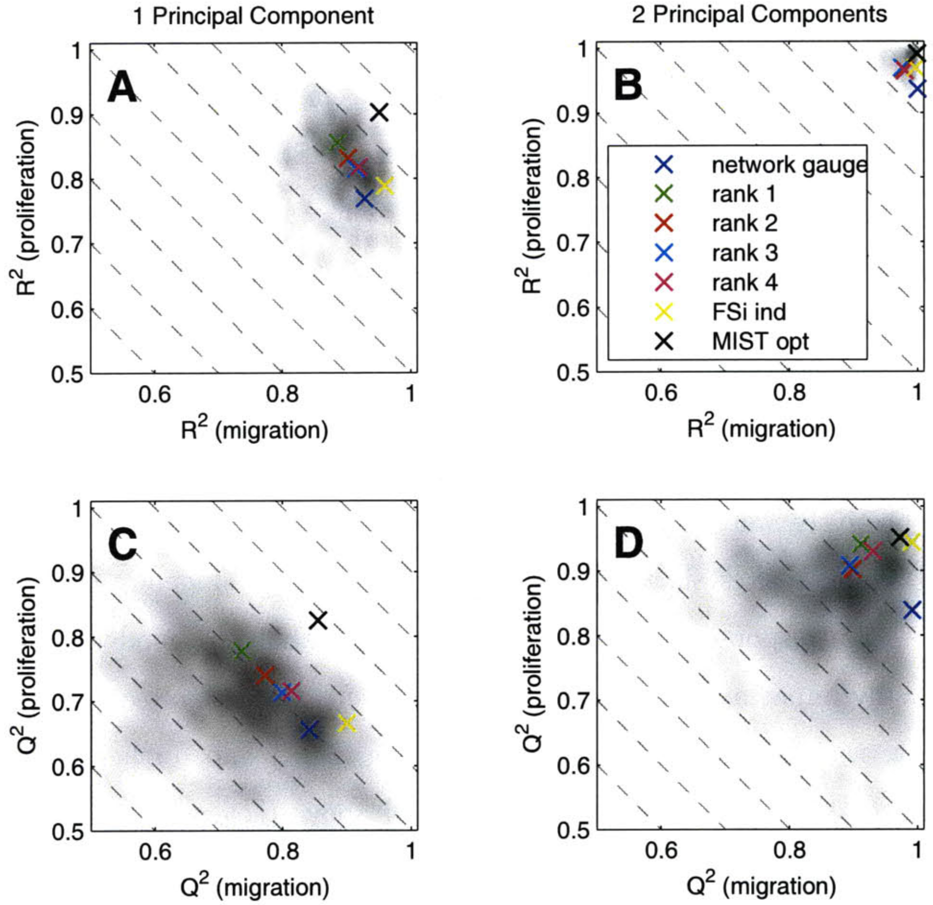


Figure 3-4: **PLSR models built with subsets**: We built PLSR models mapping the 36 measures in each selected feature set shown in Figure 3-3 against the proliferation or migration response data. Models containing one (panels A and C) or two principal components were then scored by their fit (panels A and B) and cross-validation (panels C and D) performance against migration (x-axis) or proliferation (y-axis). Dashed lines indicate contours of equal average performance for the two outputs.

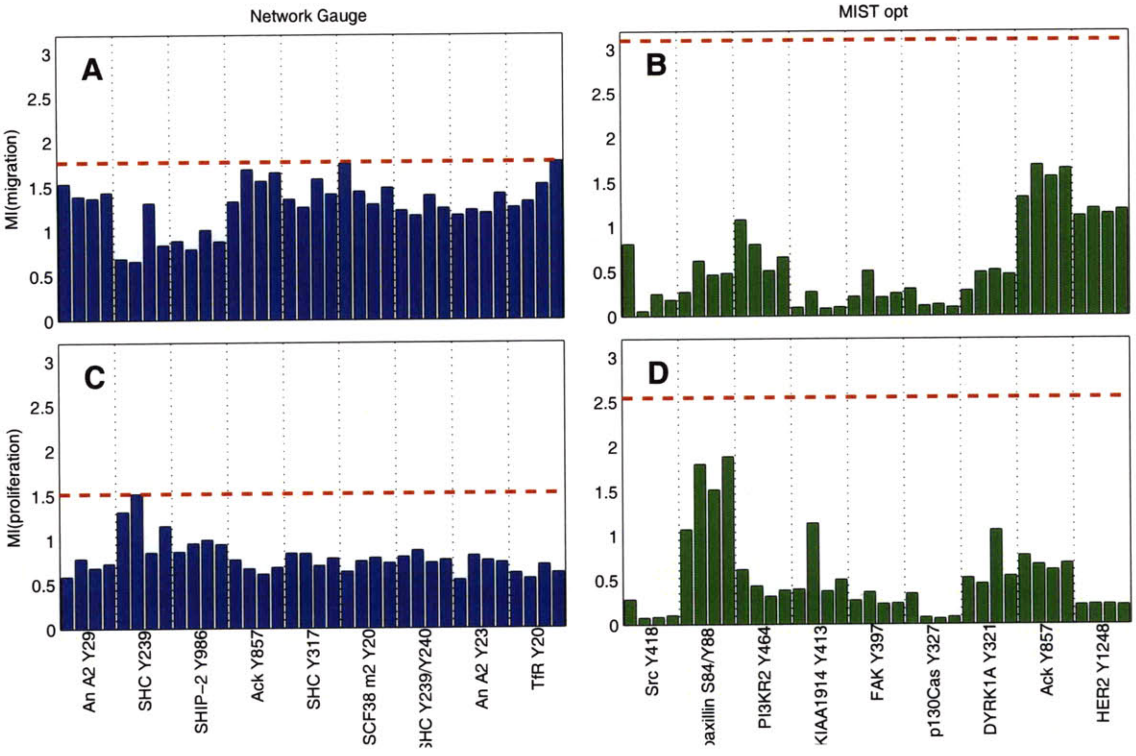


Figure 3-5: **MI of feature sets without output**: The MIs for migration (panels A and B) or proliferation (panels C and D) from each of the four measures for each of the nine phospho-peptides in the network gauge (panels A and C) or MIST-selected (panels B and D) features sets are shown. For each signal, the four bars represent, in order, the 5-, 10-, and 30-minute time points, and the integrated signal. The dashed red lines show the total mutual information between all 36 variables shown in each panel and the output of focus.

information content about the two outputs according to MIST generates significantly improved models in all tested scenarios. The fact that none of the ranking metrics performed significantly better than random and that the FSi ind and network gauge showed poorer performance for proliferation demonstrates the difficulty of choosing sets of signals that will reliably generate improved models. Alternatively, the good performance of randomly selected sets speaks to the rich information content of many of the signals in the data.

3.3.4 Properties of the optimal MIST set

Having validated that the maximally informative set generates improved models, in addition to being computed to have high mutual information by MIST, we examined the properties of the chosen signals in the context of the rest of the data set. In particular, as noted in Section 3.3.2, only two of the nine signals were computed to be particularly informative on their own. Furthermore, these two most informative signals were included in all of the other selection schemes except for the network gauge, which included one of the two. Despite this, the MIST opt set generated enhanced models compared to all other selection schemes, suggesting that these low-information signals were in fact important to the performance. A detailed view showing the information that each of the 36 measures in each feature set have about each output can be seen in Figures 3-5 and B-1. In this view, it is clear that the MIST set (green bars in Figure 3-5) are on average less individually informative than those in the network gauge (blue bars). In contrast, the total computed information between the full set of 36 measures with each output (dashed red lines), shows the MIST set to be significantly more informative overall. For the network gauge, the most informative individual measure for each output (integrated signal of TfR Y20 for migration, and 10 minute time point of SHC Y239 for proliferation), is as informative as the full set. In other words, the other 35 measures seem to provide redundant information about the outputs once the first measure is included. In contrast, the total information of the MIST set is significantly higher than any individual signal. In this case, the multiple signals, though lower in information on their own, provide unique information so as to improve the overall information content of the full set.

Given that the signals in the network gauge seemed to be providing redundant information, we next looked at the relationships shared between the constituent signals. The mutual information of all pairs in each set is shown in Figures 3-6 and B-2. Focusing on the comparison between the network gauge and the MIST chosen set (Figure 3-6) it is clear that the signals in the network gauge (panel A) are highly coupled with each other whereas the signals from the optimal MIST set (panel

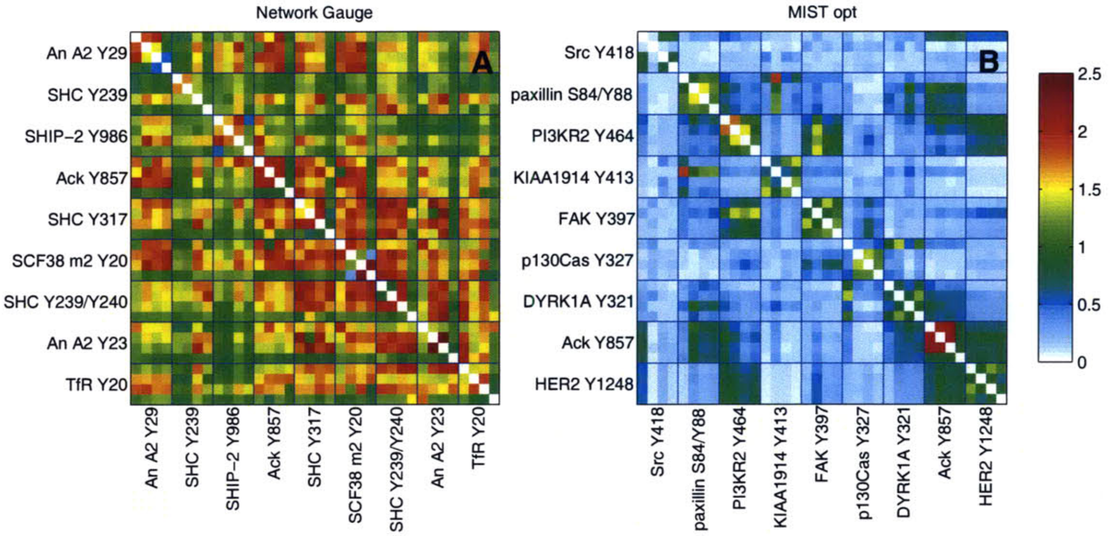


Figure 3-6: **MI matrices of feature sets**: The MI between each pair of signals in the network gauge (panel A) or optimal set by MIST (panel B) is shown. The signals in the network gauge are highly informative of each other compared to the relatively independent signals in the MIST set.

B) are relatively independent. These matrices are consistent with the explanation that a good set of signals needs not only to be informative of the outputs, but also sufficiently distinct from each other so as to provide unique information. The high-dimensional phrasing enabled by MIST was able to appropriately weight these two aspects against each other to generate a highly informative set comprised of some individually informative signals as well as some less informative species with unique information.

The trends seen in these MI matrices are also reflected in previous work. Among the analyses carried out in the original publication of this data was the application of self-organizing maps to identify clusters of phospho-sites exhibiting similar trends across the treatment and cell-line conditions. Of the nine sites selected by MIST, no more than two were seen in any of the four major clusters: paxillin S84/Y88 co-clustered with Ack Y875, and FAK Y397 with PI3K Y464. In contrast, all nine of the sites from the network gauge appeared in the same cluster as each other. These results, coupled with the validation of the MIST set by PLSR, further demonstrate that independence of the signals may be an important property of good feature sets.

3.3.5 Biological relevance of optimal MIST set

In contrast to the statistical independence of the phospho-peptides identified by MIST, the biological functions of these signals present a fairly coherent picture. Of the nine signals selected, seven are known to be associated with cell migration. FAK-Y397 is known to act as a docking site for both Src and PI3K and is required for the phosphorylation of p130cas and paxillin in mediating migration [8]. The phosphorylation of four of these five molecules (excluding PI3K) were previously identified as key responses to HRG stimulation of HER2-overexpressing cells in this same data set [88]. Tyrosine 1248 on HER2 has been shown to be necessary for migration in the context of breast cancer cell lines [24], and was also identified as a key upstream activator of migration in the PLSR modeling work performed on this data [88]. Ack has been shown to regulate cell spreading in HeLa cells through p130cas and CrkII [14] and has also been identified as an early transducer of multiple stimuli, including EGF [30]. The remaining two signals, KIAA1914-Y413 and DYRK1A-Y321 have yet to be well characterized in the context of EGFR signaling, although DYRK1A has been shown to be involved in MAPK signaling through Ras–Raf–MEK [43]. The notable absence of proliferation-specific signals is somewhat surprising, but may be explained by the fact that EGF stimulation (which is the dominant driver of proliferation in these data) was observed to activate a multitude of different pathways in this data set [88]. As such, the monitoring of any of a number of signals downstream of EGFR may provide sufficient information to accurately represent the proliferation response.

3.4 Discussion

The collection and analysis of large multivariate data sets is at the core of systems biology research. The monitoring of sets of molecular species on a near global scale, however, has not been accompanied by the ability to examine enough different conditions to directly query the multivariate relationships in these systems. The so-called Curse of Dimensionality suggests that the number of experimental conditions needed to directly model such relationships scales exponentially with the number of species. We have previously demonstrated that many of the multivariate relationships in biological data can be reasonably well approximated by appropriately combining relationships of lower order, such as the mutual information between each pair of species [46]. Here, we have applied this approximation framework to analyze an existing phospho-proteomic data set with significantly more signals (272) than experimental conditions (6). Within the context of our approximation framework, we have identified a set of 36 signals, representing the state of 9 phospho-peptides, that can accurately model the migration and proliferation response across multiple conditions. The success of this set of signals seems to be crucially tied not only to a pair of individually informative signals, but also to a set of lower-information signals that might otherwise be overlooked. Furthermore, these lower-information signals include four sites that were previously identified as being key responders to heregulin in the context of HER2 overexpression in this data [88]. This work highlights the importance of using phrasings that can capture multivariate statistical relationships when the questions being addressed are fundamentally multivariate in nature (e.g., identifying a maximally informative set of signals).

The results of these feature selection schemes may also have implications to the related field of biomarker identification. In particular, in cases where a single biomarker with sufficient statistical power cannot be identified, the relationships between multiple candidate biomarkers may become important. In the current work, the features that separated the optimal MIST set from the other sets were not the most individually informative species, but instead were lower information species that provided complementary information. In larger scale studies where multiple hypothesis corrections are required, these intermediate species may be overlooked. The results of this work suggest that such species may merit re-examination in the context of previously chosen biomarkers. While we have not examined it here, additional high-dimensional phrasings might also be relevant to biomarker selection. For example, information theory provides phrasings to account for confounding variables that can introduce spurious statistical relationships. When such confounders are known to exist, they could be explicitly accounted for by conditioning all information terms on the confounders during the selection scheme. In so doing, biomarkers could be selected that provide information about the outputs that is independent of variation explained by the confounders. While we have not examined such phrasings directly, the generality of our approximation framework can enable them, even when data sizes are limited.

In addition to analytical insights gained by selecting an informative subset, the modeling performance of this set helps to validate the approximation framework in the context of this data set. We have therefore used the approximation to combine various dimensions of the data set to enable simpler interpretation. Figure 3-2 shows an example of such analysis in which we have combined the four measures associated with each signal to generate a single metric representing the relationships between phospho-peptide pairs, as opposed to the previously available 16 relationships between all of the individual time points as shown in Figure 3-1. As discussed in Section 3.3.1, the strongest peptide–peptide relationships in the grouped MI matrix reflect known biology, including trivial results (such as multiple sites on EGFR) as well as phosphosites on proteins that are known to interact (SHC and CrkL), or known to participate in a coherent signaling cascade (EGFR and Ack).

We have also used high-order mutual information calculations enabled by MIST to quantify the statistical relationships between the phospho-peptide profiles and migratory or proliferatory responses across the experimental conditions, as shown in Figure 3-3. This analysis identifies signals that are particularly informative of the cellular response in the context of the data set. While many of the highly informative signals are predominantly informative of only one of the two responses, a few show strong information with both migration and proliferation. In particular, the third-most informative signal overall, SHC Y239, is computed to be equally informative of both outputs. This result is consistent with the established role of SHC as a key signal downstream of EGFR in mediating a variety of responses [74]. The highly informative signals according to the MIST calculations also agree surprisingly well with the previously identified network gauge. Given that the network gauge was selected based upon the behavior of the signals in a regression model, and not on any information theoretic statistics, this consistency serves as a validation of both modeling approaches.

In addition, the distribution of the selected feature sets across the spectrum of information values further demonstrates the non-obvious results that multivariate phrasings can provide. Whereas all other examined selection schemes heavily favored individually informative signals, the optimal set chosen by MIST included many signals that are not computed to be particularly informative individually. Given that many of the other feature sets also included the most informative metrics in the MIST opt set, the significantly better performance of models built using the MIST opt set (Figure 3-4) seems to be a result of these lower-information signals. The relevance of the MIST set is also supported by the good concordance with previous observations that paxillin, Src Y418, FAK Y397, and p130cas Y327 constitute a set of sites that exhibit a unique response to heregulin stimulation in the context of HER2 overexpression [88]. Furthermore, two other members of the MIST set, HER2 Y1248 and PI3K Y464 are also implicated in cellular migration through this same pathway.

While we have not examined them here, a variety of related analyses are enabled by our information theoretic phrasing. For example, the phospho-peptides present in a single protein or known to be substrates of a particular kinase could be grouped together, providing an overall view of the information content of these sets of related signals. Alternatively, one could group multiple species at a specific time point to track information flow through time in the network. These types of analyses may also prove useful for experimental design applications, where a subset of time points, measurements, or experimental conditions must be selected prior to collection of a full data set. We have previously examined such applications in the context of idiosyncratic drug toxicity studies across multiple cell systems, finding that well selected subsets of experimental conditions provided similar information content to the full data set [16]. All of these phrasings would traditionally require the calculation of high-dimensional statistics which cannot be reliably computed directly given the relatively small number of experimental conditions. The MIST approximation, however, enables such groupings even when sample sizes are small. Furthermore, as more data become available, MIST provides a series of approximations (described in Chapter 2) that provide more accurate calculations, without altering the fundamental problem phrasing.

Chapter 4

Efficient calculation of molecular configurational entropies using an information theoretic approximation

4.1 Introduction

A fundamental goal of computational chemistry is the calculation of thermodynamic properties of molecules, such as the chemical potential, enthalpy, and entropy. Accurate calculation of such properties can enable computational design and screening at a scale infeasible in experimental systems, and provides tools for detailed computational analysis of molecules of interest. While early work focused largely on characterizing single configurations, often representing the global minimum energy conformation, advances in computing technology have increasingly enabled the investigation of configurational ensemble properties [10, 45, 41]. This work, as well as recent experimental studies using NMR, highlight the importance of configurational solute entropy in a variety of systems [10, 52]. As such, improving the accuracy and speed of molecular ensemble based calculations, particularly in larger systems, is an area of active research.

One class of approaches for computing configurational averages centers around the use of sampling based simulations such as molecular dynamics (MD) and Monte Carlo. Such methods may be particularly well suited for larger systems, such as proteins, where explicit enumeration and characterization of all relevant minima is infeasible [45]. One of the better known methods in this field is the quasiharmonic approximation, which approximates the system as a multidimensional Gaussian using the covariance matrix computed across aligned simulation frames [41]. While successful in many cases, the quasiharmonic approximation has been shown to significantly overestimate entropies in systems containing multiple unconnected minima, which are poorly modeled by a single Gaussian [9, 36]. Recent phrasings have instead focused on estimating probability densities over the configuration space of a molecule using the frames from MD simulations [37, 45]. As system size grows, however, direct estimation of the density over all molecular degrees of freedom (DOF) becomes infeasible. To address this issue, a mutual information expansion (MIE) of the configurational entropy has recently been developed that enables approximation of configurational entropies as a function of lower-dimensional marginal entropies [45]. The MIE framework has proved accurate in the analysis of a variety of small molecule systems [45], and has been combined with nearest-neighbor methods to improve convergence [38]. It has also been used in the analysis of side-chain configurational entropies to identify residue-residue coupling in allosteric protein systems [63].

As discussed in Chapters 2 and 3, in parallel work developed in the context of gene expression and cell signaling data, we have generated a similar framework, MIST, that provides an upper bound to Shannon's information entropy as a function of lower-order marginal entropy terms [46]. For multiple synthetic and biological data sets, we found that, in addition to acting as a bound, the MIST approximations generated useful estimates of the joint entropy. Due to the mathematical relationships between information theory and statistical mechanics, application of MIST to the calculation of molecular entropies proved feasible with relatively little adaptation. While similar in spirit to MIE, MIST represents a distinct framework for approximating high-dimensional entropies by combining associated low-order marginal entropies. In this chapter, we examine the behavior of MIST when used to calculate molecular configurational entropies from MD simulation data, and in the context of idealized rotameric systems.

We start by evaluating MIST in the analysis of MD simulations of a series of small linear alkane systems where MIE has previously been shown to agree well with established methods. We observe that MIST demonstrates larger deviations than MIE when compared against the Mining Minima (M2) method that is currently among the most accurate tools for computing ensemble properties of small molecules [10]. We also observe, however, that MIST converges considerably faster than MIE as a function of simulation time, particularly for higher approximation orders. As such, for the systems tested, while the converged MIE approximation shows better concordance with M2 than does MIST, sampling regimes exist for which MIST provides closer agreement.

Although M2 provides a reasonable "gold standard" to compare against, approximations inherent in the M2 method result in inconsistencies with the underlying molecular ensembles sampled by MD simulations [45]. We therefore also examine MIST and MIE in the context of a series of idealized rotameric molecular systems in which the marginal entropies can be computed exactly. These systems enable evaluation of the approximation frameworks separate from the errors introduced from sampling. In contrast to the MD results, whereas MIST exhibits small errors for all discrete systems, MIE demonstrates erratic convergence with increasing approximation order, even when the marginal terms are determined exactly. These differences are particularly pronounced in constrained systems in which molecules are bound to rigid proteins. Finally, we examine the convergence properties of MIST and MIE by sampling from the discretized systems. Unlike the MD systems, evaluation of the convergence to the analytically exact value for each approximation is possible. Similar to the MD results, MIST exhibits improved convergence relative to MIE for all systems.

Having multiple distinct methods for computing configurational ensemble properties may prove useful when standards such as M2 and the enumerated rotameric systems are unavailable, as is likely to be the case for larger systems such as proteins. Even in the cases where the MIE accuracy proves to be superior, the MIST framework contains guarantees that may be useful for future applications. In particular, MIST demonstrates monotonically decreasing approximation error with order of approximation, and bounding of the entropy, when the system is well converged. Additionally, a variety of enhancements and applications of MIE have been explored recently [38, 63]. While we have not pursued them here, the existing literature is likely to be extensible to MIST, and may yield interesting results in that context.

4.2 Theory

In this section, we review the MIST approximation in the context of configurational entropies. Additional details of MIST have been published previously in the context of analyzing mRNA expression data for cancer classification [46], and are presented in Chapter 2. Here we primarily highlight the theoretical differences between MIST and MIE.

The information theoretic phrasing of the calculation of configurational entropies has been well described previously [45]. The key step of the phrasing comes from representing the partial molar configurational entropy of a molecule as

$$-TS^{\circ} = -RT \ln \frac{8\pi^2}{C^{\circ}} + RT \int \rho(\boldsymbol{r}) \ln \rho(\boldsymbol{r}) d\boldsymbol{r}, \tag{4.1}$$

where R is the gas constant, T is the temperature, C° is the standard state concentration, and ρ is the probability density over the configurational degrees of freedom, $\boldsymbol{r}$. For the purposes of this paper, $\boldsymbol{r}$ is represented in a bond-angle-torsion (BAT) coordinate system, as opposed to Cartesian coordinates. BAT coordinates tend to be less coupled than Cartesian coordinates for molecular systems, and are thus well suited for low-order approximations [73]. The first term of the RHS represents the entropic contribution of the six rigid translational and rotational degrees of freedom, and is found via analytical integration, assuming no external field. The second term, when negated, is identical to RT times the information entropy, S, as originally developed by Shannon [77], providing the equation

$$-TS^{\circ} = -RT \ln \frac{8\pi^2}{C^{\circ}} - RTS, \tag{4.2}$$

$$S = -\int \rho(\boldsymbol{r}) \ln \rho(\boldsymbol{r}) d\boldsymbol{r}. \tag{4.3}$$

This relationship allows techniques developed in the context of information theory to be used for the calculation of configurational entropies.

The MIST framework provides an upper bound to the Shannon information entropy using marginal entropies of arbitrarily low order. The approximation arises from an exact expansion of the entropy as a series of conditional entropies, or alternatively, as a series of mutual information terms,

$$S_n\left(\boldsymbol{r}\right) = \sum_{i=1}^{n} S_i\left(r_i | \boldsymbol{r}_{\boldsymbol{1}\ldots\boldsymbol{i-1}}\right) = \sum_{i=1}^{n} \left[S_1\left(r_i\right) - I_i\left(r_i; \boldsymbol{r}_{\boldsymbol{1}\ldots\boldsymbol{i-1}}\right)\right], \tag{4.4}$$

$$I\left(\boldsymbol{x}; \boldsymbol{y}\right) = \int \rho_{x,y}(\boldsymbol{x}, \boldsymbol{y}) \frac{\rho_{x,y}(\boldsymbol{x}, \boldsymbol{y})}{\rho_x(\boldsymbol{x})\rho_y(\boldsymbol{y})} d\boldsymbol{x} d\boldsymbol{y}, \tag{4.5}$$

where $I_i\left(r_i; \boldsymbol{r}_{\boldsymbol{1}\ldots\boldsymbol{i-1}}\right)$ is the mutual information (MI) between DOF r_i and all DOF that have already been included in the sum. Throughout this section, subscripts on S or I indicate the order of the term, i.e., the number of dimensions in the PDF needed to compute the term. The MI phrasing can be thought of as adding in the entropy of each DOF one at a time (S_1 terms in Equation 4.4), then removing a term corresponding to the coupling between that DOF and all previously considered DOF (I_i terms). The MIST approximation consists of limiting the number of DOF in the information term. For example, for the first-order approximation, all coupling is ignored, and the I term is completely omitted from the formulation. For the second-order approximation, when each DOF is added, its coupling with a single previously chosen DOF is accounted for, as opposed to considering the coupling with all previously included terms

$$S_n(\boldsymbol{r}) \leq S_n^{MIST_2}(\boldsymbol{r}) = \sum_{i=1}^{n} S_1(r_i) - \sum_{i=1}^{n} \max_{j<i} I_2(r_i; r_j) \tag{4.6}$$

Because removing terms cannot increase the information, the RHS is an upper bound on the entropy. Furthermore, all ordering of indices i and choices of conditioning terms j provide valid upper bounds. As such, we can optimize for the order and conditioning terms that minimize the RHS to generate the tightest bound consistent with the framework. To generate approximations of arbitrarily high order, k, we include an increasing number of DOF in the mutual information,

$$S_n(\boldsymbol{r}) \leq S_n^{MIST_k}(\boldsymbol{r}) = \sum_{i=1}^{n} \left[S_1(r_i) - \max_{\boldsymbol{j}<i} I_k(r_i; \boldsymbol{r_j}) \right]; |\boldsymbol{r_j}| \leq k-1 \tag{4.7}$$

where $\boldsymbol{r_j}$ is a vector of length $k-1$ representing any subset of DOF $\in \{r_1 \ldots r_{i-1}\}$.

In the context of approximations to thermodynamic ensemble properties, MIST bears a strong resemblance to the Bethe free energy (also know as the Bethe approximation) [5]. In fact, the second-order MIST approximation is equivalent to the Bethe approximation, and the full MIST framework may thus be thought of as a high-order generalization to the Bethe free energy. While a full comparison of MIST and Bethe approximation is outside the scope of the current work, a number of modifications and applications of the Bethe approximation have been explored that may be extensible to MIST [66, 90]

In contrast to MIST, MIE [45] expands the entropy as a series of increasingly high-order information terms, as previously formulated by Matsuda [62]:

$$S_n(\boldsymbol{r}) = \sum_{i=1}^{n} S_1(r_i) - \sum_{i=1}^{n} \sum_{j=i+1}^{n} I_2(r_i; r_j) + \sum_{i=1}^{n} \sum_{j=i+1}^{n} \sum_{k=j+1}^{n} I_3(r_i; r_j; r_k) - \ldots, \tag{4.8}$$

where I is defined as

$$I_n(r_1; \ldots; r_n) = \sum_{k=1}^{n} (-1)^{k+1} \sum_{i_1 < \ldots < i_k} S_k(r_{i_1}, \ldots, r_{i_k}), \tag{4.9}$$

and the second summation runs over all possible combinations of k DOF from the full set of $\{r_1 \ldots r_n\}$. MIE generates an approximation to the full entropy by truncating all terms of order larger than k in Equation 4.8. The approximation will converge to the true entropy when no relationships directly involving more than k DOF exist in the system. Notably, MIE does not carry any bounding guarantees, but it does not require the optimization utilized in MIST.

Despite relying on differing expansions, MIST and MIE share many similarities. The first-order approximation is identical in both cases (summing all first-order entropies). For the second-order approximation, MIE adds in all first-order entropies and subtracts off all possible pairwise mutual information terms,

$$S_n(\boldsymbol{r}) \approx S_n^{MIE_2}(\boldsymbol{r}) = \sum_{i=1}^{n} S_1(r_i) - \sum_{i=1}^{n} \sum_{j=i+1}^{n} I_2(r_i; r_j) \tag{4.10}$$

In contrast, MIST adds in all first-order entropies, and then subtracts off $n-1$ of the information terms (where n is the number of DOF in the system), as is seen in Equation 4.6. These terms are chosen to account for as much information as possible, while still guaranteeing an upper bound. The second-order approximations highlight the theoretical differences between MIST and MIE. Whereas MIE removes all pairwise couplings, effectively assuming that no higher-order relationships exist, MIST removes a subset of couplings, effectively assuming some structure about the system. In particular, MIST will provide a good approximation if the majority of the degrees of freedom in the system are directly coupled only to a small number of other DOF. Such a system can be well covered by the $n-1$ terms included in MIST. In contrast, MIE may not provide a good approximation in such a system due to indirect couplings that are likely to exist between DOF, and must be removed by higher-order terms. Alternatively, in systems containing a larger number of direct pairwise interactions and relatively few higher-order couplings, MIST may provide a poor approximation relative to MIE. Given these differences in representation, we have performed a series of computational experiments to evaluate the performance of the MIST and MIE in a variety of molecular systems, which have helped to reveal how coupled coordinates contribute to configurational entropy.

4.3 Methods

4.3.1 Molecular dynamics simulations

All molecular dynamics simulations were run using the program CHARMM [7] with the CHARMm22 all atom parameter set [59, 60]. Partial atomic charges were fit using the program GAUSSIAN03 [29]. All simulations were run at a temperature of 1000 K using a distance-dependent dielectric of four with a one fs time-step with Langevin dynamics and the leapfrog integrator. A 1 ns equilibration was performed prior to a 50 ns production run from which frames were extracted at a frequency of 1 frame per 10 fs, yielding 5 million frames per simulation.

For each system, an internal coordinate representation consisting of the selection of three seed atoms, as well as a single bond, angle, and dihedral term for each subsequent atom was chosen so as to use improper dihedrals whenever possible, and to place heavy atoms prior to hydrogens. Only bonds, angle, and dihedral terms between chemically bonded atoms were allowed as coordinates. Other than these restrictions, the specific coordinates were chosen arbitrarily. The values of each bond, angle, and dihedral were extracted from the simulations and binned. Marginal probability density functions (PDFs) of all single, pairs, and triplets of coordinates were computed using the frequencies from the simulation. These PDFs were then used to compute the first-, second-, and third-order entropies and information terms. All first- and second-order terms were computed using 120 bins per dimension, and all third-order terms were computed using 60 bins per dimension. For MIE, third-order information terms containing any bond or angle DOF were set to zero, as was done previously [45]. For MIST, all third-order terms were included, as doing so did not dramatically impact numerical stability. All calculations included a Jacobian term of $\prod b_i^2 \sin \theta_i$ where b_i and θ_i are the bond length and angle used to place atom i, and the product runs over all DOF included in the marginal term.

As a point of comparison, we used previously reported values of $-TS^{\circ}$ computed using the Mining Minima (M2) method [10, 45]. Although small differences between our energy function and that used to generate the M2 results exist, the good agreement between our recomputed MIE results and the reported MIE values (see Figure C-1), as well as with M2 (see Figure 4-1) suggests that the M2 results remain a valid comparison. To enable comparison to M2, a factor of $\ln 3$ for each methyl group, and $\ln 2$ for cyclohexane, was subtracted from S to account for the symmetry of methyl rotations and cyclic flip states, respectively.

4.3.2 Discrete rotameric systems

Discrete rotameric systems representing four candidate drug molecules, either unbound or in the binding pocket of a rigid HIV-1 protease were generated. Each system consists of the 5×10^4 lowest energy rotameric configurations, accounting for $> 99\%$ of the contributions to the free energy at 300 K in all cases. For the current work, these 5×10^4 configurations were treated as the only accessible states of the system, enabling exact calculation of all ensemble properties.

The low energy configurations were determined via a two step, grid based, enumerative configurational search. All ligands are comprised of a common chemical scaffold with variable functional groups at 5 possible positions (see Figure 4-5). We first collected an ensemble of low energy scaffold conformations using an enumerative Monte Carlo (MC) search. Ten independent simulations of 5×10^4 steps were performed for each ligand in both the bound and unbound states, and the external and scaffold degrees of freedom of all collected configurations were idealized to a uniform grid with a resolution of 0.1 Å and 10°/20° (bound/unbound). All simulations were performed using CHARMM [7] with the CHARMm22 force field [65] and a distance-dependent dielectric constant of four. The result of the first step was a set of energetically accessible rotameric scaffold configurations.

The second step exhaustively searched the configurational space of the remaining functional group degrees of freedom for each collected scaffold using a combination of the dead-end-elimination (DEE) [21, 19, 20] and A* algorithms [51] as described previously [3]. For high throughput energy evaluations, a pair-wise decomposable energy function was used that included all pairwise Van der Waals and Coulombic, intra- and inter-molecular interactions, computed with the CHARMm22 force field and a distance-dependent dielectric. Uniformly sampled rotamer libraries for each functional group with resolutions of 15° or 60° for the bound or unbound states, respectively, were used. The 5×10^4 lowest-energy configurations across all scaffolds were enumerated, and their energies computed.

The top 5×10^4 low-energy configurations from each ensemble were re-evaluated using a higher resolution energy function to account for solvation effects and obtain a more accurate measure of the energy. The enhanced energy function included all pairwise Van der Waals interactions, continuum electrostatic solvation energies collected from a converged linearized Poisson-Boltzmann calculation calculated using the Delphi computer program [31, 68], as well as solvent accessible surface area energies to model the hydrophobic effect [18]. Solvation energies were calculated using an internal dielectric of 4 and a solvent dielectric of 80. A grid resolution of $129 \times 129 \times 129$ with focusing boundry conditions [57] was used, along with a Stern layer of 2.0 Å and an ionic strength of 0.145 M.

Given the energies of all configurations in the idealized rotameric systems, entropies of arbitrary order were computed analytically by integrating through the Boltzmann distribution. To evaluate the convergence properties of the metrics in the context of the discrete rotameric systems, we randomly drew from the 5×10^4 structures representing each system with replacement according to the Boltzmann weighted distribution. The resulting samples were then used to estimate the single, pair, and triple PDFs as for the MD systems. Because the exact marginal entropies are analytically computable, convergence for these systems was examined with respect to the same approximation computed using the analytically-determined marginal terms. No symmetry adjustments were applied for the discrete systems.

4.4 Results

4.4.1 Molecular dynamics simulations of small alkanes

To investigate the behavior of the MIST framework in the context of configurational entropies, we first examined a series of linear alkanes (butane – octane), as well as cyclohexane. Configurational entropies for all of these systems have been previously computed using MIE and were shown to agree well with M2 calculations [45]. As was done in these studies, we collected 5×10^6 frames from a 50 ns molecular dynamics trajectory for each molecule and computed the single, pair, and triplet entropies of all BAT degrees of freedom. We then combined these marginal entropies according to the MIST (Equation 4.7) or MIE (Equation 4.8) framework, using approximation orders of one, two, or three. The resulting values for the entropic contribution to the free energies, $-TS^\circ$ (computed using Equation 4.2), are shown in Figure 4-1.

As seen in the previous studies of MIE (red bars), the second-order approximation (MIE_2) shows good agreement with M2 (dashed line) for all linear alkanes, with a maximum difference of 1.2 kcal/mol. MIE_1 and MIE_3 generally show worse agreement with M2 ($>$ 10 kcal/mol in some cases) as previously reported. As with previous studies, none of the approximation orders agree well with M2 for cyclohexane. The MIST approximations (blue bars), show somewhat different behavior than MIE. As guaranteed by the theory, the first order MIST and MIE approximations are identical. MIST_2, however, shows considerably larger deviations from M2 for the linear alkanes (3–7 kcal/mol) than does MIE_2. Also, whereas MIE_3 generally showed worse agreement with M2 than MIE_2, MIST_3 improves upon MIST_2 for all systems, showing deviations from M2 between 2 and 4 kcal/mol for linear alkanes. While MIST_3 is guaranteed to yield at least as accurate of a result as MIST_2 when both are fully converged, it is important to see it in the context of finite sample sizes. As with MIE, none of the MIST approximations compare favorably with M2 for cyclohexane.

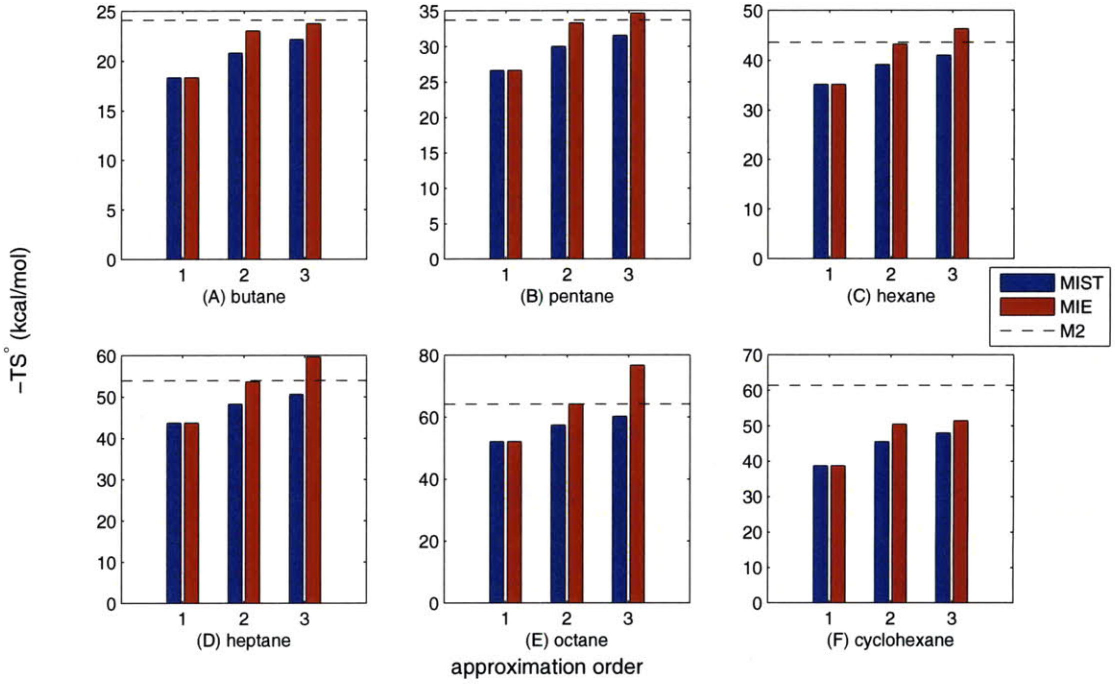


Figure 4-1: **MIST and MIE results for small alkanes**: Five linear alkanes (butane–octane) as well as cyclohexane were simulated using MD, and the resulting 5×10^6 frames were used to estimate the marginal entropies. These entropies were then combined according to MIST (blue bars) or MIE (red bars) to generate the first-, second-, or third-order approximation to the configurational entropy of each molecule. Results are compared to published calculations [45] using the Mining Minima method (dashed black line).

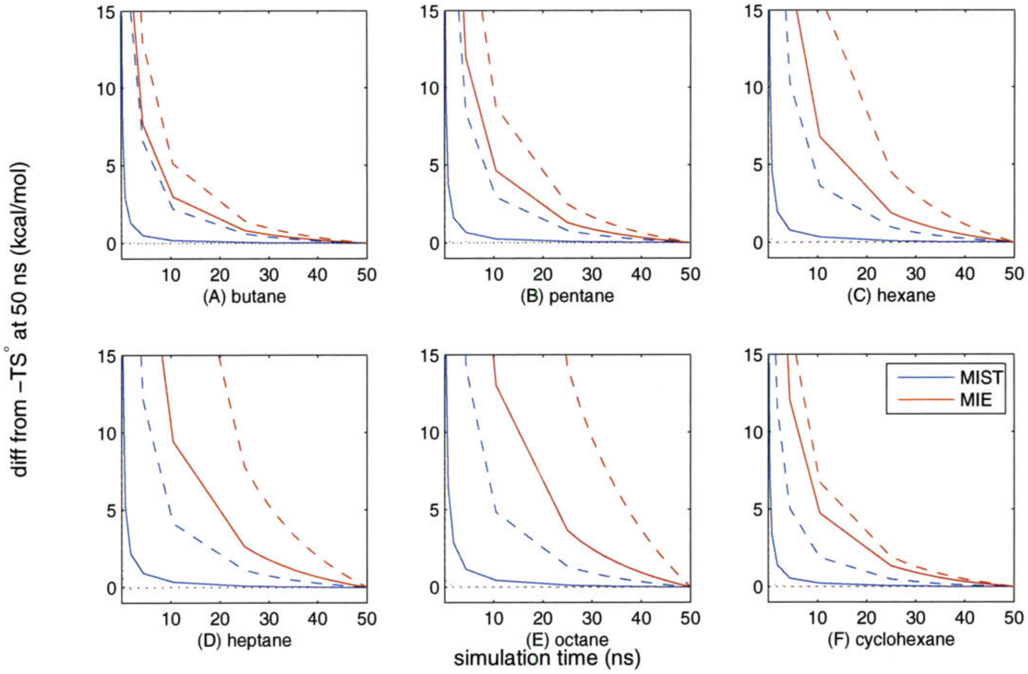


Figure 4-2: **Convergence of MIST and MIE for small alkanes**: MD simulations of various linear alkanes or cyclohexane were subsampled to include frames corresponding to shorter simulation times, and the resulting sets of frames were used to compute the MIST (blue lines), and MIE (red lines) approximations. The convergence of first- (dotted line), second- (solid lines), and third-order (dashed lines) approximations is shown. Each line shows the deviation from the same value computed using the full 50 ns trajectory.

4.4.2 Convergence for small alkanes

In addition to looking at the MIE and MIST values computed using the full 50 ns simulation, we also examined the behavior of the approximations when using only frames corresponding to shorter simulation times. Because each approximation order is converging to a different value, and the fully converged values are not known, we track the approach to the value computed with the full 50 ns. The results are shown in Figure 4-2 and Table 4.1. For all systems, MIST (blue lines) exhibits faster convergence than MIE (red lines). While the third-order approximations (dashed lines) converge more slowly than the corresponding second-order (solid lines), MIST_3 still demonstrates faster convergence than MIE_2, particularly for larger systems.

The MIE_2 convergence results are somewhat surprising given the good agreement with M2 for these systems. For example, despite the fact that MIE_2 and M2 agree within 0.02 kcal/mol for octane, the computed value for MIE_2 changed by nearly 1 kcal/mol in the last 10 ns of the simulation. Similar, though less pronounced, behavior is seen for the smaller alkanes. Given the consistent downward trend of the convergence plots, this suggests that the converged MIE_2 values are unlikely to agree as closely with M2 as the values computed at 50 ns do. While the same is technically true for the MIST approximations, the effect is likely to be much smaller given that MIST_2 and MIST_3 changed by only 0.03 and 0.34 kcal/mol, respectively in the last 10 ns of the octane simulation.

Table 4.1: Change in estimation of $-TS^\circ$ from 40 ns–50 ns (kcal/mol)

molecule	$\text{MIST}_1/\text{MIE}_1$	MIST_2	MIST_3	MIE_2	MIE_3
butane	0.00	-0.01	-0.15	-0.21	-0.37
pentane	-0.01	-0.03	-0.20	-0.34	-0.64
hexane	0.00	-0.02	-0.23	-0.48	-1.15
heptane	-0.01	-0.03	-0.29	-0.68	-2.01
octane	0.00	-0.03	-0.34	-0.93	-3.67
cyclohexane	0.00	-0.01	-0.11	-0.33	-0.48

Previous work showed that MIE_3 was poorly converged for many of the alkanes, particularly the larger ones, as is observed here [45]. Over the last 10 ns of the hexane, heptane, and octane simulations, the MIE_3 estimate changes by 1.0–3.5 kcal/mol. Notably, the third-order MIE approximation already omits a number of terms to improve numerical stability (all three-way information terms containing a bond or an angle are set to zero). In contrast, the third-order MIST implementation shown here includes all of these terms, and still demonstrates significantly faster convergence. Though we have not explored higher-order MIST approximations for these systems, the good convergence of MIST_3 suggests that fourth- or fifth-order approximations may be feasible.

Taken together with the previous section demonstrating the agreement between MIST, MIE, and M2, we can see that sampling regimes may exist in which any of the MIE or MIST approximations give the smallest error. To get a sense of how the approximations may behave in this regard, we can treat M2 as a comparison point.

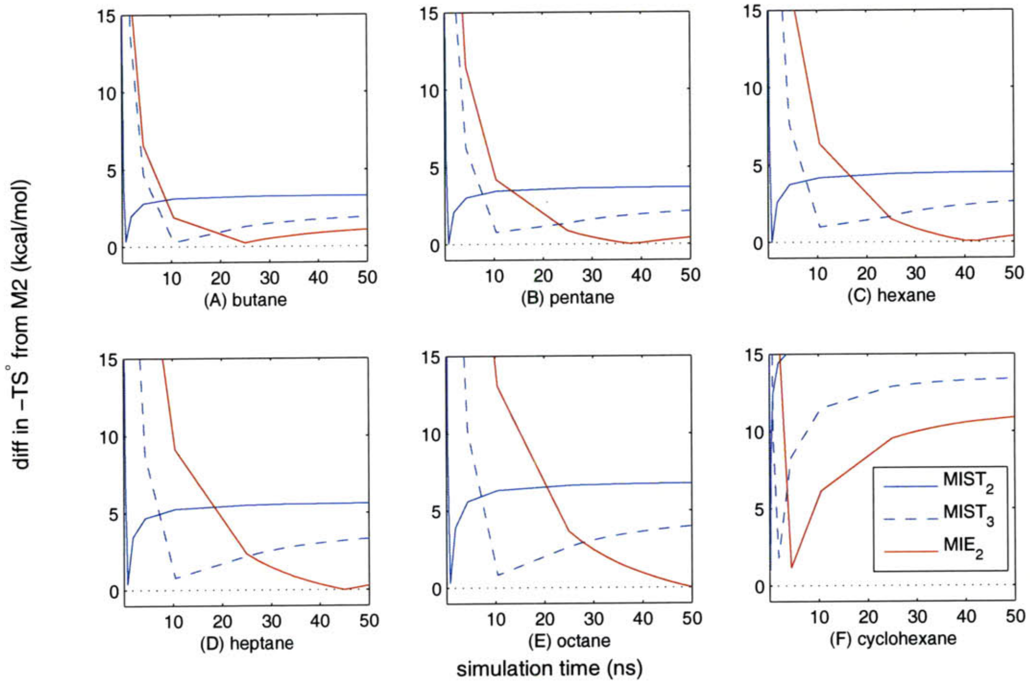


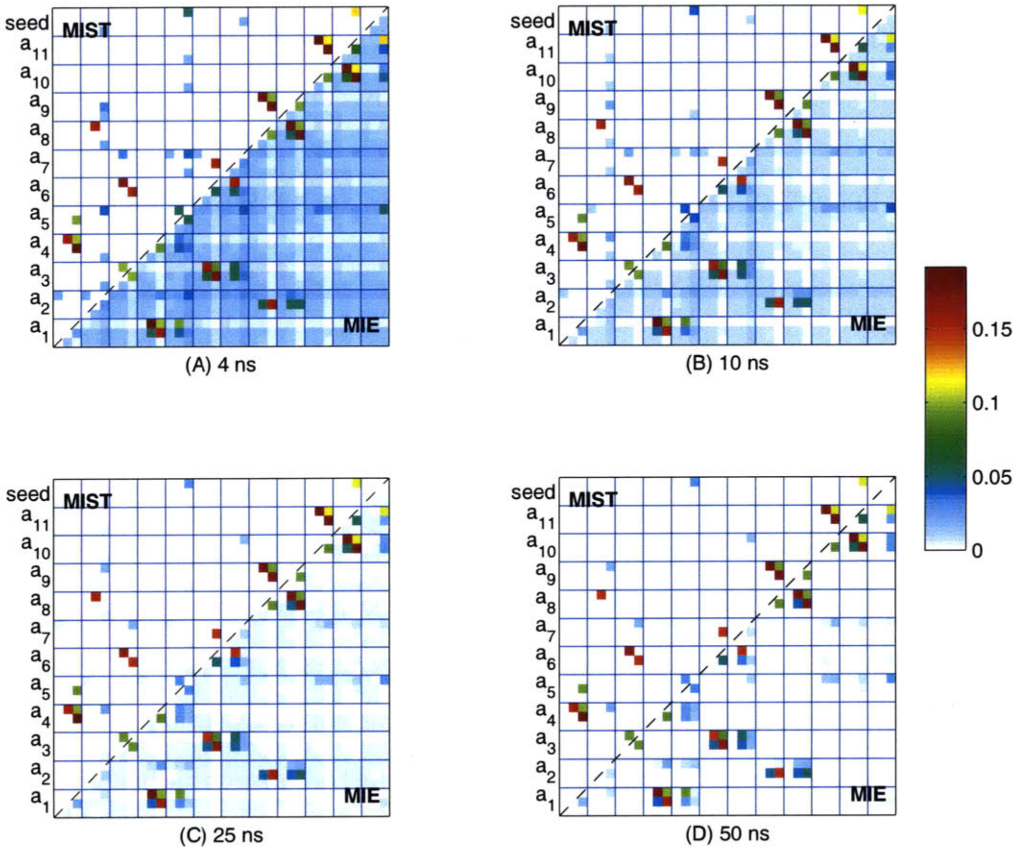
Figure 4-3: **Agreement with M2 across sampling regimes**: MIST (blue lines) and MIE (red lines) approximations were computed as a function of simulation times as described in Figure 4-2, and the absolute deviation from published M2 results were plotted, demonstrating that different approximations provide the best agreement with M2 in different sampling regimes.

Although the M2 result may not be equivalent to the full entropy to which MIE and MIST would ultimately converge, treating it as a standard may be instructive about the combined behavior of the methods when weighing accuracy and convergence. To this end, Figure 4-3 shows the error of the approximations as a function of simulation time when treating M2 as a gold standard. For all linear alkane systems, regimes exist for which MIST_2, MIST_3, or MIE_2 provide the smallest error. In particular, the rapid convergence of MIST_2 produces the best agreement with M2 for simulation times < 9 ns. Around this point, MIST_3 tends to reach good enough convergence to provide the best estimate until $\sim$ 25 ns at which point MIE_2 converges to the point that it provides the closest agreement. Across the linear alkanes, as system size increases, the transition points tend to extend to later times, suggesting that the regimes in which the MIST approximations provide improved accuracy relative to MIE may be particularly relevant for larger systems.

4.4.3 Source of differences between MIE_2 and MIST_2 for small alkanes

To understand the differences in accuracy and convergence between MIE and MIST, we next examined the terms of the expansions that differ between the two approximation frameworks. In particular, for the second-order approximations, MIST_2 includes a subset of the mutual information terms considered by MIE_2, as can be seen in Equations 4.10 and 4.6. As such, these omitted terms are entirely responsible for the differences between the two approximations. The values of the terms used for both approximations when applied to butane are shown in Figure 4-4.

For each plot, the lower triangle of the matrix shows the pairwise mutual information between each pair of degrees of freedom, all of which are included in the calculation of MIE_2. The upper triangle shows the subset of these terms that are used by MIST_2, chosen to minimize Equation 4.6 while maintaining an upper bound on the entropy. Focusing on panel D, showing the results using the full 50 ns simulation, one can see that most of the omitted terms are relatively low in value, whereas the


(A) 4 ns

(B) 10 ns

(C) 25 ns

(D) 50 ns

Figure 4-4: **Convergence of MI matrix for butane**: The pairwise mutual information terms between all pairs of degrees of freedom in butane computed using the first (A) 4 ns, (B) 10 ns, (C) 25 ns, or (D) 50 ns are show in the lower triangles. The upper triangles indicate the terms that were chosen to be included in the second order MIST approximation, according to Equation 4.6. The dark blue lines separate the atoms from each other, with each atom being represented by three degrees of freedom associated with its placement (bond, angle, torsion from bottom to top and left to right in each box). All values are reported in kcal/mol.

high MI terms are included. Panels A–C show the same information when using the first 4, 10, or 25 ns of the simulation, respectively. In contrast to the 50 ns results, the shorter simulations show dramatic differences between $MIST_2$ and MIE_2. While roughly the same set of terms are omitted by $MIST_2$ in these cases as in the 50 ns case, the omitted terms are much larger, due to their relatively slow convergence. These plots indicate that slow convergence of MIE_2 relative to $MIST_2$ is likely a result of the many terms in the MI matrix that are slowly converging to very small values.

Table 4.2: Percent of (MIE_2 − $MIST_2$) accounted for by terms of various magnitudes

molecule	$x \geq .05$	$.05 > x \geq .01$	$x < .01$
butane	29.7	30.4	39.9
pentane	28.4	30.1	41.5
hexane	24.4	26.7	49.0
heptane	19.5	26.3	54.2
octane	17.5	23.8	58.7

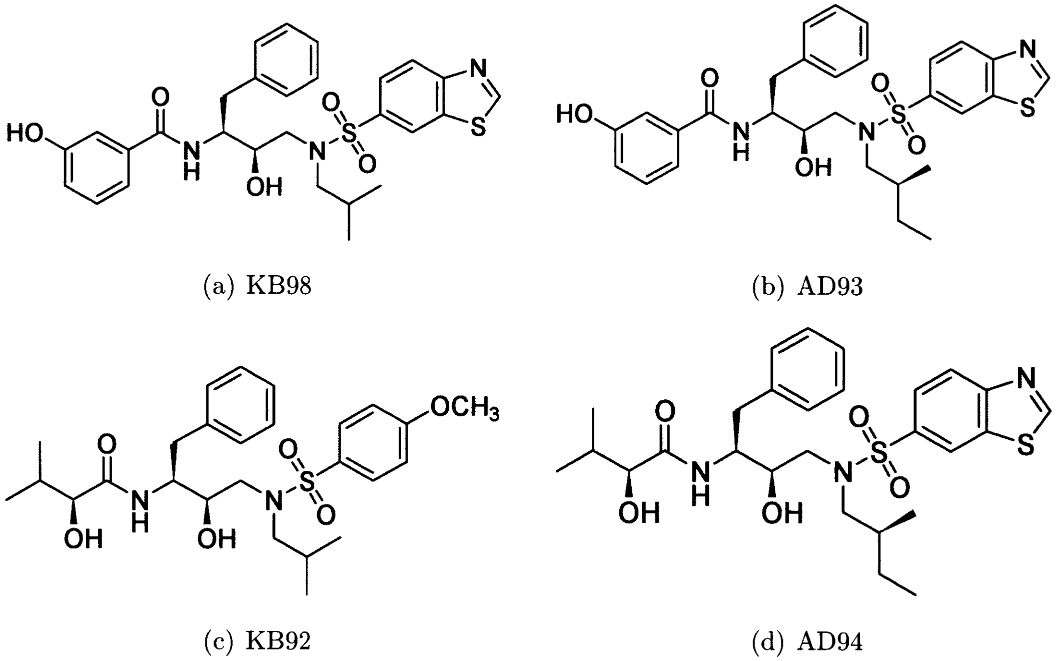
To further examine the source of differences between $MIST_2$ and MIE_2, we looked at how much of the difference between the approximations was accounted for by terms of various sizes for the linear alkanes. The results of this analysis using the full 50 ns simulations are shown in Table 4.2. As suggested by Figure 4.6, much of the difference between $MIST_2$ and MIE_2 comes from the large number of omitted small terms. For example, for butane, 39.9% of the 2.22 kcal/mol difference comes from MI terms with magnitudes less than 0.01 kcal/mol. Furthermore, the importance of these small terms grows as the system size increase, accounting for nearly 60% of the disparity for octane. Taken in conjunction with the slow convergence of these small terms, these results suggest that, while some real representational differences do exist between MIE and MIST, much of the difference may in fact be explained by differences in convergence even at 50 ns.

4.4.4 Discretized drug molecules as an analytical test case

While the good agreement that both MIST and MIE show with the M2 results is an important validation step in evaluating the overall accuracy of the approximations, some fundamental differences in the methodology can make the results somewhat difficult to evaluate. There are two primary issues that can confound the interpretation. Firstly, M2 calculations and MD simulations represent similar but ultimately different energy landscapes. Whereas the MD landscape represents the exact energy function used in the simulation, M2 approximates the landscape by linearizing the system about a set of relevant minima. Although mode-scanning is employed to account for some anharmonicities in the systems, M2 still operates on an approximation of the energy landscape sampled during MD. As such, even given infinite samples, and without making any truncation approximations (i.e., directly generating $\rho(\boldsymbol{r})$ for use in Equation 4.1), the entropy estimate would not necessarily converge to the M2 result. Secondly, because application of MIST and MIE relies upon estimating the low-order marginal entropies from a finite number of MD frames, it is difficult to separate the error introduced by the approximation framework from the error introduced by estimating the marginal terms.

To address these issues, we examined MIST and MIE in the context of a series of discrete rotameric systems in which the energy of all relevant states was calculated directly. Given this distribution of rotameric states, the full configurational entropy and all marginal entropies can then be computed exactly. As such, in these systems, we can separately evaluate the approximation errors due to the MIST or MIE frameworks, as well as sampling errors due to estimating the marginal terms. These discrete ensembles were originally generated to analyze a series of candidate HIV-1 protease inhibitors [3], but their primary importance for the current work is as a test case in which entropies of arbitrary order can be computed exactly. The chemical structures of the four drugs can be seen in Figure 4-5. Additional details on the generation of these systems can be found in Section 4.3.2.

We employed eight different discrete ensembles, representing bound and unbound states of the four molecules. All bonds, angles, and non-torsional dihedrals were idealized and fixed, leaving 13–15 torsional degrees of freedom in the systems. We also included an additional single external degree of freedom in the bound cases to model the position of the molecule with respect to the rigid binding pocket. For each



(a) KB98

(b) AD93

(c) KB92

(d) AD94

Figure 4-5: **Chemical structures of idealized discrete molecules**: The four molecules shown were previously designed as candidate HIV-1 protease inhibitors [3]. For the current work, idealized rotameric systems in which the exact energies of 50,000 rotameric states were generated in both bound and unbound states, as described and in Section 4.3.2. All torsional degrees of freedom for each drug were rotamerized, and all other DOF (bonds, angles, impropers) were fixed to idealized values.

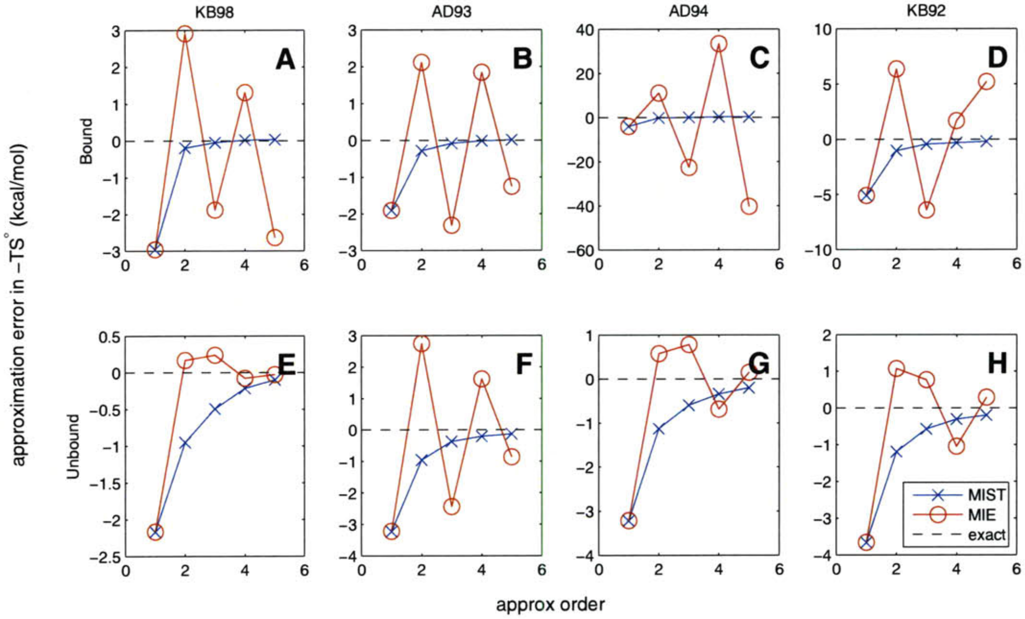


Figure 4-6: **Accuracy in rotameric systems**: For each of the four molecules, either in the unbound state (bottom row), or in the context of a rigid binding pocket (top row), we computed the exact marginal entropies for all combinations of 1–5 torsions, according to the Boltzmann distribution across the 5 × 10^4 configurations representing each system. Using these exact marginal entropies, we computed the MIST (blue lines) or MIE (red lines) approximations to the entropy of each system. The convergence as a function of approximation order is shown in comparison to the analytically determined entropy of the full system (dashed black line).

system, we computed exactly all entropy terms containing 1, 2, 3, 4, or 5 degrees of freedom by marginalizing the full Boltzmann distribution. We then computed approximations to the total entropy of each system using either MIST or MIE. As such, we were able to examine the approximation error associated with both methods when the low-order terms are known exactly. The results are shown in Figure 4-6. For all eight systems, the MIST approximations (blue lines, ×'s) monotonically approach the full entropy (dashed black line) as the approximation order increases. All MIST approximations also provide a lower bound to the entropic free energy (or an upper bound to the associated Shannon entropy) when the low order terms are known exactly. Both of these properties are guaranteed for MIST when the marginal terms are known exactly, so seeing them hold in our test system is important, if not surprising. For all cases, the second-order MIST approximation provides an estimate within 1.2 kcal/mol of the full analytic entropic free energy, with particularly good performance in the bound systems (top row of Figure).

For the four unbound systems (bottom row of Figure 4-6), MIE (red lines, o's) shows similar accuracy to MIST, generating a lower-error estimate once (KB98, panel E), a worse estimate once (AD93, panel F), and comparable error for two cases (AD94 and KB92, panels G and H). Unlike MIST, MIE is not guaranteed to monotonically reduce the approximation error as the order increases, and in some cases, such as unbound KB98 and AD94, the third-order approximation performs worse than the second-order. In general, however, for the unbound cases the MIE approximations converge towards the true entropy as the approximation order is increased, with exact low-order terms.

In contrast to its performance in the unbound systems, MIE demonstrates erratic behavior in the bound systems. For all four systems and all approximation orders, MIST results in considerably lower error than the corresponding MIE approximations. Furthermore, increasing the approximation order does not dramatically improve the performance of MIE in the bound systems, and actually results in divergent behavior for orders 1–5 in AD94 (panel C). Notably, the bound systems represent identical molecules to those in the unbound systems; the only differences lie in the level of discretization, and the external field imposed by the rigid protein in the bound state.

4.4.5 Convergence properties in discrete systems

Having investigated the error due to the MIST and MIE approximation frameworks in our analytically exact discrete systems, we next looked to explore the errors associated with computing the approximations from a finite number of samples. To do this, we performed a series of computational experiments in which we randomly drew with replacement from the 50,000 structures representing each system according to the Boltzmann distribution determined by their energies and a temperature of 300 K. For each system, we drew 10^6 samples, and estimated the PDF over the 50,000 states using subsets of the full 10^6. These PDFs were then used to compute the marginal entropies used in MIST and MIE. For each system, this procedure was repeated 50 times to evaluate the distribution of sampling errors for the two methods.

In order to quantify the sampling error separately from the approximation error (which we previously examined in Section 4.4.4), we compared the approach of each approximation to the value computed when using the exact low-order terms (i.e., we examined the convergence of each approximation to its fully converged answer, as opposed to the true joint entropy). The results for the bound and unbound KB98 systems are shown in Figure 4-7. Results for the other molecules were similar and are shown in Figures C-2, C-3, and C-4. As expected, the lower-order approximations converge more quickly, as the low-order PDFs require fewer samples to estimate accurately. For the unbound case (bottom row), both MIE (red) and MIST (blue) exhibit consistent steady convergence for all 50 runs. For the bound case (top row), while MIST exhibits similar convergence behavior as in the unbound system, MIE shows much larger variations across the 50 runs. As with the MD analysis in Section 4.4.2, MIST demonstrates considerably faster convergence than MIE for all approximation orders examined and all systems.

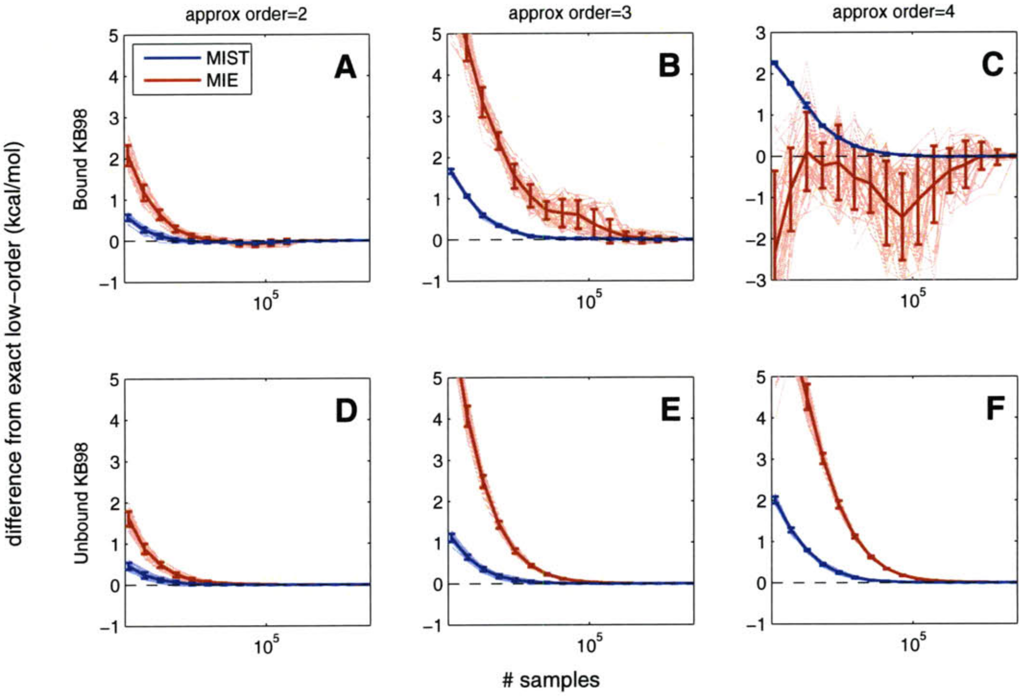


Figure 4-7: **Convergence in KB98 rotameric systems**: For each of the eight idealized rotameric systems, we sampled with replacement from the 5×10^4 configurations representing each system, according to the Boltzmann distribution determined by the relative energies of each configuration. These samples were then used to estimate the marginal entropies of all combinations of 1–4 torsions, prior to application of MIST (blue lines) or MIE (red lines). This procedure was repeated 50 times for each system, and the results of each run are shown (pale lines), as well as the mean and standard deviation across the 50 runs (thick lines). Results for bound (top row) and unbound (bottom row) KB98 are shown here. Results for other molecules were similar and can be seen in Figures C-2, C-3, and C-4

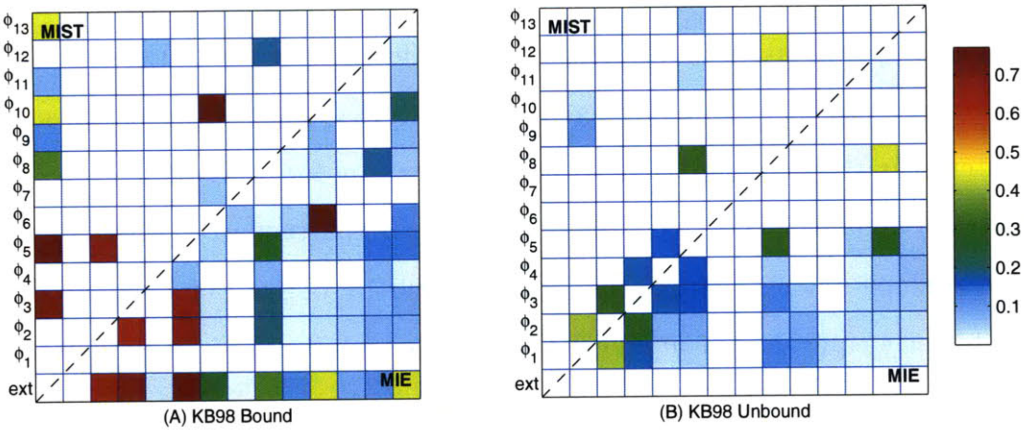


Figure 4-8: **MI matrix for discretized KB98**: The pairwise mutual information terms between all pairs of degrees of freedom in (A) bound or (B) unbound KB98 are show in the lower triangles. The upper triangles indicate the terms that were chosen to be included in the second order MIST approximation, according to Equation 4.6. All values are reported in kcal/mol.

4.4.6 Source of differences between MIE$_2$ and MIST$_2$ for discrete systems

We next examined the MI terms accounting for differences between the two approximation frameworks. As with the analysis of the alkanes (Section 4.4.3), the similarities between the second-order approximations enables a direct comparison of the MI terms that are included by MIE but omitted in MIST. Unlike the alkane studies, however, because the low-order terms can be determined directly for these discrete cases, the convergence errors, which played a important role in differences for the alkanes, can be eliminated in the current analysis. Doing so allows direct examination of the differences for the two approximation frameworks, independent of errors introduced due to sampling. The MIs between all pairs of degrees of freedom for bound and unbound KB98, as well as the terms chosen by MIST$_2$ are shown in Figure 4-8.

The results for the unbound case (panel B), for which MIE$_2$ provides lower error, are qualitatively similar to those seen for the alkanes. Most of the differences between

MIE_2 and MIST_2 in the unbound molecule arise from the omission of a number of relatively small terms, less than 0.2 kcal/mol. The larger MI terms are all included in both approximations. In contrast, the differences between the two methods for the bound case come from a different source: MIST_2 omits three of the seven largest MI terms in the bound system, together accounting for nearly 2 kcal/mol of the 2.91 kcal/mol difference between MIE_2 and MIST_2. In particular, whereas all six pairwise relationships among the external, ϕ_2, ϕ_3, and ϕ_5 degrees of freedom show strong (and nearly equivalent) couplings, MIST_2 only includes three of these terms (as it is restricted to avoid cycles in order to maintain bounding guarantees).

The qualitative differences in the terms accounting for the disparity between MIST_2 and MIE_2 in bound KB98 compared the unbound KB98 and the alkanes may be particularly relevant given the relatively poor accuracy of MIE for the bound systems. The strong couplings between the four degrees of freedom of focus (external, ϕ_2, ϕ_3, and ϕ_5), suggest a high-dimensional transition in which all four DOF are tightly coupled to each other and must change in concert to adopt different energetically relevant states. In particular, the values of the couplings, all of which are near $\ln 2$, are consistent with these four degrees of freedom together occupying two dominant states. Such high-order couplings could be responsible for the poor performance of MIE in the bound systems.

4.5 Discussion

Here we have examined the behavior of our Maximum Information Spanning Trees (MIST) approximation framework in the context of computing molecular configurational entropies. Though we originally developed MIST in order to pursue high-dimensional information theoretic phrasings in the analysis of experimental biological data, the generality of the method, coupled with the mathematical relationships between information theory and statistical mechanics, enabled application to this system with relatively little modification. The adaptation of the method was largely inspired by the similar approach taken previously with the Mutual Information Expansion (MIE) method [45]. As such we have compared against both MIE and the well established Mining Minima (M2) method in the context of MD simulations of linear alkanes. Although the MIST approximations did not demonstrate as close agreement with M2 as that seen with the second-order MIE method, we did observe agreement within 2–4 kcal/mol, as well as significantly improved convergence, even for higher-order MIST approximations, which may prove valuable when investigating larger systems. Even in the context of the relatively small linear alkanes investigated here, we identified sampling regimes in which the MIST approximations generated better agreement with M2 compared to MIE. The size of these regimes (roughly the first 25 ns of simulation time) suggests that MIST may be particularly useful for larger systems in which simulation time may be limiting.

While the agreement with M2 is an important validation for the overall accuracy of the methods, it does not provide an ideal testing framework, as M2 and the MD simulations represent different energy landscapes. As such, separate examination of the errors due to approximation and sampling was not possible. To address this, we also examined MIST and MIE in the context of a series of idealized rotameric systems in which the exact entropies could be computed directly. In these systems, we observed that while MIE and MIST both showed good behavior in systems representing unbound molecules, MIE demonstrated poor accuracy in the more restricted bound systems, even for the fifth-order approximation with exactly determined marginal terms. In contrast, MIST exhibited small approximation errors in the bound systems, even for the second-order approximation. Furthermore, when sampling from the known analytical distribution, the fast convergence of MIST relative to MIE seen in the MD systems was also observed for these discretized molecular systems.

In addition to improved convergence, MIST carries useful properties that are not shared by MIE. For fully converged systems, the approximation error of MIST is guaranteed to monotonically decrease with increasing approximation order. This behavior can be easily seen for the discrete systems in Figure 4-6, and stands in contrast to the behavior of MIE in the same systems. In application to novel systems where the behavior of the approximations is untested, this property means that the highest approximation order to have reached convergence provides the best estimate of the full entropy. In the absence of such a guarantee, it is unclear how to select the appropriate approximation order.

Furthermore, all converged MIST approximations provide a lower bound on the entropic contribution to the free energy, $-TS^{\circ}$ (or an upper bound on the Shannon information entropy, S). The bounding behavior may prove particularly useful in identifying optimal coordinate representations. In the previous MIE work, the choice of coordinate system has been demonstrated to significantly impact the quality of the approximation [45]. In particular, removing high-order couplings between coordinates, such as those present in Cartesian coordinates, can dramatically improve the accuracy of low-order approximations like MIST and MIE. Because MIST applied to any valid coordinate set will still provide a lower bound on $-TS^{\circ}$, a variety of coordinate sets may be tested, and the one that yields the largest converged answer is guaranteed to be the most accurate. While additional work is needed to fully enable such a method, even brute-force enumeration is likely to improve performance.

The results of MIE and MIST in the context of the discrete systems also highlights the ability of MIST to provide a good approximation at low orders, even when direct high-order couplings are known to exist. As has been described previously [45, 62], low-order MIE approximations truncate terms in Equation 4.8 representing only direct high-order relationships. The poor accuracy of low-order MIE metrics for the bound idealized systems therefore implies that these systems contain significant high-order terms. Despite the presence of such complex couplings, MIST still provides a good approximation in these same systems. For systems such as proteins that are known to exhibit high-dimensional couplings, the ability to capture high-order relationships in the context of a low-order approximation may prove crucial.

Since the original development of the MIE framework, additional work has been done to extend and apply the method. Nearest-neighbor (NN) entropy estimation has been used to compute the low-order marginal terms utilized by the MIE framework, resulting in significantly improved convergence [38]. Given that MIST relies upon the same low-order marginal terms as MIE, it is likely that NN methods would also be useful in the context of MIST. MIE has also been used to analyze residue side-chain configurational freedom from protein simulations [63]. These studies were able to identify biologically relevant couplings between distal residues in allosteric proteins. Given the relative computational costs of simulating large proteins, and the strong high-dimensional couplings that surely exist in the context of proteins, application of MIST in similar studies may be particularly useful. Preliminary results from ongoing studies have proved promising in the calculation of residue side-chain configurational entropies in the active site of HIV-1 protease.

In summary, we have adapted our existing information theoretic-based approximation framework to enable calculation of configurational entropies from molecular simulation data. Having characterized its behavior in a variety of molecular systems, we believe MIST can serve as a complement to existing methods, particularly in poorly sampled regimes. A variety of existing extensions and applications for MIE are also likely to be useful in the context of MIST, though further exploration is needed. Finally, in addition to improved convergence, MIST carries monotonicity and bounding guarantees that may prove valuable for future applications.

Appendix A

Supporting materials for Chapter 2

Table A.1: Microarray datasets for cancer classification

Tissue	# Samples	# Genes	Class Type	Ref
breast	295	70	good/bad prog	[86]
leukemia	72	7070	AML/ALL	[35]
colon	62	2000	normal/tumor	[2]
prostate	102	12600	normal/tumor	[81]

Table A.2: Genes selected by MIST$_2$

Tissue	#	Gene ID	Repro %	Cancer Rel	Other Studies
breast	1	NM_003981	91.0*	[78]	[54, 12, 83]
	2	AI918032	91.0*		[12]
	3	NM_003239	85.5*	[25]	[12]
	4	AW024884	52.0*		
	5	AA404325	68.5*		
	6	AF055033	77.0*		[12, 83]
	7	AW014921	77.0*		
	8	AL080059	49.5*		[91]
	9	AI738508	1.5		
	10	AK000745	17.0		
leukemia	1	M27891	33.0*		[28, 4, 13, 91, 6, 23]
	2	U29175	3.5*		[4, 23]
	3	U72621	19.0*	[1]	[4]
	4	U88047	7.5*		[23]
	5	M92287	24.0*	[80]	[4, 6, 23]
	6	M19507	2.0		[4, 13, 6, 23]
	7	D84294	0.5		
	8	HG3549-HT3751	6.5*		
	9	M32304	6.5*		[4]
	10	AF005043	1.0		
colon	1	M63391	22.0*	[22]	[4, 6, 23]
	2	U30825	3.5		[4, 23]
	3	T57468	4.5*		[23]
	4	T47377	21.5*		[4, 6, 23]
	5	M26383	19.0*		[4, 6, 23]
	6	R39209	24.5*		[23]
	7	M76378	5.5*		[4, 6, 23]
	8	M80815	3.0		[4, 23]
	9	Y00097	4.5*	[79]	
	10	X90858	1.0	[40]	[4]
prostate	1	X07732	90.0*	[42]	[13, 89, 84]
	2	U24577	33.0*		
	3	M62895	6.0*	[75]	
	4	U12472	14.0*		
	5	D80010	17.5*		
	6	AB014545	15.0*		
	7	AB023204	27.0*		
	8	U67615	23.5*		
	9	M21536	12.5*	[72]	
	10	AF038451	4.0*	[85]	

*Bonferroni adjusted pval$\leq$0.01 for gene occuring this often in 200 random runs.

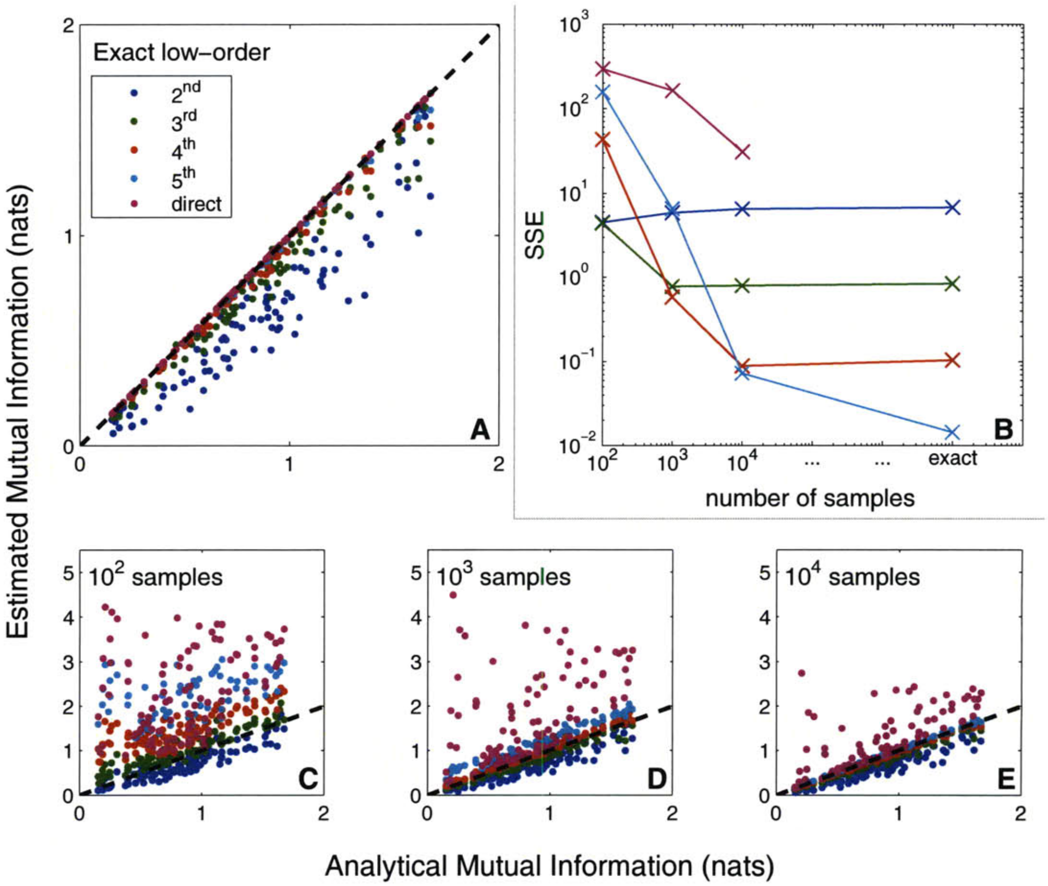


Figure A-1: **Direct validation of MIST MI approximation.** To evaluate the MIST framework, we simulated 100 randomly generated networks with analytically computable joint entropies and applied the metrics using a range of sample sizes. Half of each network was randomly chosen and the MI between one half and the other was computed analytically or using the MIST approximation of various orders. When the analytical entropies are known exactly (A), the higher-order approximations performing increasingly well. When the entropies are estimated from a finite sample, however (C–E), the approximations provide the best estimates, with the higher-order approximations performing better as more data become available. This behavior is quantified by computing the sum-of-squared error of each metric as a function of the sampling regime (B). The best approximation to use depends upon the amount of data available, but for all cases examined with finite sample size, the approximations outperform direct estimation and the second-order approximation provides a good estimate.

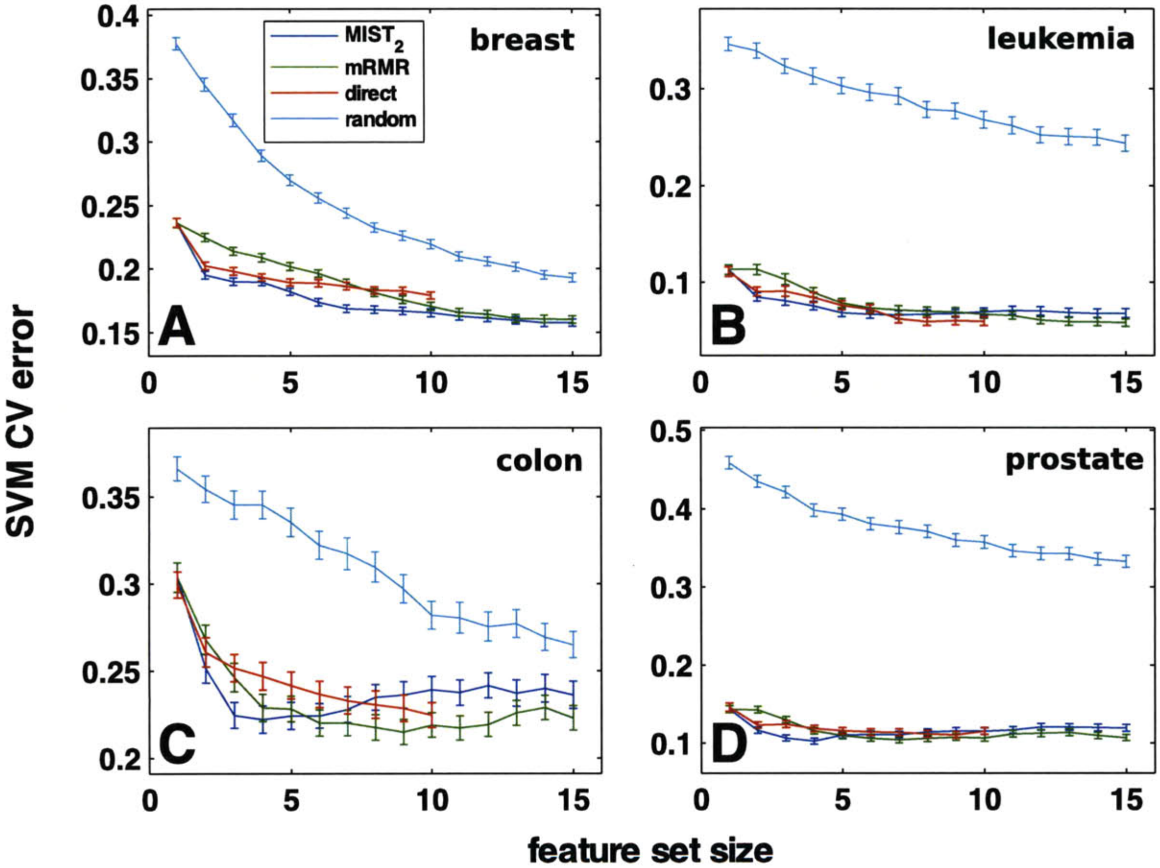


Figure A-2: **Gene subset selection for cancer classification.** Subsets of gene expression levels were chosen incrementally to maximize the information content with the cancer class variable according to MIST$_2$, direct estimation of MI, mRMR, or at random and the chosen sets were scored by the cross-validation error of an SVM classifier trained to discriminate the cancer type. For all data sets, 75% of the data was separated and used to select features and train the model; the classifier was then used to classify the remaining 25% of the samples. The mean classification error and standard error of the mean for 200 such training/testing partitioning are reported. Genes were selected for data sets relating to (A) breast, (B) leukemia, (C) colon, and (D) prostate cancer.

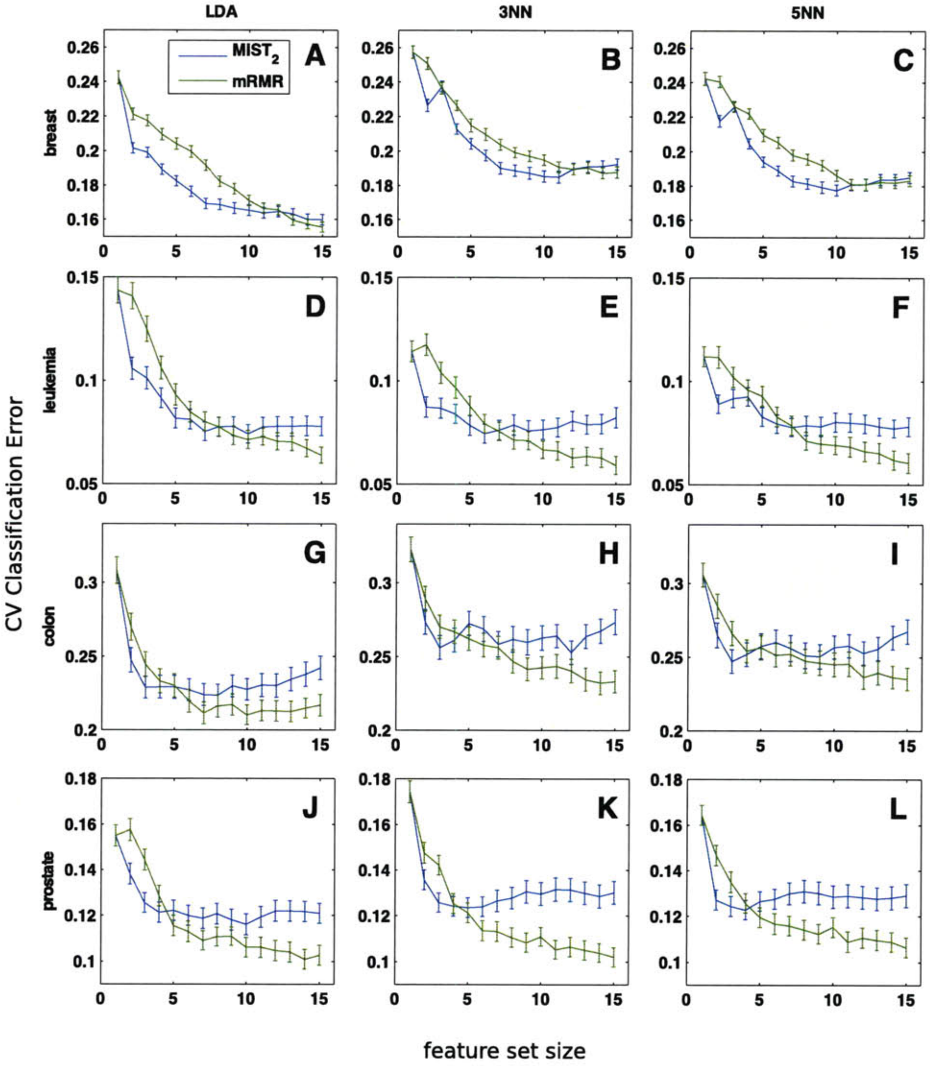


Figure A-3: **Gene subset selection for cancer classification.** Subsets of gene expression levels were chosen incrementally to maximize the information content with the cancer class variable according to MIST$_2$ or mRMR and the chosen sets were scored by the cross-validation error of an LDA (A,D,G,J), 3NN (B,E,H,K), or 5NN (C,F,I,L) classifier trained to discriminate the cancer type. For all datasets, 75% of the data was separated and used to select features and train the model; the classifier was then used to classify the remaining 25% of the samples. The mean classification error and standard error of the mean for 200 such training/testing partitioning are reported. Genes were selected for four datasets relating to (A,B,C) breast, (D,E,F) leukemia, (G,H,I) colon, and (J,K,L) prostate cancer. Results using an SVM classifier and including direct estimation-based feature selection are shown in Figure 4.

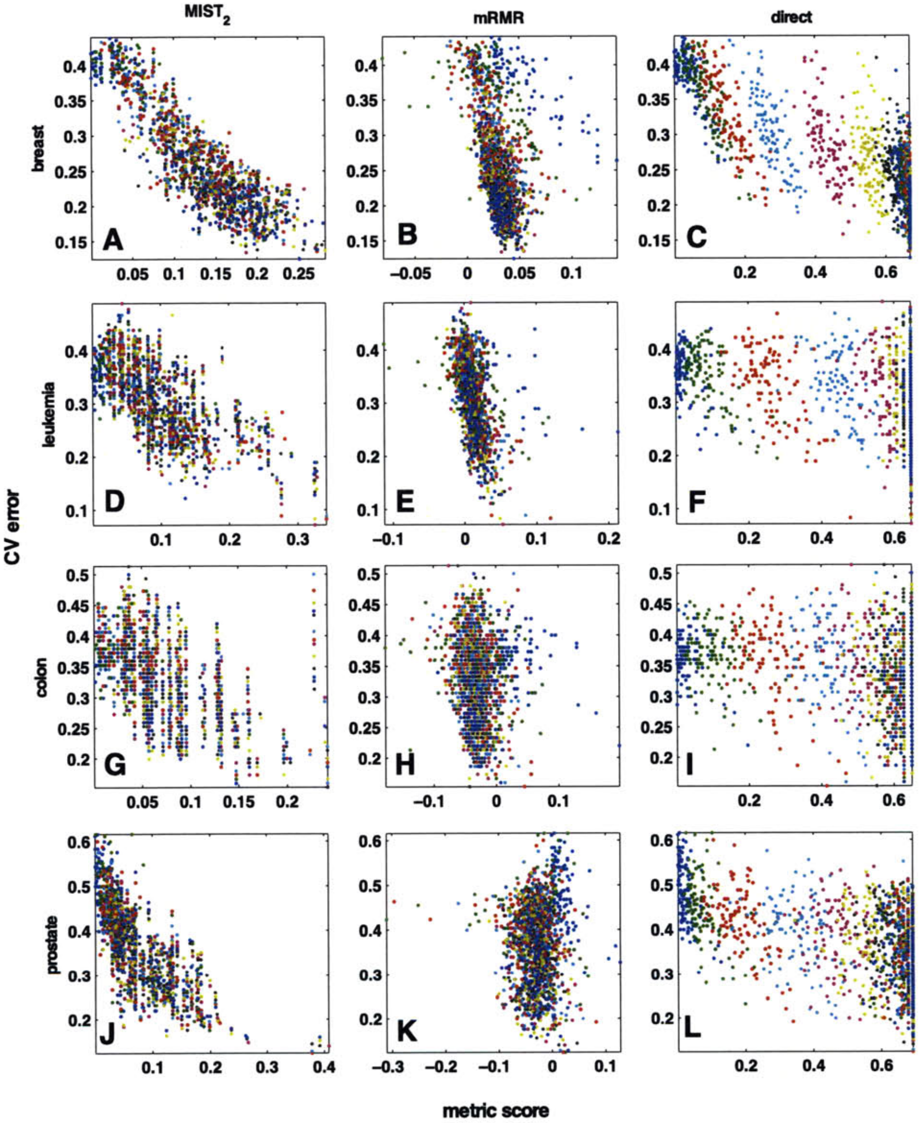


Figure A-4: **Correlation of classification error and MI metrics.** The classification error of randomly chosen subsets of 1–15 genes was computed through cross-validation with an SVM based classifier. The same sets were then scored by MIST_2 (A,D,G,J), MI computed with direct estimation (B,E,H,K), and mRMR (C,F,I,L) and these metrics are shown plotted against the CV classification error. The color of the points relates to the size of the feature set, cycling through blue, green, red, cyan, magenta, yellow, black for increasing set size. The correlation coefficients between metrics as a function of set size is shown in Figure 3. Notably, MIST_2 has strong negative correlation across all feature set sizes.

Appendix B

Supporting materials for Chapter 3

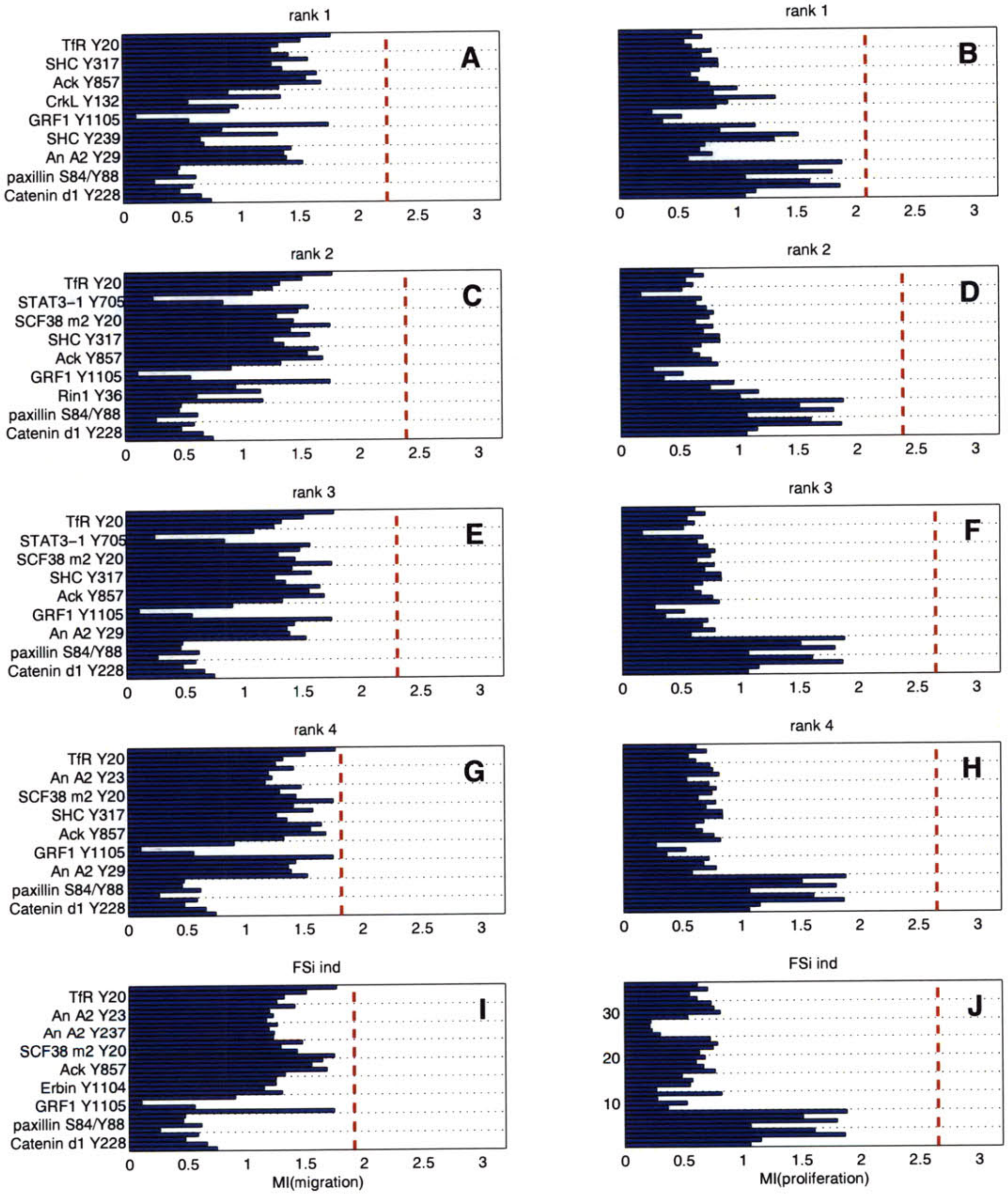


Figure B-1: **MI of feature sets with outputs**: The MIs for migration (left column) or proliferation (right column) from each of the four measures for each of the nine phospho-peptides in each feature set is shown. Additional details, as well as results for network gauge and MIST opt can be seen in Figure 3-5.

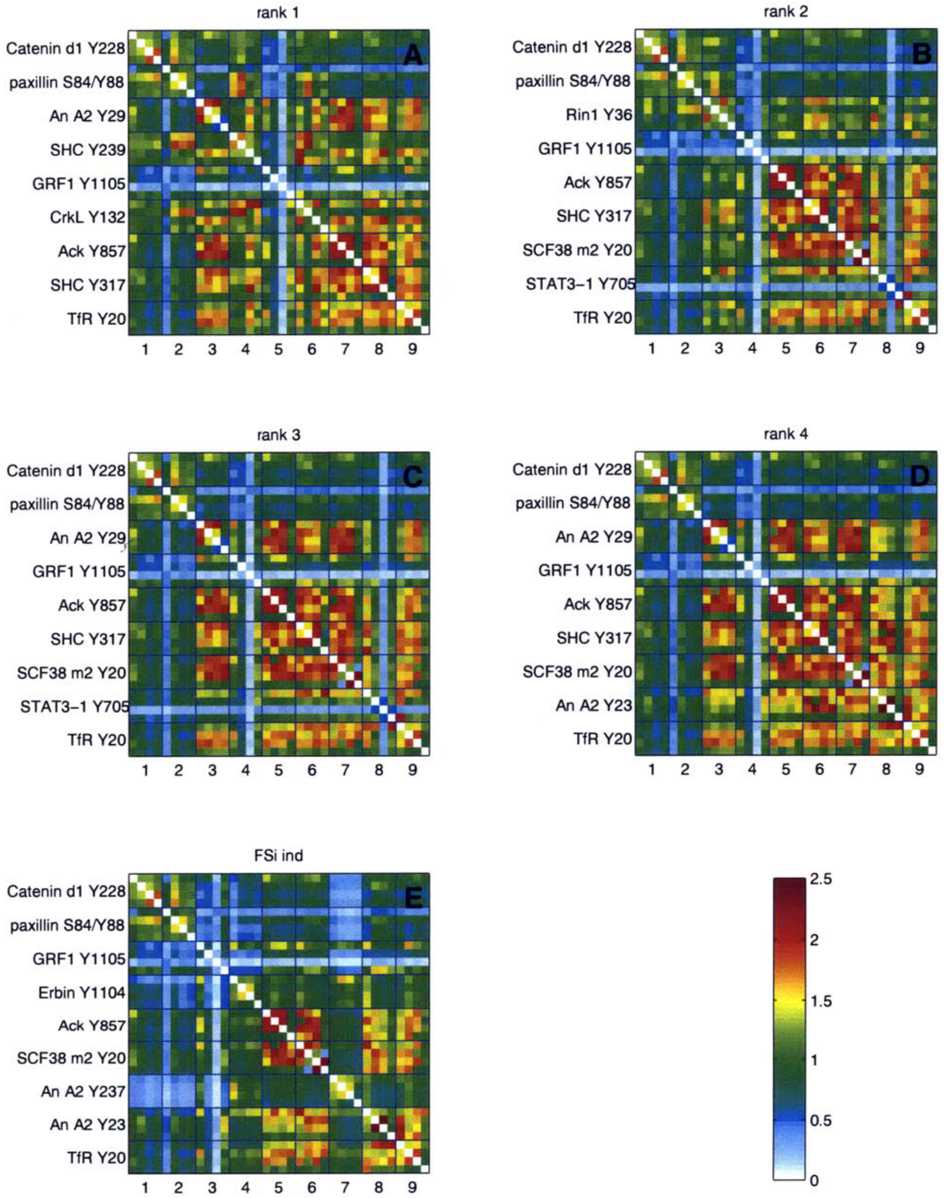


Figure B-2: **MI matrices of feature sets**: The MI between each pair of signals in the indicated feature sets are shown. Results for the network gauge and MIST opt sets can be found in Figure 3-6.

Appendix C

Supporting materials for Chapter 4

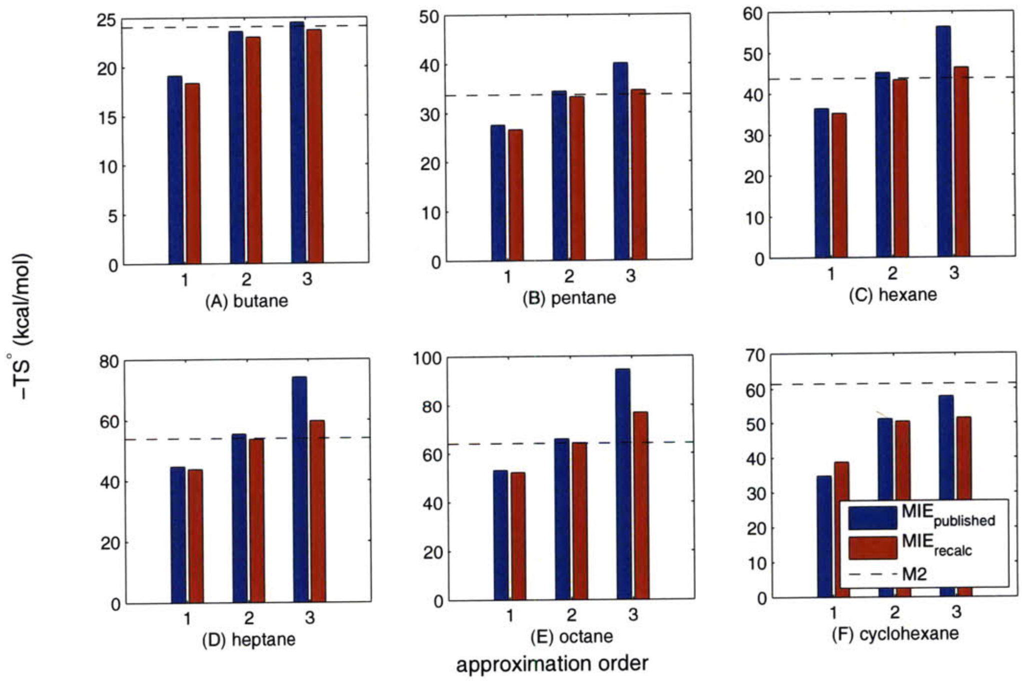


Figure C-1: **Regeneration of published MIE results**: MD simulations of the indicated alkanes were run and analyzed as described in Methods section 4.3.1 and summarized in the caption of Figure 4-1. Our recomputed MIE results are compared against those published previously CITE. Both the first- and second-order recalculated values agree well with published results. Deviations in the third-order are likely a result of the poor convergence for both our numbers and those reported.

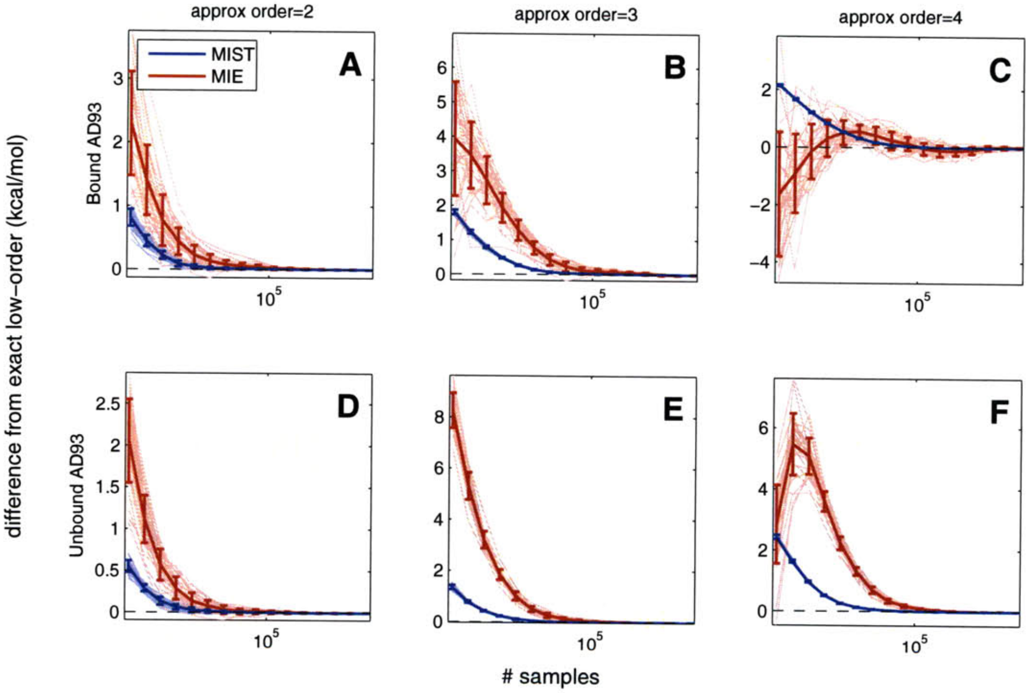


Figure C-2: **Convergence in AD93 rotameric systems**: The convergence of MIST and MIE in idealized rotameric systems was computed as described in Methods section 4.3.2 and summarized in the caption of Figure 4-7. Results for bound and unbound AD93 are shown.

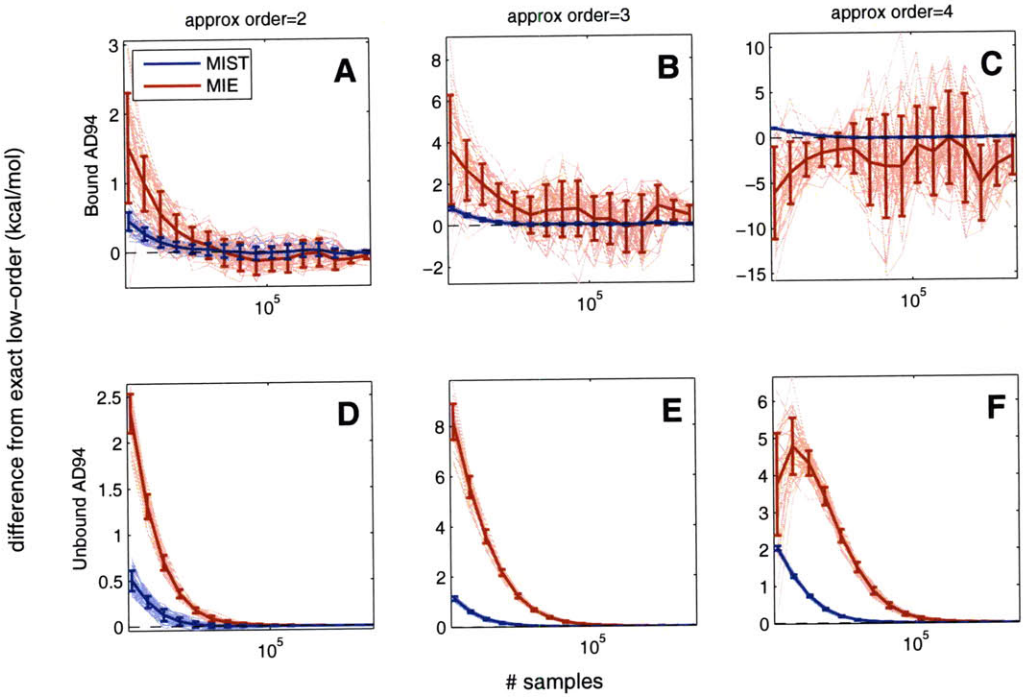


Figure C-3: **Convergence in AD94 rotameric systems**: The convergence of MIST and MIE in idealized rotameric systems was computed as described in Methods section 4.3.2 and summarized in the caption of Figure 4-7. Results for bound and unbound AD94 are shown.

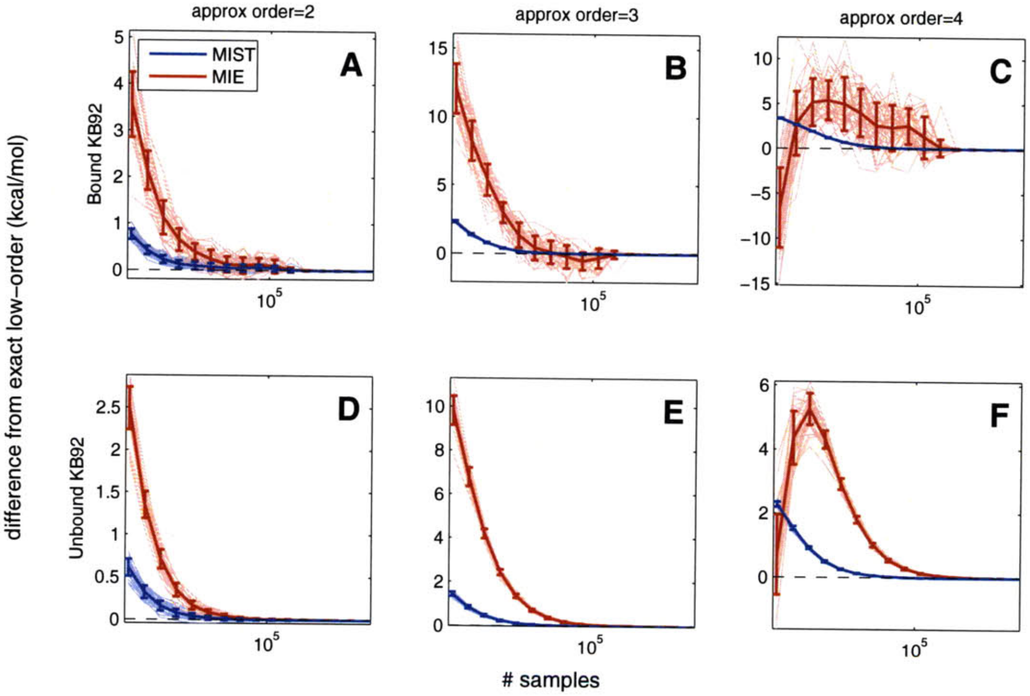


Figure C-4: **Convergence in KB92 rotameric systems**: The convergence of MIST and MIE in idealized rotameric systems was computed as described in Methods section 4.3.2 and summarized in the caption of Figure 4-7. Results for bound and unbound KB92 are shown.

Bibliography

[1] A. Abdollahi, A. K. Godwin, P. D. Miller, L. A. Getts, D. C. Schultz, T. Taguchi, J. R. Testa, and T. C. Hamilton. Identification of a gene containing zinc-finger motifs based on lost expression in malignantly transformed rat ovarian surface epithelial cells. *Cancer Res*, 57(10):2029–2034, 1997.

[2] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proc Natl Acad Sci USA*, 96(12):6745–6750, 1999.

[3] M. D. Altman, A. Ali, G. S. K. K. Reddy, M. N. L. Nalam, S. G. Anjum, H. Cao, S. Chellappan, V. Kairys, M. X. Fernandes, M. K. Gilson, C. A. Schiffer, T. M. Rana, and B. Tidor. HIV-1 protease inhibitors from inverse design in the substrate envelope exhibit subnanomolar binding to drug-resistant variants. *J Am Chem Soc*, 130(19):6099–6113, 2008.

[4] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *J Comput Biol*, 7(3-4):559–583, 2000.

[5] H. A. Bethe. Statistical theory of superlattices. *Proc. R. Soc. Lond. A*, 150:552–575, 1935.

[6] T. Bo and I. Jonassen. New feature subset selection procedures for classification of expression profiles. *Genome Biol*, 3(4):RESEARCH0017, 2002.

[7] B. R. Brooks, R. E. Bruccoleri, B. D. Olafson, D. J. States, S. Swaminathan, and M. Karplus. CHARMM: a program for macromolecular energy, minimization, and dynamics calculations. *Journal of Computational Chemistry*, 4(2):187–217, 1983.

[8] L. A. Cary, D. C. Han, and J. L. Guan. Integrin-mediated signal transduction pathways. *Histol Histopathol*, 14(3):1001–1009, 1999.

[9] C. Chang, W. Chen, and M. K. Gilson. Evaluating the accuracy of the quasiharmonic approximation. *Journal of Chemical Theory and Computation*, 1(5):1017–1028, 2005.

[10] C. Chang and M. K. Gilson. Free energy, entropy, and induced fit in host-guest recognition: Calculations with the second-generation mining minima algorithm. *Journal of the American Chemical Society*, 126(40):13156–13164, Oct. 2004.

[11] H. Chin, T. Saito, A. Arai, K. Yamamoto, R. Kamiyama, N. Miyasaka, and O. Miura. Erythropoietin and IL-3 induce tyrosine phosphorylation of CrkL and its association with shc, SHP-2, and cbl in hematopoietic cells. *Biochemical and Biophysical Research Communications*, 239(2):412–417, Oct. 1997.

[12] A. Choudhary, M. Brun, J. Hua, J. Lowey, E. Suh, and E. R. Dougherty. Genetic test bed for feature selection. *Bioinformatics*, 22(7):837–842, 2006.

[13] W. Chu, Z. Ghahramani, F. Falciani, and D. L. Wild. Biomarker discovery in microarray gene expression data with Gaussian processes. *Bioinformatics*, 21(16):3385–3393, 2005.

[14] M. Coon and R. Herrera. Modulation of HeLa cells spreading by the non-receptor tyrosine kinase ACK-2. *Journal of Cellular Biochemistry*, 84(4):655–665, 2002. PMID: 11835391.

[15] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms.* MIT Press, Cambridge, Mass., 2nd edition, 2001.

[16] B. D. Cosgrove, B. M. King, M. A. Hasan, L. G. Alexopoulos, P. A. Farazi, B. S. Hendriks, L. G. Griffith, P. K. Sorger, B. Tidor, J. J. Xu, and D. A. Lauffenburger. Synergistic drug-cytokine induction of hepatocellular death as an in vitro approach for the study of inflammation-associated idiosyncratic drug hepatotoxicity. *Toxicol Appl Pharmacol*, 237(3):317–330, 2009.

[17] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* Wiley-Interscience, Hoboken, N.J., 2nd edition, 2006.

[18] K. A. S. D. Sitkoff and B. Honig. Accurate calculation of hydration free-energies using macroscopic solvent models. *J Phys Chem*, 98:1978–1988, 1994.

[19] B. I. Dahiyat and S. L. Mayo. Protein design automation. *Protein Sci*, 5(5):895–903, 1996.

[20] B. I. Dahiyat and S. L. Mayo. De novo protein design: fully automated sequence selection. *Science*, 278(5335):82–87, 1997.

[21] J. Desmet, M. D. Maeyer, B. Hazes, and I. Lasters. The dead-end elimination theorem and its use in protein side-chain positioning. *Nature*, 356(6369):539–542, 1992.

[22] P. Dias, P. Kumar, H. B. Marsden, P. H. Morris-Jones, J. Birch, R. Swindell, and S. Kumar. Evaluation of desmin as a diagnostic and prognostic marker of childhood rhabdomyosarcomas and embryonal sarcomas. *Br J Cancer*, 56(3):361–365, 1987.

[23] C. Ding and H. Peng. Minimum redundancy feature selection from microarray gene expression data. *J Bioinform Comput Biol*, 3(2):185–205, 2005.

[24] T. Dittmar, A. Husemann, Y. Schewe, J. Nofer, B. Niggemann, K. S. Zanker, and B. H. Brandt. Induction of cancer cell migration by epidermal growth factor is initiated by specific phosphorylation of tyrosine 1248 of c-erbB-2 receptor via epidermal growth factor receptor. *FASEB J.*, pages 02–0096fje, Sept. 2002.

[25] T.-V. Do, L. A. Kubba, H. Du, C. D. Sturgis, and T. K. Woodruff. Transforming growth factor-beta1, transforming growth factor-beta2, and transforming growth factor-beta3 enhance ovarian cancer metastatic potential by inducing a smad3-dependent epithelial-to-mesenchymal transition. *Mol Cancer Res*, 6(5):695–705, 2008.

[26] M. Draminski, A. Rada-Iglesias, S. Enroth, C. Wadelius, J. Koronacki, and J. Komorowski. Monte Carlo feature selection for supervised classification. *Bioinformatics*, 24(1):110–117, 2008.

[27] S. Dunn, L. Wahl, and G. Gloor. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. *Bioinformatics*, 24(3):333–340, Feb. 2008.

[28] J. Fand and J. Grzymala-Busse. *Leukemia Prediction from Gene Expression Data—A Rough Set Approach*, volume 4029 of *Lecture Notes in Computer Science*, pages 1611–3349. Springer, Berlin, 2006.

[29] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, J. A. Montgomery, Jr., T. Vreven, K. N. Kudin, J. C. Burant, J. M. Millam, S. S. Iyengar, J. Tomasi, V. Barone, B. Mennucci, M. Cossi, G. Scalmani, N. Rega, G. A. Petersson, H. Nakatsuji, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, M. Klene, X. Li, J. E. Knox, H. P. Hratchian, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, P. Y. Ayala, K. Morokuma, G. A. Voth, P. Salvador, J. J. Dannenberg, V. G. Zakrzewski, S. Dapprich, A. D. Daniels, M. C. Strain, O. Farkas, D. K. Malick, A. D. Rabuck, K. Raghavachari, J. B. Foresman, J. V. Ortiz, Q. Cui, A. G. Baboul, S. Clifford, J. Cioslowski, B. B. Stefanov, G. Liu, A. Liashenko, P. Piskorz, I. Komaromi, R. L. Martin, D. J. Fox, T. Keith, M. A. Al-Laham, C. Y. Peng, A. Nanayakkara, M. Challacombe, P. M. W. Gill, B. Johnson, W. Chen, M. W. Wong, C. Gonzalez, and J. A. Pople. Gaussian 03, Gaussian, Inc., Wallingford, CT, 2004.

[30] M. L. Galisteo, Y. Yang, J. U. na, and J. Schlessinger. Activation of the nonreceptor protein tyrosine kinase Ack by multiple extracellular stimuli. *Proceedings of the National Academy of Sciences*, 103(26):9796–9801, June 2006.

[31] M. K. Gilson and B. Honig. Calculation of the total electrostatic energy of a macromolecular system: solvation energies, binding energies, and conformational analysis. *Proteins*, 4(1):7–18, 1988.

[32] G. B. Gloor, L. C. Martin, L. M. Wahl, and S. D. Dunn. Mutual information in protein multiple sequence alignments reveals two classes of coevolving positions. *Biochemistry*, 44(19):7156–7165, May 2005.

[33] L. Goh and N. Kasabov. An integrated feature selection and classification method to select minimum number of variables on the case study of gene expression data. *J Bioinform Comput Biol*, 3(5):1107–1136, 2005.

[34] I. Gokcen and J. Peng. *Advances in Information Systems*, volume 2457 of *Lecture Notes in Computer Science*, pages 104–113. Springer, Berlin, 2002.

[35] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[36] V. Hnizdo, E. Darian, A. Fedorowicz, E. Demchuk, S. Li, and H. Singh. Nearest-neighbor nonparametric method for estimating the configurational entropy of complex molecules. *Journal of Computational Chemistry*, 28(3):655–668, 2007.

[37] V. Hnizdo, A. Fedorowicz, H. Singh, and E. Demchuk. Statistical thermodynamics of internal rotation in a hindering potential of mean force obtained from computer simulations. *J Comput Chem*, 24(10):1172–1183, 2003.

[38] V. Hnizdo, J. Tan, B. J. Killian, and M. K. Gilson. Efficient calculation of configurational entropy from molecular simulations by combining the Mutual-Information expansion and Nearest-Neighbor methods. *Journal of computational chemistry*, 29(10):1605–1614, July 2008. PMC2620139.

[39] K. A. Janes, J. G. Albeck, S. Gaudet, P. K. Sorger, D. A. Lauffenburger, and M. B. Yaffe. A systems model of signaling identifies a molecular basis set for cytokine-induced apoptosis. *Science*, 310(5754):1646–1653, 2005.

[40] A. Kanzaki, Y. Takebayashi, H. Bando, J. F. Eliason, S.-i. Watanabe Si, H. Miyashita, M. Fukumoto, M. Toi, and T. Uchida. Expression of uridine and thymidine phosphorylase genes in human breast carcinoma. *Int J Cancer*, 97(5):631–635, 2002.

[41] M. Karplus and J. N. Kushick. Method for estimating the configurational entropy of macromolecules. *Macromolecules*, 14(2):325–332, 1981.

[42] K. A. Kelly, S. R. Setlur, R. Ross, R. Anbazhagan, P. Waterman, M. A. Rubin, and R. Weissleder. Detection of early prostate cancer using a hepsin-targeted imaging agent. *Cancer Res*, 68(7):2286–2291, 2008.

[43] P. A. Kelly and Z. Rahmani. DYRK1A enhances the mitogen-activated protein kinase cascade in PC12 cells by forming a complex with Ras, B-Raf, and MEK1. *Mol. Biol. Cell*, 16(8):3562–3573, Aug. 2005.

[44] M. L. Kemp, L. Wille, C. L. Lewis, L. B. Nicholson, and D. A. Lauffenburger. Quantitative network signal combinations downstream of TCR activation can predict IL-2 production response. *J Immunol*, 178(8):4984–4992, 2007.

[45] B. J. Killian, J. Y. Kravitz, and M. K. Gilson. Extraction of configurational entropy from molecular simulations via an expansion approximation. *The Journal of Chemical Physics*, 127(2):024107–16, July 2007.

[46] B. M. King and B. Tidor. MIST: maximum information spanning trees for dimension reduction of biological data sets. *Bioinformatics*, 25(9):1165–1172, May 2009.

[47] N. Kumar, A. Wolf-Yadlin, F. M. White, and D. A. Lauffenburger. Modeling HER2 effects on cell behavior from mass spectrometry phosphotyrosine data. *PLoS Comput Biol*, 3(1):e4, 2007.

[48] N. Kwak and C.-H. Choi. Input feature selection by mutual information based on Parzen window. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:1667–1671, 2002.

[49] M. R. Landon and S. E. Schaus. JEDA: Joint entropy diversity analysis. An information-theoretic method for choosing diverse and representative subsets from combinatorial libraries. *Mol Divers*, 10(3):333–339, 2006.

[50] M. T. Laub and M. Goulian. Specificity in Two-Component signal transduction pathways. *Annual Review of Genetics*, 41:121–145, Dec. 2007.

[51] A. R. Leach and A. P. Lemon. Exploring the conformational space of protein side chains using dead-end elimination and the A* algorithm. *Proteins*, 33(2):227–239, 1998.

[52] A. L. Lee, S. A. Kinnear, and A. J. Wand. Redistribution and loss of side chain entropy upon formation of a calmodulin-peptide complex. *Nat Struct Biol*, 7(1):72–77, 2000.

[53] S. Liang, S. Fuhrman, and R. Somogyi. REVEAL, a general reverse engineering algorithm for inference of genetic network architectures. *Pac Symp Biocomput*, pages 18–29, 1998.

[54] H. Liu, J. Li, and L. Wong. Use of extreme patient samples for outcome prediction from gene expression data. *Bioinformatics*, 21(16):3377–3384, 2005.

[55] J. J. Liu, G. Cutler, W. Li, Z. Pan, S. Peng, T. Hoey, L. Chen, and X. B. Ling. Multiclass cancer classification and biomarker discovery using GA-based algorithms. *Bioinformatics*, 21(11):2691–2697, 2005.

[56] N. Ludtke, S. Panzeri, M. Brown, D. S. Broomhead, J. Knowles, M. A. Montemurro, and D. B. Kell. Information-theoretic sensitivity analysis: a general method for credit assignment in complex networks. *J R Soc Interface*, 5(19):223–235, 2008.

[57] B. H. H. M. K. Gilson, K.A. Sharp. Calculating the electrostatic potential of molecules in solution – method and error assessment. *J Comp Chem*, 9:327–335, 1988.

[58] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms.* Cambridge University Press, Cambridge, UK, 2003.

[59] A. D. MacKerell, Jr., D. Bashford, M. Bellott, R. L. Dunbrack Jr., J. D. Evanseck, M. J. Field, S. Fischer, J. Gao, H. Guo, S. Ha, D. Joseph-McCarthy, L. Kuchnir, K. Kuczera, F. T. K. Lau, C. Mattos, S. Michnick, T. Ngo, D. T. Nguyen, B. Prodhom, W. E. Reiher, III, B. Roux, M. Schlenkrich, J. Smith, R. Stote, J. Straub, M. Watanabe, J. Wiorkiewicz-Kuczera, D. Yin, and M. Karplus. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J Phys Chem B*, 102:3586–3616, 1998.

[60] A. D. Mackerell Jr, M. Feig, and C. L. Brooks III. Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations. *Journal of Computational Chemistry*, 25(11):1400–1415, 2004.

[61] A. A. Margolin, I. Nemenman, K. Basso, C. Wiggins, G. Stolovitzky, R. Dalla Favera, and A. Califano. ARACNE: An algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, 7 Suppl 1:S7, 2006.

[62] H. Matsuda. Physical nature of higher-order mutual information: Intrinsic correlations and frustration. *Physical Review E*, 62(3):3096, 2000.

[63] C. McClendon, G. Friedland, D. Mobley, H. Amirkhani, and M. Jacobson. Quantifying correlations between allosteric sites in thermodynamic ensembles. *J Chem Theory Comput*, 5(9):2486–2502, 2009.

[64] P. E. Meyer, K. Kontos, F. Lafitte, and G. Bontempi. Information-theoretic inference of large transcriptional regulatory networks. *EURASIP J Bioinform Syst Biol*, page 79879, 2007.

[65] F. A. Momany and R. Rone. Validation of the general purpose QUANTA3.2/CHARMm force field. *Journal of Computational Chemistry*, 13(7):888–900, 1992.

[66] A. Montanari and T. Rizzo. How to compute loop corrections to the Bethe approximation. *Journal of Statistical Mechanics: Theory and Experiment*, 2005(10):P10011, 2005.

[67] H. Ney. *Pattern Recognition and Image Analysis*, chapter On the Relationship between Classification Error Bounds and Training Criteria in Statistical Pattern Recognition, pages 636–645. Springer Berlin/Heidelberg, 2003.

[68] A. Nicholls and B. Honig. A rapid finite-difference algorithm, utilizing successive over-relaxation to solve the poisson-boltzmann equation. *J Comp Chem*, 12:435–445, 1991.

[69] L. Paninski. Estimation of entropy and mutual information. *Neural Computation*, 15(6):1191–1253, June 2003.

[70] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(8):1226–1238, 2005.

[71] H. Peng, F. Long, and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans Pattern Anal Mach Intell*, 27(8):1226–1238, 2005.

[72] G. Petrovics, A. Liu, S. Shaheduzzaman, B. Furusato, C. Sun, Y. Chen, M. Nau, L. Ravindranath, Y. Chen, A. Dobi, V. Srikantan, I. A. Sesterhenn, D. G. McLeod, M. Vahey, J. W. Moul, and S. Srivastava. Frequent overexpression of ETS-related gene-1 (ERG1) in prostate cancer transcriptome. *Oncogene*, 24(23):3847–3852, 2005.

[73] M. J. Potter and M. K. Gilson. Coordinate systems and the calculation of molecular properties. *The Journal of Physical Chemistry A*, 106(3):563–566, 2002.

[74] K. S. Ravichandran. Signaling via Shc family adapter proteins. *Oncogene*, 20(44):6322–6330, Oct. 2001. PMID: 11607835.

[75] S. A. Reeves, C. Chavez-Kappel, R. Davis, M. Rosenblum, and M. A. Israel. Developmental regulation of annexin II (Lipocortin 2) in human brain and expression in high grade glioma. *Cancer Res*, 52(24):6871–6876, 1992.

[76] B. Rothhut. Participation of annexins in protein phosphorylation. *Cell Mol Life Sci*, 53(6):522–526, 1997.

[77] C. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423,623–656, 1948.

[78] A. Shimo, T. Nishidate, T. Ohta, M. Fukuda, Y. Nakamura, and T. Katagiri. Elevated expression of protein regulator of cytokinesis 1, involved in the growth of breast cancer cells. *Cancer Sci*, 98(2):174–181, 2007.

[79] S. Shin, K. L. Rossow, J. P. Grande, and R. Janknecht. Involvement of RNA helicases p68 and p72 in colon cancer. *Cancer Res*, 67(16):7572–7578, 2007.

[80] E. Sicinska, I. Aifantis, L. Le Cam, W. Swat, C. Borowski, Q. Yu, A. A. Ferrando, S. D. Levin, Y. Geng, H. von Boehmer, and P. Sicinski. Requirement for cyclin D3 in lymphocyte development and T cell leukemias. *Cancer Cell*, 4(6):451–461, 2003.

[81] D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kantoff, T. R. Golub, and W. R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, 2002.

[82] N. Slonim, G. S. Atwal, G. Tkacik, and W. Bialek. Information-based clustering. *Proc Natl Acad Sci USA*, 102(51):18297–18302, 2005.

[83] Z. Su, H. Hong, H. Fang, L. Shi, R. Perkins, and W. Tong. Very important pool (VIP) genes–an application for microarray-based molecular signatures. *BMC Bioinformatics*, 9 Suppl 9:S9, 2008.

[84] Y. Tang, Y.-Q. Zhang, Z. Huang, X. Hu, and Y. Zhao. Recursive fuzzy granulation for gene subsets extraction and cancer classification. *IEEE Trans Inf Technol Biomed*, 12(6):723–730, 2008.

[85] D. A. Thompson and R. J. Weigel. hAG-2, the human homologue of the Xenopus laevis cement gland gene XAG-2, is coexpressed with estrogen receptor in breast cancer cell lines. *Biochem Biophys Res Commun*, 251(1):111–116, 1998.

[86] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. M. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med*, 347(25):1999–2009, 2002.

[87] J. van der Greef, S. Martin, P. Juhasz, A. Adourian, T. Plasterer, E. R. Verheij, and R. N. McBurney. The art and practice of systems biology in medicine: Mapping patterns of relationships. *J Proteome Res*, 6(4):1540–1559, 2007.

[88] A. Wolf-Yadlin, N. Kumar, Y. Zhang, S. Hautaniemi, M. Zaman, H. Kim, V. Grantcharova, D. A. Lauffenburger, and F. M. White. Effects of HER2 overexpression on cell signaling networks governing proliferation and migration. *Mol Syst Biol*, 2, Oct. 2006.

[89] Y. Yap, X. Zhang, M. T. Ling, X. Wang, Y. C. Wong, and A. Danchin. Classification between normal and tumor tissues based on the pair-wise gene expression ratio. *BMC Cancer*, 4:72, 2004.

[90] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Bethe free energy, Kikuchi approximations and belief propagation algorithms, 2000.

[91] K. Y. Yeung, R. E. Bumgarner, and A. E. Raftery. Bayesian model averaging: development of an improved multi-class, gene selection and classification tool for microarray data. *Bioinformatics*, 21(10):2394–2402, 2005.