THE UNIVERSITY OF CHICAGO

INVERSE STATISTICAL PHYSICS: CONNECTIONS TO MACHINE LEARNING AND APPLICATIONS IN BIOLOGY

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF PHYSICS

BY
PETER WILLIAM FIELDS

CHICAGO, ILLINOIS
MARCH 2026

for nana and papa

Modern mentality reduces reason to a group of categories in which reality is forced to find a place, and whatever does not fall into these categories is defined as irrational. But reason is like an eye staring at reality, greedily taking it in, recording its connections and implications, penetrating reality, moving from one thing to another yet conserving all of them in memory, trying to embrace everything...

Realism requires a certain method for observing and coming to know an object, and this method must not be imagined, thought of or organized and created by the subject: it must be imposed *by the object*... This means that the method of knowing an object is dictated by the object itself and cannot be defined by me.

— Luigi Giussani, *The Religious Sense* (Montreal: McGill–Queen's University Press, 1997)

Who are you who so fill my heart with your absence?

Who fill the entire world with your absence?

— Pär Lagerkvist, "My Friend Is a Stranger"

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to give a very special thanks to my advisor, Stephanie E. Palmer, whose guidance and expertise were indispensable to this work and to my growth as a scientist and as a person. Her kindness and support cannot be overstated. I am also grateful to the other members of my committee—David Schwab, Rama Ranganathan, Arvind Murugan, and Wendy Zhang—for their feedback and counsel. I am especially grateful to have collaborated with Wave Ngampruetikorn, whose scientific rigor never ceased to expand my idea of what it means to be a good researcher.

I am happy to have received support from the Physics Frontier Center for Living Systems through the National Science Foundation award NSF PHY-2317138; the NSF-Simons National Institute for Theory and Mathematics in Biology, awards NSF DMS-2235451 and Simons Foundation MP-TMPS-00005320; the University of Chicago Materials Research Science and Engineering Center, award NSF DMR-2011854; the Center for the Physics of Biological Function, NSF PHY-1734030; the University of Chicago Center for Physcis of Evolving Systems; and by a grant from ICAM, the Institute for Complex Adaptive Matter to Peter Fields.

I am blessed to have many friends and family who accompanied me before and through my PhD journey: my wonderful parents, Jonathan and Susan; my siblings Elizabeth and Justin, along with his wife Emily and their three beautiful children—Joel, David, and my God-daughter Anna; my large, loving, and hilarious extended family—both Fields and Flansburgs; all the fun and friendly members of the Palmer Lab; my close friends Cheyne Weis, Adam Kline, Daine Danielson, Kyle Bojanek, Sophie Colt, Brian Billion and Colin McCarthy; my 2019 UChicago physics graduate student cohort; the always-welcoming Ranganathan lab; the members of the Calvert House Choir and their ever-uplifting voices; my whole Church community and the many reliable friendships it affords me, especially that of Clare Chiodini, Matteo Sabato, and Jamie Ascenzo; my early-graduate and undergraduate advisors, William

Irvine, Robert Messinger, and Paolo Privitera; the various professors who encouraged me in my early physics career, especially Vincent Pauchard, Myriam Sarachik, Mark Shattuck, V. Parameswaran Nair, and James Hedberg; all my friends from home who have stayed in touch over the years—Jane Oberman, Malachi Provenzano, Kaylana Sareen, Owen Kunhardt, Sami Vukelj, Albert Connolly, Antonio Franciosa, Dylan Ditucci-Cappiello, Grace Ronan, Francesco Costanzo, Chris Vath, the Maniscalco, Ronan, Weiner, Kuzmin families, and far too many others to count; the loving memory of departed friends and family, especially Uncle Frank Simmonds, Aunt Mary Vasquez, Grandpa Ted Flansburg, Monsignor Lorenzo Albacete, and my lovely Nana and Papa, Joel and Barbara Fields, to whom this dissertation is dedicated; and a good deal many more friends, who have taught me the meaning of trust in a Power greater than myself and a plan beyond my reckoning.

ABSTRACT

Statistical mechanics was historically significant for its ability to link the microscopic descriptions of matter and energy to the macroscopic observations of thermodynamics. Recent cross-disciplinary work has utilized insights from statistical mechanics to solve the inverse problem—going from observable phenomenon to underlying interactions and principles—and has done so from applications in protein research to theory in neuroscience. Inverse statistical physics relies on a key property of information entropy: that fitting a distribution to data such that it obeys given observed constraints (but is otherwise as maximally entropic as possible) leads to a probabilistic model of the system that is least biased from unobserved assumptions, which leads to maximal predictive power. These maximum entropy models belong to a wider class of regularly used architectures in machine learning, known as energy-based models.

When such models are fit to real data of complex multi-dimensional systems, ideally the learned distribution is able to generate states exemplary of the ground truth. In practice, however, this is often not the case; specialized sampling must be performed to generate desired outputs. For example, energy-based models of sequences of evolutionary related families of proteins have the ability to learn the generic constraints necessary to make novel functional sequences, which have been validated by *in vivo* experiments. However, these learned energy functions require re-scaling by a temperature parameter in order to sample novel functional sequences.

Here we utilize minimal and physically motivated energy-based models in order to systematically interrogate the differences between the data-generation processes of ground truth and learned models sampled at varying temperatures. This lends itself well to an examination of the surprising ability of temperature tuning of learned energy functions—a poorly understood heuristic used across machine learning—to improve sampling performance. Whether the post-hoc sampling temperature need be raised or lowered, and by how much, depends on

several factors: choice of objective function, amount of training data, and most importantly, properties of order and disorder inherent to the true system. Crucially, we show that the need to lower temperature to improve generative performance arises from a tendency of fit models to overestimate the probability mass on excited states when the number of training data is low and the ground truth is characterized by a strong preference for producing few ground states—induced by large "energy gaps" or low ground truth "temperature." Additionally, we show via a minimal setting that the temperature tuning phenomenon may be directly linked to a wide array of empirical evidence for a synergistic cluster of amino acids, or sector, within a sequence that is responsible for determining the functionality of that protein sequence.

CHAPTER 1

INTRODUCTION

Statistical mechanics stands out from the other major fields of physics for its ability to link microscopic descriptions to macroscopic observations. Relations among such things as temperature, heat, and pressure are derived in terms of the statistics of ensembles of particles. Careful coarse-graining schemes of microscopic models through Renormalization Group theory leads to universal characterizations of macroscopically observed phase transitions.

Jaynes [1957] made the remarkable connection between statistical physics and the (then nascent) field of information theory by linking the concepts of thermodynamic entropy and information theoretic entropy (understood as a quantification of uncertainty)—showing that the Boltzmann distribution maximizes this entropy subject to observed constraints on mean energy, particle number, etc. That is, the Boltzmann distribution is that which captures the statistics of observable phenomena and is the "most uncertain" with respect to unavailable information, thereby representing a theory free of any arbitrary assumption beyond those inferrable from observation.

This introduced a new paradigm into our understanding of statistical physics; Jaynes [1957] notes that introducing entropy in this manner frees statistical mechanics from the need to motivate the use of ensembles via arguments of ergodicity, as it is not necessary to attain the functional form of the Boltzmann distribution. Jaynes, however, was still working with notions of energy as an observed conserved quantity, known to be some function of the physical state of affairs: positions and velocities of particles, orientations of magnetic dipoles, and so forth.

In recent years, the Maximum Entropy (MaxEnt) principle of Jaynes has been extended to systems beyond those of traditional physics, particularly in the life sciences [Weigt et al., 2009; Morcos et al., 2011b; Kleeorin et al., 2023b; Mora et al., 2010; Harte, 2011; Bialek et al., 2012; Mora et al., 2016; Lezon et al., 2006; Locasale and Wolf-Yadlin, 2009; Schneidman

et al., 2006; Tkačik et al., 2014; Berry II et al., 2013; Lynn et al., 2025; Russ et al., 2020b; De Martino and De Martino, 2018]. Take, for example, the work of Schneidman et al. [2006], where the authors defined their states (or "spins") as a population of neurons, each taking on a value of 0 or 1 representing the absence or presence of a neural spike. The MaxEnt distribution with respect to observed spike patterns—captured by observed single-neuron means and pairwise correlations of spike rates—then takes on the functional form of an Ising distribution. Importantly, *the interaction terms between "spins" and their "field strengths" were not defined a priori but were inferred from the data.* This led to deeper insight into their system (Schneidman et al. [2006] suggest these discovered interaction networks are indicative of error-correcting properties in the neural code).

This is an inversion of the typical workflow of statistical physics; whereas traditionally we have gone from Hamiltonian to Boltzmann distribution to expectation values we now go from our observables back to a Hamiltonian via the MaxEnt principle. The theoretical power of inverse statistical physics—and its applications beyond traditional systems—is borne out by the syntactic nature of information theory itself. Its theorems hold regardless of whatever realities its symbols happen to be grounded in. $P(x = 1)$ may mean x as magnetic dipole is oriented up or x as neuron is spiking—with some given probability. Consequently, the applications of inverse statistical physics and the MaxEnt principle are far reaching, and stand to bear extension to further discovery of underlying principles in whatever settings they may be applied.

This dissertation is primarily motivated by this above consideration. In particular, it expands on recent results from Russ et al. [2005b]. In this work, the authors fit a pairwise MaxEnt distribution to a dataset comprised of aligned amino-acid sequences corresponding to proteins in an evolutionarily related family of enzymes found across various species. The observables used were the single-site and pairwise frequencies of amino-acid types at each position in the sequence. Interestingly, when the fit models were sampled, the resulting

synthetic proteins were tested *in vivo* and found to recapitulate functional enzymatic activity. However, these functional sequences were only attained when the learned-from-data energy function was rescaled by a *post hoc* (and un-physical) "temperature" parameter. *During* fitting to data, $T = 1$ without loss of generality, as it sets the units of the "energy" function, and the parameters will always rescale accordingly during learning. Russ et al. [2005b] found that lowering temperature *after* fitting allows for functional (and never-before-seen-in-nature) proteins to be generated from the model.

The following chapters are an investigation into this temperature tuning phenomena, what it reveals about the true distributions we are trying to learn, and how this is relevant for protein research, the life sciences, and applications in machine learning where similar tuning phenomena occur. Before we explore these key findings, we first review the core concepts of entropy and model fitting, and how they are well-equipped for problems in quantitative biology.

1.1 Model fitting and information theoretic entropy

Physics, in general, proceeds from a twofold game in which the researcher negotiates between necessary relations among mathematical entities—grounded in observations—and comparison to further observation suggested by those relations.

Consider a simple kinematics problem looking for the time it takes a ball to come back down after being thrown upwards; solving the proper kinematic equation (the one that is quadratic in time) yields a positive and negative value for time-elapsed. The negative value is rejected not for any mathematical reason, but simply because it does not make sense according to the physical story originally told. This back-and-forth between mathematical and physical storytelling characterizes the physicist's mode of thinking. General relativity goes from the equivalence principle to black holes with the aid of differential geometry. The panoply of fundamental particles relies heavily on arguments from symmetry and group

theory—both lines of reasoning confirmed by comparison to continued observation.

When Maxwell originally wrote down his distribution for the velocity of particles in a gas, he could reasonably assume conservation of energy and the isotropy of space (no preferred direction for the particles), helping him towards the functional form of the probability mass function—and all without taking a single measurement of a particle's speed.

Intuition and preference for simplicity have been the chief guides of physicists through this math-to-observation back-and-forth, just as it was for Maxwell. This approach falters in systems as complex as a population of neurons or protein sequences, where simplifying assumptions are rarely justified *a priori*. The advantage of the MaxEnt principle is clear. It gives a formalized approach to going from observation to underlying interactions. But how *exactly* does maximizing informational entropy achieve this?

Let us begin by introducing Shannon's definition of entropy for a distribution $p(x)$:

$$H[p] = -\sum_x p(x) \log_2 p(x), \tag{1.1}$$

where the sum runs over all states in the support of x.

This quantity may best be understood through an example. Consider a distribution of N fair coins (which are necessarily uncorrelated). We have $x \in \{0,1\}^N$ where 0 represents tails and 1 heads. The distribution is uniform over all possible 2^N configurations, and since all coins are independent, we have $p(x) = \frac{1}{2^N}$ for all x. Plugging into (1.1) we have

$$H[p] = -\sum_x \frac{1}{2^N} \log_2 \frac{1}{2^N} = \log_2 2^N = N. \tag{1.2}$$

We can understand entropy as the average minimum number of yes/no questions that must be made in order to specify the state. This makes intuitive sense when considering the coin example. We must ask for each of the N coins whether it is heads or tails in order to learn the state of all of them. We may also see how $\log_2$ enters the equation, as for every new coin

there are 2 times as many possible states of x, and therefore $\log_2 2 = 1$ more question must be asked; with each new coin, the number of possible configurations grows exponentially and the number of questions grows linearly. This also resonates with the notion of entropy as uncertainty. The more questions we must ask, the more uncertain we must be. Note that the choice of base 2 for the logarithm only corresponds to a choice of units. Base 2 gives us units of bits (number of yes/no questions), base 10 gives us dits (number of questions with 10 answers each), and so forth. For the rest of this work we shall work in units of nats given by the natural logarithm, denoted simply as log.

If the coins are somehow biased or correlated than we can come up with a scheme that discovers the states of each coin in less than N questions, on average. For examples, if we have 2 coins that are biased towards heads and correlated with each other, we will often manage to discover the state of both coins by asking 1 question: are they both heads?

We may now see how maximizing the entropy of a distribution while constraining it to reproduce observed statistics gives the least assumptive probability mass function. Let us define the following Lagrangian for the distribution $p(x)$:

$$\begin{aligned}\mathscr{L}[p] = -\sum_x p(x)\log p(x) - \sum_k \lambda_k \left(\langle O_k(x)\rangle_{\text{emp}} - \left(\sum_x p(x) O_k(x) \right) \right) \\ - \mu \left(\sum_x p(x) - 1 \right),\end{aligned} \tag{1.3}$$

where the first term is the entropy to be maximized; the second term is a constraint on the expectation values of observables, $O_k(x)$, that are scalar functions of x and $\langle\cdot\rangle_{\text{emp}}$ is an expectation value with respect to observed states; and the third term is a normalization constraint, with λ_k and μ as the Lagrange multipliers.

Extremizing over p gives us

$$p(x) = \frac{1}{Z} \exp\left(\sum_i \lambda_k O_k(x) \right) \tag{1.4}$$

where we changed variables from μ to Z when defining the normalization constant. We identify the term $-E(x) = \sum_k \lambda_k O_k(x)$ as the energy function. The goal of learning this distribution is to fit the parameters λ_k. If we define the observables as $O(x) = x_j$ and $O(x) = x_i x_j$, and the corresponding Lagrange multipliers as h_i and J_{ij} we obtain

$$p(x) = \frac{1}{Z} \exp\left(\sum_i h_i x_i + \sum_{i<j} J_{ij} x_i x_j \right), \tag{1.5}$$

which is the familiar Ising distribution from statistical physics. The proper choice of the h_i and J_{ij} yield the maximum entropy distribution that respects the observed expectation values according to Eq. (1.3). *By optimizing over (1.3) we have found the distribution that corresponds to the fewest number of questions we may expect to ask in order to specify the state of any new x's that are obtained. Importantly, including any observations that are not measured would yield a necessarily greater number of questions.* To reiterate, the MaxEnt principle gives us the most uncertain $p(x)$ subject to given constraints.

How exactly are the Lagrange multipliers estimated? The objective function to be optimized that yields $\sum_x p(x) O_k(x) = \langle O_k(x) \rangle_{\text{emp}}$ as desired is

$$\mathcal{L}(\mathcal{D}|h_i, J_{ij}) = \frac{1}{M} \sum_{x \in \mathcal{D}} \log p(x|h_i, J_{ij}) = -\langle E(x) \rangle_{\text{emp}} - \log Z(h_i, J_{ij}) \tag{1.6}$$

where the dataset of M samples is given by $\mathcal{D} = \{x^{(1)}, x^{(2)}, ..., x^{(M)}\}$. Minimizing with respect to h_i, J_{ij} gives us the gradients

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial h_i} &= \langle x_i \rangle_{\text{model}} - \langle x_i \rangle_{\text{emp}} \\ \frac{\partial \mathcal{L}}{\partial J_{ij}} &= \langle x_i x_j \rangle_{\text{model}} - \langle x_i x_j \rangle_{\text{emp}}, \end{aligned} \tag{1.7}$$

where $\langle \cdot \rangle_{\text{model}}$ is an expectation with respect to the model we are fitting. Setting these gradients equal to 0 yields the desired constraints. These expectations are in general intractable

to calculate analytically due to the intractability of the partition function Z, and must be obtained via gradient descent methods (see [LeCun et al., 2006; Nguyen et al., 2017b; Mehta et al., 2019] for more information).

It is worth noting some equivalencies of (1.6) with other interesting quantities. We define the cross-entropy as

$$H[f, g] = -\sum_x f(x) \log g(x). \tag{1.8}$$

for two distributions f and g.

This is the extra number of bits (nats) that must be used to encode x if we are using the wrong code via g. To "encode x via some distribution f" is the same as simply coming up with a scheme of questions to discover the state of x. If x occurs with probability $f(x)$, then we can expect to ask $-\log f(x)$ questions to discover the actual value of this state. If the entropy of our distribution is low, we may set up some encoding where we need to ask fewer and fewer questions; that is, make some code-book that relates each state x to a codeword of shorter and shorter length (each symbol in the codeword corresponds to an answer to a question). This codeword is on average of length $\mathbb{E}_f[-\log f(x)]$ if we are using codewords optimized for f, and will intuitively be necessarily longer if we are using codewords optimized for some other distribution g, which is exactly what Eq. (1.8) measures.

We can see that

$$-\mathcal{L}(\mathcal{D}|h_i, J_{ij}) = \sum_x p_{\mathcal{D}}(x) \log p(x|h_i, J_{ij}) = H[p_{\mathcal{D}}, p]$$

where identify $p_{\mathcal{D}}$ as an empirical distribution defined by our data samples.

The relative entropy is the difference of the cross-entropy and entropy and is defined by the Kullback-Leibler divergence

$$D_{\mathrm{KL}}(f\|g) = \sum_x f(x) \log\left(\frac{f(x)}{g(x)}\right) = H[f, g] - H[f] \tag{1.9}$$

where we note that minimizing this divergence between $p_{\mathcal{D}}$ and p, $D_{\mathrm{KL}}(p_{\mathcal{D}}\|p) = H[p_{\mathcal{D}}, p] - H[p_{\mathcal{D}}]$ is equivalent to maximizing (1.6), as $H[p_{\mathcal{D}}]$ is a constant that does not depend on the model parameters.

In summary, fitting a model to data via the MaxEnt principle involves solving the following:

$$\hat{h}_i, \hat{J}_{ij} = \underset{h_i, J_{ij}}{\operatorname{argmax}} \mathcal{L}(\mathcal{D}|h_i, J_{ij}) \tag{1.10}$$

where $\hat{h}_i, \hat{J}_{ij}$ are the parameters obtained at the end of fitting.

1.2 Relevance to problems in the life sciences

One may justifiably object: too intricate are the underlying principles that require so much machinery to uncover, and ought not be considered principles at all. This objection is especially cogent in the life sciences, which is often characterized by seemingly bespoke and unique phenomenon.

But life itself is characterized by a strong degree of sameness within its variability. This is true in the sense that many similar traits can be observed within such a variability, and also in the sense that such characteristics are capable of adapting to a wide variety of circumstances while maintaining their essential function.

Take for example the person whose arms are filled with a few too many things to carry: books, groceries, laundry, and so forth. A good deal many different kinds of people can execute this task, albeit to different degrees, but the essential function is maintained. Furthermore, humans can do this complex task and (often though not always) manage others: pick up the phone, open a door, and so on. Such robustness within a diversity of expression suggests *something(s)* that make it the same.

To return again to the results of Schneidman et al. [2006], their central result indicated the ability of weak pairwise interactions in a MaxEnt distribution (and little to no higher-order

interactions beyond that) to capture the statistics of population activity. In their particular experimental setting, they recorded spike patterns from retinal neural ganglion cells from salamander and guinea pigs (the neurons responsible for reading out an eye's photo-receptors and sending this signal downstream to the rest of the brain), and from a culture of cortical neurons.

The ganglion cells were recorded in response to visual stimuli. And it was surprising that the population activity could be well-described by a weakly-interacting pairwise coupling model—without any recourse to higher-order interactions. This collective activity at the population level suggested error-correcting behaviors, that is, noisy firing patterns and noisy stimuli could be compensated for by correlated structure inherent to the neural population itself, making it advantageous for predictive encoding of stimuli and downstream readout by the rest of the brain. Furthermore, that only weak pairwise interactions were necessary is illustrative of a biological system finding the simplest (or perhaps, only as complicated as necessary) solution to a problem—in this case, how to efficiently encode visual stimuli with limited energy resources on a time budget.

In [Russ et al., 2020a], the authors took a dataset consisting of an alignment of amino-acid sequences corresponding to different instantiations of the chorismate mutase enzyme from different species. Each sequence was represented by a string of symbols, one for each amino acid at each position in the sequence. No information about the physico-chemical properties of the amino-acids, nor any information about the folded structure of each protein were retained in the dataset. Importantly, this alignment of sequences represented a family of evolutionarily related sequences. The sequences themselves implicitly represented combinations of building blocks that survived the process of natural selection.

A similar MaxEnt model was fit to the single-site and pairwise statistics of this alignment: the Potts model—similar to the Ising model but each "spin" (amino acid random variable) can take on more than 2 states. This model was used to generate sequences, and the ability

of these sequences to function—carry out enzymatic activity—inside of *E. coli* cells not only meant that pairwise interactions could capture the necessary constraints leading to function; it also meant that these necessary interactions were encoded in the sequences themselves via evolution and its fitness requirements. MaxEnt models for proteins give us a window into the nature of evolution.

The above examples illustrate the ability of inverse statistical physics to elucidate the *discoverable* principles leading to life's most robust traits.[1]

1.3 Overview

Equipped with these fundamentals and insights, we are now able to approach the question of temperature tuning of energy functions learned via MaxEnt. Chapter 2 will investigate a minimal model used to generate synthetic protein data and train models *in silico*. This minimal setting is motivated by a large body of work indicating the presence of a small subset of amino acids in a sequence whose patterns are known to determine function, known as the sector [Lockless and Ranganathan, 1999; Russ et al., 2005a; Socolich et al., 2005; Halabi et al., 2009a; Reynolds et al., 2011; McLaughlin Jr et al., 2012; Reynolds et al., 2013; Teşileanu et al., 2015; Rivoire et al., 2016a; Raman et al., 2016; Salinas and Ranganathan, 2018a]. Importantly, Ch. 2 will show that the empirically observed sector phenomenon can be directly linked to the need to tune temperature to improve the ability to generate novel and functional sequences.

In Ch. 3, we explore more broadly the role of temperature tuning as a means to improve sampling performance. MaxEnt models are only one of a larger class of models known as *energy-based models*, which is a machine learning procedure that fits a distribution to data with an energy function not necessarily parameterized by MaxEnt Lagrange multipliers and observables, but by a large neural networks [LeCun et al., 2006]. Furthermore, energy-based

1. Many other examples of MaxEnt's ability to discover principles exist (see citations on page 1).

modeling is part of a larger category of machine learning techniques known as *generative modeling* [Bond-Taylor et al., 2021], whereby a probability distribution is learned from data so that new states exemplary of (but not identical to) the training set may be generated. Such models are currently used commercially in large-language models and text-to-image generation. Within the context of machine learning (and especially generative modeling), temperature tuning for improved performance is broadly utilized [Hinton et al., 2015; Qi et al., 2023; Qiu et al., 2023; Guo et al., 2017; Wenzel et al., 2020; Zhang et al., 2024b,a; Ramsauer et al., 2021].

Chapter 3 motivates a reasonable metric for measuring generative performance, and leverages intuitive physics-like properties of energy-based models to illuminate how and why temperature tuning is required in the first place, and under what circumstances it is more beneficial to raise or lower the sampling temperature. Crucially, the results illuminate a non-trivial relation between the lack of a sufficient number of training samples and the nature of the ground truth distribution from which the training set is drawn—namely, that when this data-generating ground truth is in a sufficiently "ordered" or "cold" phase, the resulting fit model will *over*estimate the probability mass on excited states, leading to the requirement for the *sampling temperature* to be lowered in order to improve generative performance.

Finally, Ch. 4 will outline possible future directions of research suggested by the preceding results. First to be shown are preliminary results indicating the possible existence of scaling relations between the optimal *sampling* temperature (to be denoted as τ^*), the number of training samples M, and the *ground truth* temperature, T. Such relations are of interest as it opens up the possibility of more general rules for the relationship between τ^*, M, and T to be discovered across differing ground truth distributions—ones that may either be higher-dimensional or have different *true underlying* energy landscapes.

Second to be discussed in Ch. 4 is the possibility to leverage this new understanding of the temperature tuning effect for improving training and generating when the ground truth

is not in possession, as is the case for all real-world applications of energy-based models. Specifically, the relationship between a fit model's heat capacity and its corresponding optimal sampling temperature can possibly be exploited as a means to ensure better generative performance. Lastly, the ramifications and possible advantages for using these techniques for protein-like distributions (and real protein data itself) are discussed.

CHAPTER 2

UNDERSTANDING ENERGY-BASED MODELING OF PROTEINS VIA AN EMPIRICALLY MOTIVATED MINIMAL GROUND TRUTH MODEL

This chapter is adapted from the paper "Understanding Energy-Based Modeling of Proteins via an Empirically Motivated Minimal Ground Truth Model" *(Fields, Ngampruetikorn, Ranganathan, Schwab, and Palmer, 2023), which was presented at the* ICML 2023 Workshop on Synergy of Scientific and Machine Learning Modeling. *The workshop is non-archival; authors retain full copyright.*

2.1 Introduction

Proteins have evolved flexible ways to achieve the same function with significantly different amino acid sequences across many species [Starr and Thornton, 2016; Cocco et al., 2018a]. Understanding what principles underpin function yet support variation is of both fundamental and technological interest, see, e.g., [Göbel et al., 1994; Neher, 1994; Halabi et al., 2009a; Morcos et al., 2011a; Fowler and Fields, 2014]. Answering this question requires not only a deep biological understanding but also the tools to unlock relevant insights hidden in the data. This has inspired cross-disciplinary efforts among scientists, bioengineers and machine learning practitioners, see, e.g., Rao et al. [2021a,b]; Biswas et al. [2021]; Marks et al. [2011]; Hawkins-Hooker et al. [2021]; Trinquier et al. [2021]; Rives et al. [2021a]; Jumper et al. [2021]; Lin et al. [2023]; Lian et al. [2023]; Madani et al. [2023]; Ziegler et al. [2023]; Sgarbossa et al. [2023]; Lipsh-Sokolik et al. [2023].

Here we focus on energy-based modeling (EBM),[1] which is a current method used to build generative models of protein sequence data [Morcos et al., 2011a; Cocco et al., 2011; Tubiana et al., 2019; Tagasovska et al., 2022]. In particular, we concentrate on EBMs with a pairwise

1. For a review of EBMs, see, e.g., Lecun et al. [2006].

interacting energy function, which underlie direct coupling analysis (DCA), a commonly used technique for fitting a multiple sequence alignment (MSA) of protein sequences from a family of evolutionarily related organisms [Uguzzoni et al., 2017; Russ et al., 2020c; Barrat-Charlaix et al., 2021]. This model, also known as the Potts model in statistical physics [Nguyen et al., 2017a; Cocco et al., 2018a], is characterized by the energy function,

$$E(\mathbf{v} \mid \hat{\theta}) = -\sum\nolimits_i \hat{h}_i(v_i) - \sum\nolimits_{i<j} \hat{J}_{ij}(v_i, v_j) \tag{2.1}$$

where $\mathbf{v} = (v_1, v_2, \ldots, v_N)$ is a sequence with N positions and each position takes q discrete values, $v_i \in \{1, 2, \ldots, q\}$ with $q = 21$ for proteins (20 possible amino acids and one gap state). We let $\hat{\theta} = \{\hat{\mathbf{h}}, \hat{\mathbf{J}}\}$ denote the model parameters. The probability of a sequence is related to its energy via

$$\hat{P}(\mathbf{v} \mid \hat{\theta}) \propto \exp\left(-E(\mathbf{v} \mid \hat{\theta})\right).$$

Importantly, this does not explicitly model physico-chemical potentials governing atomic interactions. Instead, it captures only the information that is encoded in sequence statistics. This class of models has proved surprisingly successful in extracting the relevant constraints of functional proteins, leading to predictions of novel sequences that are verified to be functional in *in vivo* experiments [Russ et al., 2020c].

Training EBMs on real MSAs and taking samples from them is a modeling and empirical challenge, however.[2] Although Potts models are surprisingly expressive and have led to new insights about sequence data [Schug et al., 2009; Bravi et al., 2020], they place an unverifiable assumption on the interactions in the system. In fact, higher-order interactions are likely responsible for critical functions of proteins [Starr and Thornton, 2016; Poelwijk et al., 2019]. Additionally, the high dimensionality and undersampling of sequence data necessitate regularization, which has a nontrivial effect on inference performance [Kleeorin

2. See Song and Kingma [2021] for a recent review of EBM training in more general settings.

et al., 2023a]. Furthermore, EBMs trained on finite samples struggle to imitate the sharp disparity between functional and nonfunctional proteins. The latter do not survive *in vivo* and thus correspond to zero probability or, equivalently, infinite energy, see Fig. 2.4. In practice, fitted models assign finite energy to all sequences, including nonfunctional ones, although with mostly higher energy; as a result, the synthesis of new sequences often returns many nonfunctional proteins. Increasing the fraction of functional sequences requires an *ad hoc* rescaling of the fitted energy function, $\hat{P}(\mathbf{v} \mid \hat{\theta}) \propto \exp\left(-E(\mathbf{v} \mid \hat{\theta})/T\right)$ with the temperature T set after training to be smaller than one, see Russ et al. [2020c].

To explore the learning behaviors and generative performance of EBMs for sequence data, we develop a minimal model of a protein sequence based on the "sector hypothesis" [Lockless and Ranganathan, 1999; Russ et al., 2005a; Socolich et al., 2005; Halabi et al., 2009a; Reynolds et al., 2011; McLaughlin Jr et al., 2012; Reynolds et al., 2013; Teşileanu et al., 2015; Rivoire et al., 2016a; Raman et al., 2016; Salinas and Ranganathan, 2018a]. Briefly, this hypothesis posits that some highly-correlated subset of amino acids (roughly 10-20%) are responsible for determining the function of a sequence in a given protein family through their collective state. Our model of functional sequences captures this important aspect by explicitly disallowing sequences that exceed the mutation threshold away from archetypal 'functional' patterns. We use the sequences generated from this 'ground-truth' model as the training data for Potts models. This setting allows for a relatively controlled investigation of the generative performance of the fitted Potts models. In particular, we ask how often the learned models produce novel functional sequences and how diverse the generated sequences are.

We show that our setting captures the salient learning behavior of EBMs fitted to real sequence data. We explore the generative performance and its dependence on the interplay between model selection via cross-validation and post-training temperature adjustments. We quantify the trade-off between functionality and novelty in sampled sequences by computing

the false positive rate and entropy of fitted models. Finally, we discuss how the lessons from our study guide insight into our understanding of sequence-function relationships.

2.2 Methods

There is much evidence that mutations with strong deleterious effects on a protein's main biochemical functions are confined to a small subset of sequence positions [McLaughlin Jr et al., 2012; Salinas and Ranganathan, 2018a; Poelwijk et al., 2019; Kleeorin et al., 2023a]. Via statistical analyses of MSAs of various protein families, it has been found that the amino acids in this subset are strongly correlated with each other and weakly correlated with those positions outside of the subset [Lockless and Ranganathan, 1999; Socolich et al., 2005; Russ et al., 2005a; Halabi et al., 2009a; Reynolds et al., 2011, 2013; Rivoire et al., 2016a; Raman et al., 2016]. This function-determining, strongly intra-correlated subset of positions, which usually comprises roughly 10-20% of the sequence, defines the sector. See Fig. 2.1B for an example. Here we introduce a minimal model that captures this important aspect of protein sequences.

As shown in Fig. 2.1A, our model consists of a ground truth distribution defined on a sequence of $N = 35$ positions with $q = 5$ amino acid types. The sequence is divided into three parts that reflect the behavior of sequence data from experiments: (i) pairwise interactions important for structure but not involved in function [Morcos et al., 2011a; Uguzzoni et al., 2017; Kleeorin et al., 2023a], (ii) uncorrelated positions, and (iii) function-determining positions modeled by higher-order correlations as defined above, (Fig. 2.1A&C). It is possible in principle for a protein to have multiple sectors, sometimes overlapping. This represents an interesting extension to our framework; as a first step, in this work, we consider only one sector and its effects on learning and generative performance. We emphasize that all the above-defined features do not represent physico-chemical potentials of the underlying interactions, but rather a correlational structure that relates directly to fitness. Unfit non-

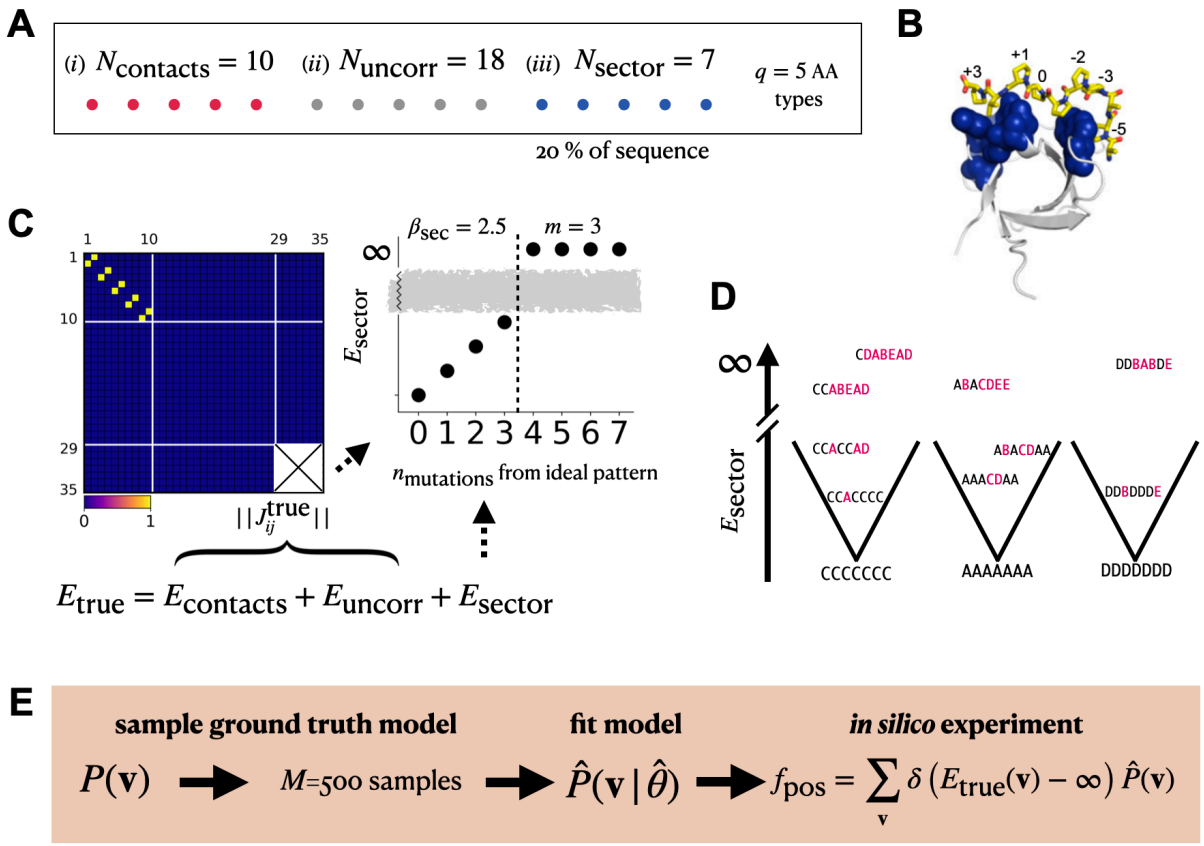

Figure 2.1: *Minimal model for protein sequences* (A) Our model captures three salient features of a sequence: (*i*) pairwise structural contacts, (*ii*) uncorrelated, independent positions, and (*iii*) a highly-correlated sector. (B) Example of a sector in the SH3 family of polyproline-binding domains (PDB ID: 2ABL). Sector positions (blue) are good predictors of ligand-binding sites (yellow). Reproduced with permission from Halabi et al., Cell 138, 774–786 (2009). © Elsevier 2009. (C) Energy function of our model. The heatmap depicts the Frobenius norm of the pairwise couplings, $\|J_{ij}\| = (\sum_{a,b} J_{ij}(a,b)^2)^{1/2}$, for each residue pair i and j. The sector (positions 29-35) cannot be described by pairwise interactions and we model this feature separately; the energy of the sector grows linearly with slope β_{sec} with the number of mutations away from ideal patterns up to a threshold m beyond which the energetic cost diverges. (D) Schematic energy landscape of the sector. Exceeding the mutation threshold (here $m = 3$) makes a sequence nonfunctional, resulting in diverging energy and vanishing probability. (E) Fitting workflow. First, we sample M sequences from our minimal model (see A&C). We use these samples to train the pairwise EBM [Eq (2.1)], from which we generate synthetic sequences. The synthetic sequences that contain more mutations than the mutation threshold contribute to the false positive rate and are *in silico* analogs of nonfunctional proteins in *in vivo* experiments [Russ et al., 2020c; Lian et al., 2023].

functional sequences therefore correspond to sequences with low probability or high energy.

Of the 35 positions, 7 of them, which is 20 percent of positions, comprise the sector, whose energy landscape has 5 energy basins, each defined by unique, non-overlapping patterns (Fig. 2.1C & D). A linear function, whose slope is determined by β_{sec}, controls the energy cost of being $n_{\text{mutations}}$ away from an ideal pattern. Importantly, if the number of mutations exceeds a maximum m, the energy goes to infinity. When this occurs, it corresponds to a function-eliminating mutation, which is to say it is a sequence that cannot rescue function in any biological context. For our experiments, we choose $m = 3$ and $\beta_{\text{sec}} = 2.5$. Results are qualitatively insensitive to these choices (See Fig. 2.6 in the Appendix at the end of this chapter). Ten positions (out of 35) correspond to structural, non-function determining interactions, modeled via Potts interactions such that $J_{ij}(a, b) = 1$ if $a = b$, and zero otherwise for five pairs of positions; all other positions (18 of 35) are uncorrelated (Fig. 2.1C). In addition, we set $h_i(a) = 0$ for all i and a.

The parameters m and β_{sec} represent the defining knobs of the minimal ground truth model that determine the shapes of the energy basins and extent of higher-order interactions within the sector. Thus, our empirically motivated model allows for direct control over the extent to which certain sequences will not be in the support of the fitted model. This ability to know which sequences are and are not in the ground truth support allows for a direct measure of how often a fitted model $\hat{P}(\mathbf{v} \mid \hat{\theta})$ will produce non-functional sequences. This false positive rate will serve as an *in silico* biological experiment, see Fig. 2.1E.

The models are fit via the ratio-matching algorithm [Hyvärinen, 2007], and are 5-fold cross validated to choose appropriate hyper-parameters. The objective and loss functions

are

$$\text{obj}(\hat{\theta}) = \text{loss}(\hat{\theta}) + \lambda_h \sum_{i,a} \|h_i(a)\|^2 + \lambda_J \sum_{i<j,a,b} \|J_{ij}(a,b)\|^2,$$
$$\text{loss}(\hat{\theta}) = \mathop{\mathbb{E}}_{\mathbf{v}\sim\text{Data}} \sum\nolimits_{\mathbf{v}'} A\left(\mathbf{v},\mathbf{v}'\right) \sigma\left(E(\mathbf{v} \mid \hat{\theta}) - E(\mathbf{v}' \mid \hat{\theta})\right)^2$$

where $\sigma(x) = 1/(1+e^{-x})$, $\mathbf{v}'$ are sequences *not* in data, and $A(\mathbf{v},\mathbf{v}') = 1$ if $\mathbf{v}$ and $\mathbf{v}'$ differ in one position and $A(\mathbf{v},\mathbf{v}') = 0$ otherwise. In the following, we take $M = 500$ samples with $\beta_{\text{sec}} = 2.5$ and $m = 3$ and set $\lambda_h = 100$ for all learning tasks.

The entropies of the corresponding fitted models are calculated via annealed importance sampling [Neal, 2001]. Entropy serves as a measure of the fitted model's estimate for the size of functional sequence space. One may state that the goal of learning, as defined by current experimental work, is to reduce the false positive rate as much as possible while keeping the entropy high. This ensures that functional sequences can be of a wide variety, and not simply a memorization of the training data.

2.3 Results

Figure 2.2 illustrates the properties of Potts models [Eq (2.1)] with parameters fitted to samples from our minimal model for protein sequences (Fig. 2.1). We see that the validation loss is smallest at an intermediate regularization strength $\lambda_J = 0.01$ (Fig. 2.2A). But this value does not yield the lowest false positive rate, which occurs at $\lambda_J = 10^{-5}$ (Fig. 2.2B). The ability to generate unseen, yet functional, sequences requires both low false positive rates and high entropy. These competing objectives make model selection nontrivial. Moreover, the standard loss function does not seem to effectively capture relevant performance measures. Taken together, our results exemplify the well-documented challenges of balancing multiple generative objectives and encoding them in a loss function.

To facilitate the ability to sample functional sequences from these learned models, the

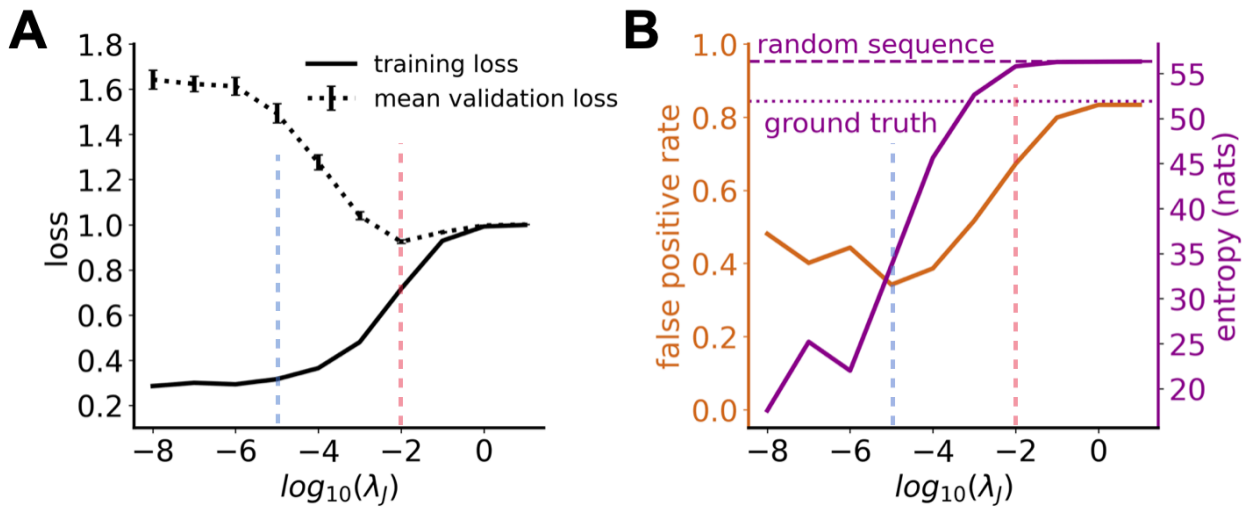

Figure 2.2: *Validation error, false-positive rate, and sequence entropy reach an optimum for different regularization values.* (A) Training (solid) and validation (dashed) losses *vs* regularization strength λ_J. (B) False positive rate (left axis) and entropy (right axis) *vs* λ_J. The vertical lines mark the minimum false positive rate (blue, $\lambda_J = 10^{-5}$) and minimum validation loss (red, $\lambda_J = 0.01$). See the Appendix at the end of this chapter for J_{ij} matrices of the corresponding fit models above.

parameters for each are systematically rescaled via temperature, T. As shown in Fig. 2.3A and B, decreasing T improves the overlap between energy distributions of the training data and sampled data. Fig. 2.3C displays the decrease in false positive rate associated with lowering T. The accompanying decrease in entropy and false positive rate mirrors empirical results found in real data and experiments, where good overlap with training data energy distributions corresponds to the ability to generate novel functional sequences [Russ et al., 2020c].

Why must T be lowered to produce functional sequences? To understand this effect further, we track its impact on sequence entropy and false positive rates for models trained at several regularization strengths. Fig. 2.3D shows the results of this analysis. Here each curve is parameterized by T. For weak regularization, decreasing temperature lowers the entropy with almost no effect on the false positive rate. At intermediate regularizations, a beneficial trade-off occurs as false positive rates decrease while entropy remains high before

dropping off steeply at the lowest temperatures. For strong regularization, the under-fit models lose the ability to generate functional sequences as temperature drops, as not enough structure in the data is found.

2.4 Discussion

Figure 2.4 provides an overview of the intuition behind post-training temperature adjustments for improved generative performance. In Fig. 2.4A, we show a schematic 'true' energy landscape, in which functional sequences (black) occupy low-energy states and non-functional sequences (red) are at infinitely higher energy ($\Delta E \approx \infty$). Ideally, a well-trained model should distinguish functional from non-functional sequences and assign a large energy gap between them. In practice, a tradeoff exists between classification correctness and confidence. Weak regularization results in overfitting, characterized by high misclassification (i.e., many nonfunctional sequences in the energy basins) at high confidence (large energy gap), see Fig. 2.4B. Strong regularization, on the other hand, yields underfitted models, which encode little relevant information in the data and thus cannot classify any sequences with confidence (no energy gap), see Fig. 2.4D. A moderately regularized model has high classification correctness but low confidence (small energy gap), see Fig. 2.4C. Sampling from such a model, despite its low misclassification, can give spurious sequences since their energies, and probabilities, can still be comparable to those of functional ones. In this case, decreasing the temperature during sampling serves as a strategy to increase the confidence of a model and reduce the false positive rate.

We have developed a minimal model that recapitulates the strong, high-order coupling between amino acids in a sector. Under-sampling from this ground truth model (as one does in real sequence analysis) and fitting the 'wrong' Potts models via standard DCA techniques has reproduced the mysterious need to lower temperature in these pairwise models in order to generate functional sequences. We offer the preliminary insight that even though validation

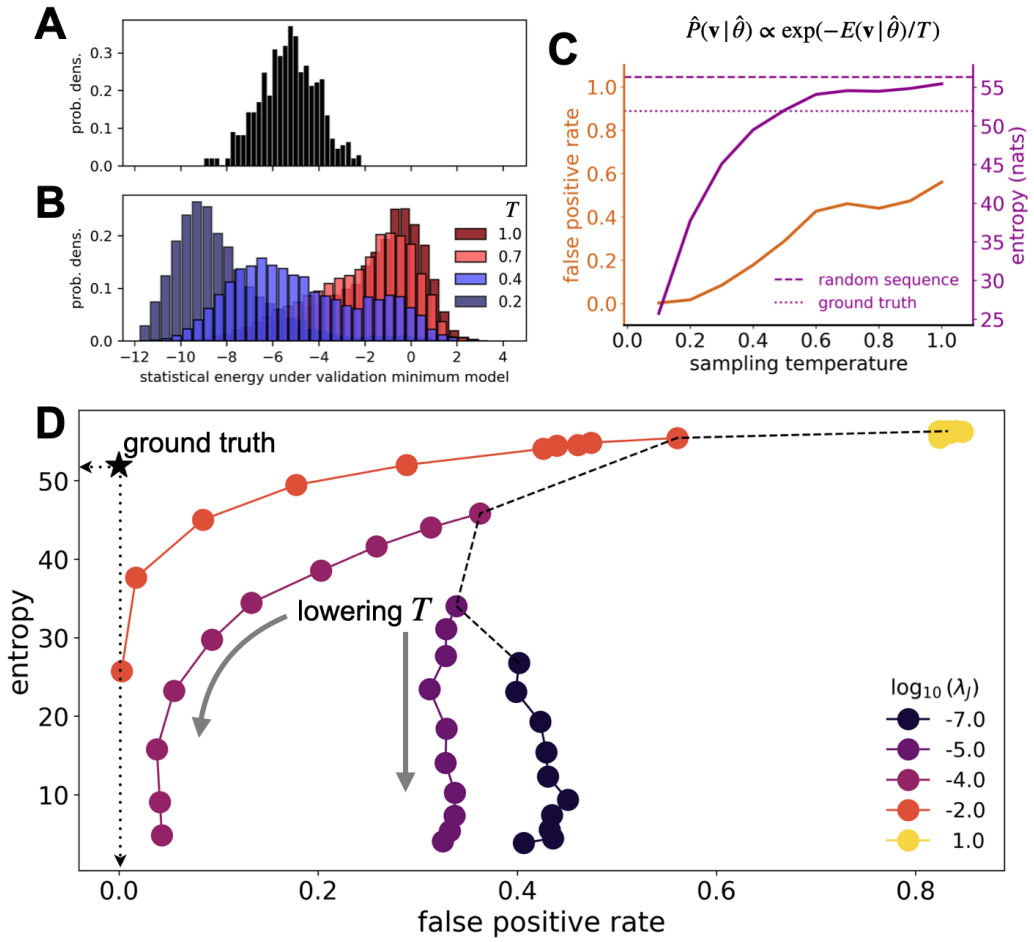

Figure 2.3: *Temperature dependence at the validation minimum.* Histograms of statistical energy of training data (A) and sampled sequences at different temperatures (B) under the fitted validation loss-minimum model. (C) Dependence of false positive rate and entropy of validation minimum model on sampling temperature. Lower temperature produces more functional sequences at the cost of lowering entropy, similar to empirical results in real data [Russ et al., 2020c]. (D) Curves of entropy vs. false positive rate, each parameterized by sampling temperature, for selected models trained at different regularizations. Models at $T = 1$ connected by dotted black line. Note the optimal trade-off between false positive rate and entropy, which corresponds to points in the upper left corner of the graph near the ground truth, occurs as the temperature is lowered on the validation minimum model corresponding to $\lambda_J = 0.01$.

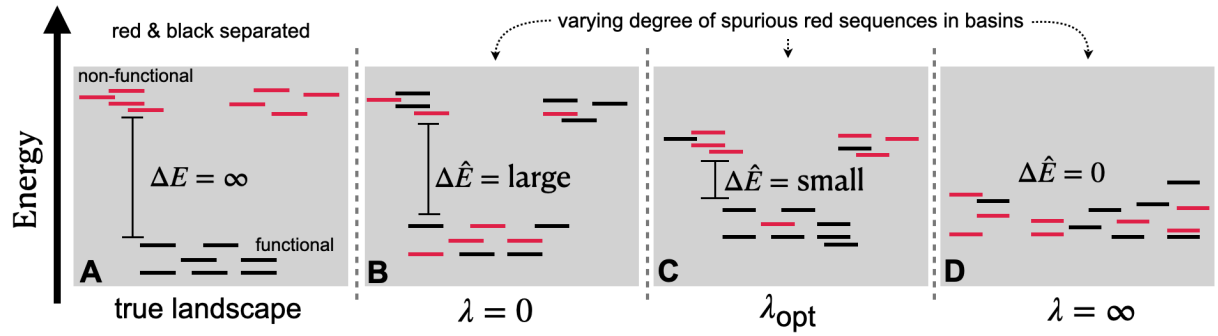

Figure 2.4: *A cartoon of how temperature and regularization interact in EBMs.* (A) Under the ground truth model, functional sequences (black) are low energy, and non-functional sequences (red) are high energy, separated by an energy gap $\Delta E \approx \infty$. (B) When undersampled, an unregularized model $\lambda = 0$, estimates a large energy gap $\Delta\hat{E}$ at the expense of misclassifying functional and non-functional sequences–red and black mixed. (C) For a λ found via cross-validation, red and black sequences are optimally placed on correct sides of the energy gap. This comes at the cost of a smaller value of $\Delta\hat{E}$, which necessitates the lowering of temperature to sample functional sequences, as shown in Fig. 2.3. (D) For no regularization, $\Delta\hat{E}$ goes to zero, and all functional and non-functional sequences have similar energies.

minimum models optimally predict what sequence must be low energy, it does not adequately segregate them from high energy states, and temperature must therefore be lowered. This suggests that our ground truth model could be further exploited in order to understand the optimal training and sampling techniques for real protein sequence data.

APPENDIX FOR CHAPTER 2

All code for generating training data, fitting, and plotting may be found at:

`https://github.com/peter-fields/toysector/`.

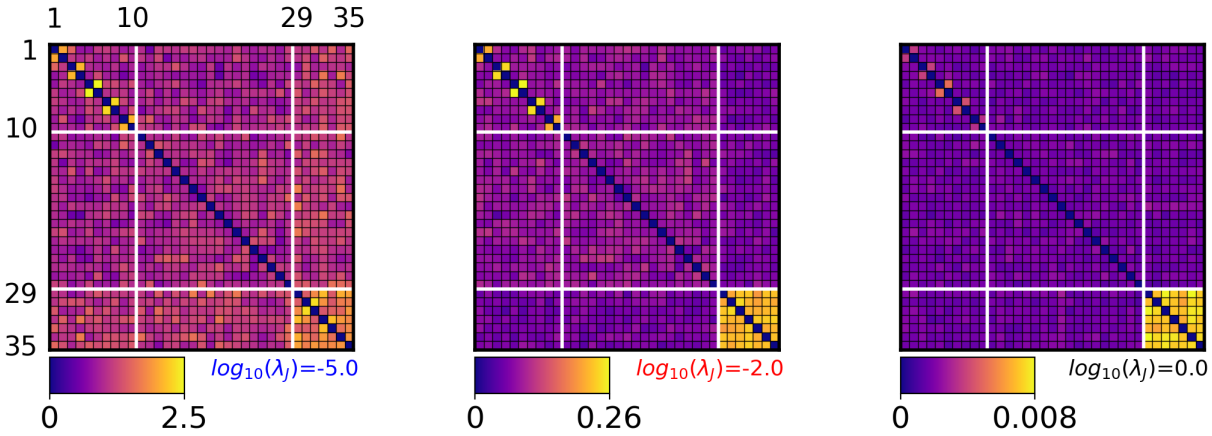

Figure 2.5: Fitted pairwise couplings $J_{ij}(a,b)$ at $\lambda_J = 10^{-5}, 0.01, 1$ (left to right). Frobenius norm over amino acids are shown for each (i,j) pair, $\sqrt{\sum_{a,b} J_{ij}(a,b)^2}$. The model with the smallest false positive rate (as indicated in Fig. 2.2) yields large fractions of erroneous couplings (left), whereas the model with minimum validation loss better separates true features from background noise (middle). The strongest regularization yields the best relative difference between sector-like and isolated-pair-like features, but this comes at the cost of an overall small magnitude of all couplings (right). These results are qualitatively in line with those of Kleeorin et al. [2023b].

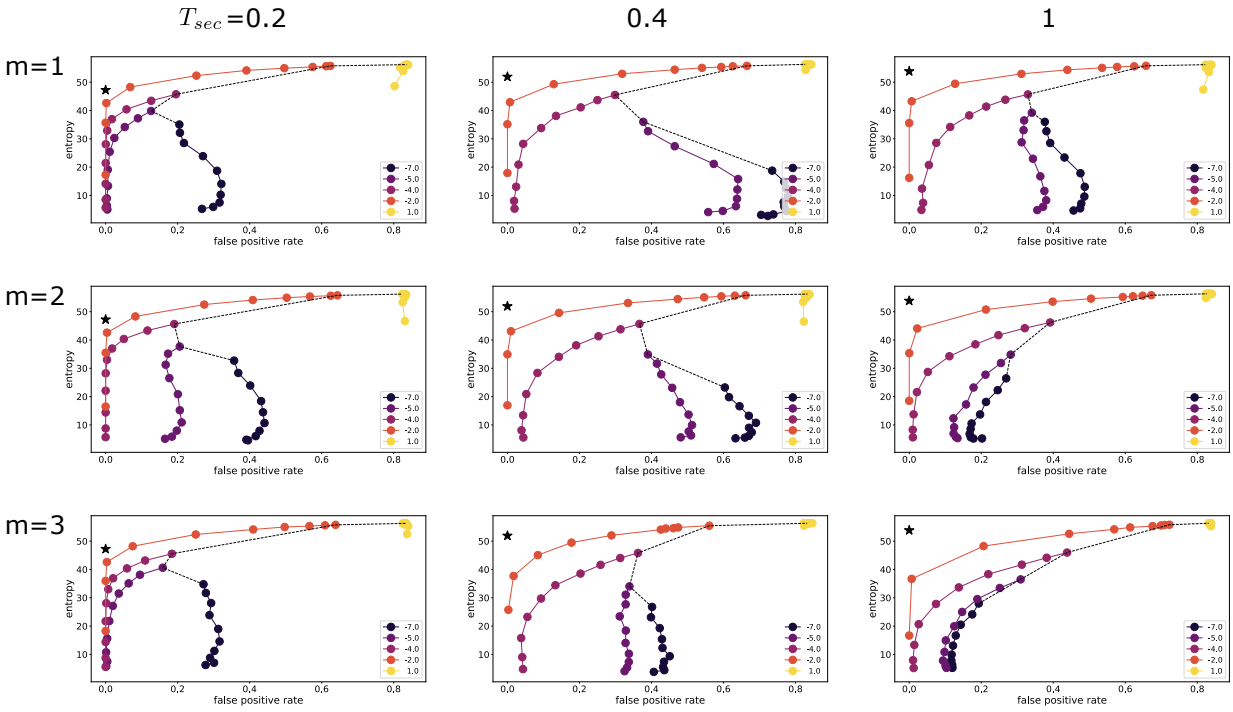

Figure 2.6: Experiments were conducted for slight variations on the ground truth. The maximal number of allowable sector mutations m, and the slope of the sector mutation cost, $\beta_{sec} = 1/T_{sec}$, were altered. Corresponding entropy versus false positive rates (as sampling temperature is swept) plots were generated for each, as done in Fig. 2.3D, where color represents regularization strength (that particular plot reproduced in bottom middle panel here for comparison). Note that results are qualitatively insensitive to choice of ground truth parameters.

CHAPTER 3

UNDERSTANDING TEMPERATURE TUNING IN ENERGY-BASED MODELS

This chapter is based on a manuscript in preparation for submission to PRX Life*, co-authored with Vudtiwat Ngampruetikorn, David J. Schwab, and Stephanie E. Palmer. The submitted manuscript will incorporate relocation of some material to appendices and other minor edits for clarity.*

3.1 Introduction

Energy-based models trained on evolutionary data can now generate novel protein sequences with custom functions [Russ et al., 2020b]. A crucial, yet poorly understood, step in these successes is the use of an artificially low sampling "temperature" to produce functional sequences from the trained model. This adjustment is often the deciding factor between generating functional enzymes and inert polypeptides. A fundamental question arises as to what necessitates temperature tuning and what it reveals about the space of functional proteins and the limits of the models trained on finite data.

Temperature tuning is a broadly used heuristic across machine learning contexts, used to improve training [Hinton et al., 2015; Qi et al., 2023; Qiu et al., 2023], generalization/generative performance [Guo et al., 2017; Wenzel et al., 2020; Zhang et al., 2024b,a], and energy-landscape dynamics for memory retrieval [Ramsauer et al., 2021]. It follows the basic intuition that one can navigate the trade-off between fidelity (producing believable, high-probability outputs at low temperature) and diversity (exploring a wide range of novel outputs at high temperature). Despite its widespread use, this practice lacks a principled, quantitative explanation and has not been systematically connected to known issues of the fitting procedure—particularly how it connects to fundamental limits in the learning process, such as biases introduced by training on finite data [Firth, 1993; Kloucek et al., 2023;

Bordelon et al., 2020; Sorscher et al., 2022; Kleeorin et al., 2023b; Fields et al., 2023].

Inspired by the success of energy-based models in protein synthesis, we investigate the temperature tuning phenomenon using interpretable, physically motivated models. Our central hypothesis traces the need for temperature tuning to a common feature of high-dimensional systems. Many such systems possess a large "energy gap" that separates a small region of meaningful, low-energy states from a vast space of high-energy, unrealistic ones. When trained on necessarily sparse data from such a system, a model develops a specific bias, systematically overestimating the probability of the high-energy states. Lowering the sampling temperature serves as a direct correction for this bias, suppressing the generation of unrealistic states and improving generative fidelity.

However, our framework reveals a richer and more complex picture. The optimal sampling temperature is not a fixed parameter but determined by a quantifiable interplay between the amount of training data and the properties of the underlying energy landscape. Crucially, lowering the temperature is not always optimal. We identify the precise conditions in which raising the sampling temperature maximizes generative performance.

In the course of deriving these results, this work combines concepts from traditionally separate domains: (i) the effect of under-sampling on model bias [Firth, 1993] and its relationship to ground truth properties [Kleeorin et al., 2023b; Fields et al., 2023; Bordelon et al., 2020; Sorscher et al., 2022; Kloucek et al., 2023], (ii) the temperature tuning phenomena [Hinton et al., 2015; Qiu et al., 2023; Wenzel et al., 2020; Zhang et al., 2024a; Ramsauer et al., 2021], and (iii) metrics for evaluating good generative performance [Theis et al., 2016; Salimans et al., 2016; Barratt and Sharma, 2018; Huszár, 2015].

To formally characterize the conditions that determine the optimal temperature, we must first establish a metric that quantitatively captures the trade-off between generative fidelity and diversity. This will allow us to define and investigate an optimal sampling temperature that maximizes generative performance across different systems and data regimes.

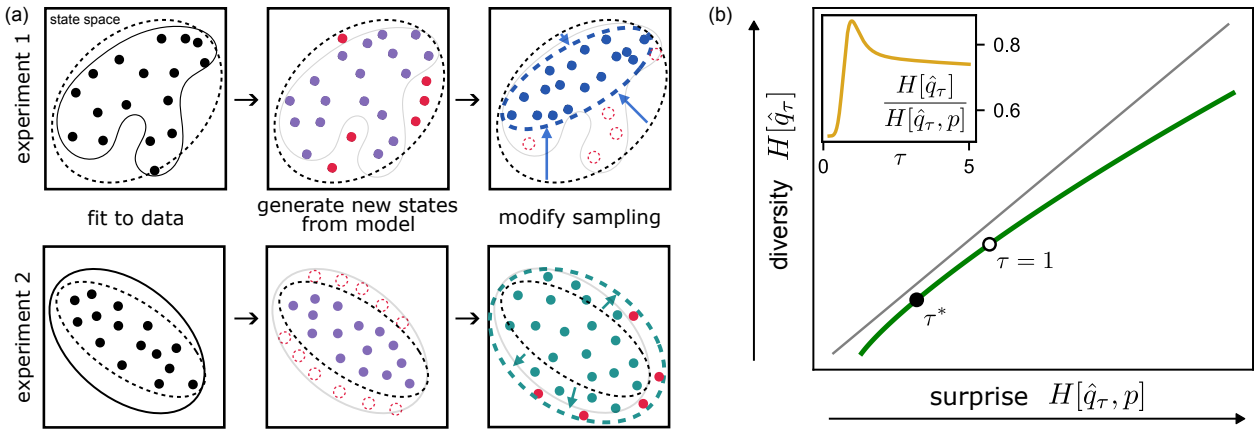

Figure 3.1: (a) Schematic of training, modifying, and sampling generative models. Each box represents the entirety of state space—each dot a data point in that space. (a-Left) Training data (black dots) are generated by a ground truth distribution, p (solid line). A model is fit to these data, $\hat{q}$ (dotted lines). (a-Middle) Generate samples from the model. Samples may either be taken from areas of high probability in the ground truth (purple dots) or from areas of low probability in the ground truth (red solid dots—false positives). Note that depending on the ground truth distribution, the fit model may also fail to generate relevant samples (red empty dots—false negatives). (a-Right) Modifying the sampling technique gives larger proportion of relevant samples. At top, this comes at the expense of missing some areas of ground truth distribution, increasing false negatives. At bottom, this increases the sampling of false positives. (b) The trade-off between sampling states from the fit model that are probable, i.e. less surprising/more believable, according to the true distribution versus the diversity of said states—captured by the trade off between entropy, $H[\hat{q}_\tau]$, and cross-entropy, $H[\hat{q}_\tau, p]$, where the above curve is parameterized by sampling temperature τ of the trained distribution $\hat{q}$. The difference of these two quantities is $D_{\mathrm{KL}}(\hat{q}_\tau || p)$ whose minimum (over τ) is denoted by τ^*. Note that τ^* need not equal 1, which would denote sampling the model "as is." (Inset) Optimal trade-off evidenced by peak in ratio of $H[\hat{q}_\tau]$ to $H[\hat{q}_\tau, p]$.

3.2 Results

3.2.1 Quantifying generative performance

A good generative model must balance two competing objectives: first, it must be able to create believable samples, i.e. states with high fidelity to the true data-generating process; second, these states must cover as wide a range of variety as possible—beyond merely memorizing the data itself. When training data are limited, it is difficult for the fit model to satisfy both goals, and specialized sampling procedures must be adopted to achieve a

trade-off between the two.

Depicted in Fig. 3.1(a) are two scenarios in which the sampling procedure must be modified to obtain optimal generative performance. At left, we can see a fit model, $\hat{q}$, to training data from a ground truth, p. At middle, we can see that sampling such models either leads to sampling false positives, that is, states that are not probable under the ground truth (top), or false negatives, that is, missing states with significant probability mass in the ground truth (bottom). Sampling can be modified such that either: (1) more believable states (according to the ground truth) are sampled (top right) at the cost of a less diverse representation or (2) a more diverse set of states are sampled at the cost of accepting unlikely false positives (bottom right).

A good metric of generative performance is one that measures this trade-off as the broadening/narrowing of sample space is conducted. Figure 3.1(b) introduces the quantities that track the diversity/believability properties inherent in the model sampling procedure, where τ represents the post-fitting "temperature" used to modify the sampling, to be formally introduced in the next section; $\tau = 1$ represents sampling the model "as is."

Consider taking N samples from $\hat{q}_\tau$, creating a synthetic data set $\hat{\mathcal{D}}_\tau$. The frequency with which these samples are on the states of p with significant probability mass may be represented via the quantity $\langle -\log p(v) \rangle_{v \sim \hat{\mathcal{D}}_\tau}$. In information-theoretic terms, the quantity $-\log p(v)$ is considered the surprise of the observation v. The lower the average of this quantity with respect to $\hat{\mathcal{D}}_\tau$, the less surprising (or the more believable) the samples may be considered. If we take the number of samples, N, to infinity, we see that

$$\begin{aligned} \lim_{N \to \infty} \langle -\log p(v) \rangle_{v \sim \hat{\mathcal{D}}_\tau} &= -\sum_v \hat{q}_\tau(v) \log p(v) \\ &= H[\hat{q}_\tau, p], \end{aligned}$$

where we identify $\langle -\log p_T(v) \rangle_{v \sim \hat{\mathcal{D}}_\tau}$ with the cross-entropy of $\hat{q}_\tau(v)$ with p. In contrast to cross-entropy, we may consider the entropy of the fit model, $H[\hat{q}_\tau]$, as it naturally captures

the variability of states *within* the model itself:

$$\begin{aligned}\lim_{N\to\infty} \langle -\log \hat{q}_\tau \rangle_{v\sim\hat{\mathcal{D}}_\tau} &= -\sum_v \hat{q}_\tau(v) \log \hat{q}_\tau(v) \\ &= H[\hat{q}_\tau],\end{aligned}$$

We note further that the difference of these two quantities,

$$D_{\mathrm{KL}}(\hat{q}_\tau || p) = H[\hat{q}_\tau, p] - H[\hat{q}_\tau], \tag{3.1}$$

is simply the Kullback-Leibler divergence of $\hat{q}_\tau$ with p, giving us the desired metric that captures this believability/diversity trade-off. It has been noted elsewhere that including $D_{\mathrm{KL}}(q||p)$ (or a proxy for it) in one's objective function causes the fit model to focus more strongly on modes of the data distribution, as opposed to using $D_{\mathrm{KL}}(p||q)$ alone, and is less susceptible to assigning probability mass to spurious states [Ishida et al., 2025; Huszár, 2015; Mehta et al., 2019; Goodfellow et al., 2014], leading to improved generative performance.

We refer to $D_{\mathrm{KL}}(\hat{q}_\tau||p)$ as the *reversed* D_{KL}, as it is reverses the arguments of the typical objective function used for fitting in many machine learning contexts, $D_{\mathrm{KL}}(p||q)$, which we shall refer to as the *forward* D_{KL}.

As shown in Fig. 3.1(b), as we tune the sampling temperature, τ, obtaining probable states comes at the expense of lacking diversity (lower entropy); having less surprising samples will not imply having diverse samples, nor vice versa. Furthermore, we can see from Fig. 3.1(b) that when the ratio between these two quantities is considered, an optimal value of τ clearly exists (inset) and that the minimum difference at τ^* corresponds to the minimum of Eq. (3.1).

$D_{\mathrm{KL}}(\hat{q}_\tau||p)$ not only gives an intuitive metric for the diversity/believability trade-off, but lends itself to a quantitative interpretation from hypothesis testing theory. Consider an experimenter trying to discern whether a set of K samples come from one of two distributions,

$\hat{q}_\tau$ or p, when the samples in fact come from $\hat{q}_\tau$; the log-likelihood ratio $L_K = \sum_{i=1}^{K} \log \frac{\hat{q}_\tau(v_i)}{p(v_i)}$ is then the optimal quantifier of the hypothesis that the samples come from $\hat{q}_\tau$ and not p [Cover and Thomas, 2006]. For each sample, we may define an increment in the amount of evidence $\delta L_i = \log \frac{\hat{q}_\tau(v_i)}{p(v_i)}$. By Wald's identity [Grimmett and Stirzaker, 2001], the expected value of the number of samples an experimenter must be given before $\sum_i \delta L_i$ passes a threshold, γ, (which they determine beforehand) and they are certain the samples do not come from p, is

$$K_{\text{detect}} \approx \frac{\gamma}{D_{\text{KL}}(\hat{q}_\tau || p)}. \tag{3.2}$$

If we consider, K_{detect} as a function of τ, then

$$R = \frac{K_{\text{detect}}(\tau^*)}{K_{\text{detect}}(\tau = 1)} = \frac{D_{\text{KL}}(\hat{q}_{\tau=1} || p)}{D_{\text{KL}}(\hat{q}_{\tau^*} || p)} \tag{3.3}$$

represents an R-fold increase in the number of samples that can be taken from $\hat{q}_\tau$ and look like they come from p (subject to the experimenter's criteria). See Appendix 3.A for derivation and further information.

To summarize, $D_{\text{KL}}(\hat{q}_\tau || p)$ (the reversed D_{KL}), when considered as a function of sampling temperature τ, captures the trade-off between the ability of a fit model to generate diverse samples versus believable samples with respect to the ground truth distribution, p. This motivates the existence of an optimal sampling temperature, τ^*, that produces states representing the most economical trade-off.

3.2.2 An illustrative model

To illustrate what necessitates the need to tune model temperature to improve generative performance, we set up a simple experiment on a toy model, as depicted in Fig. 3.2. The

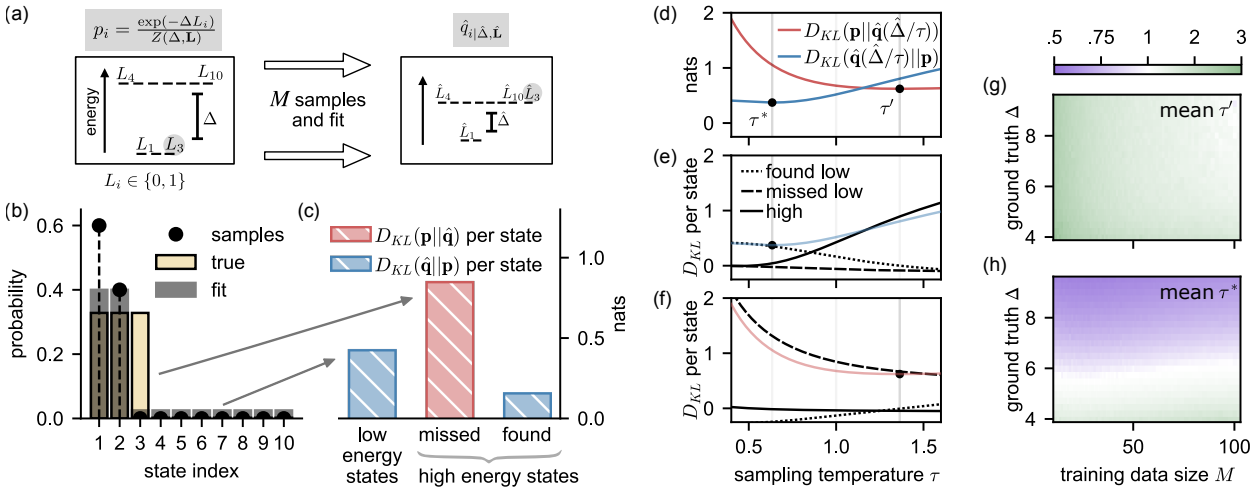

Figure 3.2: Simple toy example of when raising versus lowering temperature improves generative performance. (a) The ground truth consists of a vector, $\mathbf{L}$, that assigns each of 10 states to a low- or high-energy level. Sampling this distribution leads to an empirical distribution over states, $\mathbf{p}_{\text{data}}$, used to get maximum likelihood estimates of the energy level assignment vector and energy gap, $\hat{\mathbf{L}}$ and $\hat{\Delta}$. (b) True model in yellow with $\Delta = 4$ and data distribution, made from 10 samples, represented by dotted lines. Fitting to an under-sampled distribution causes the maximum likelihood estimates to over-estimate the probability mass on high-energy states and under-estimate the mass on the "missed" low-energy state. (c) Checking the decomposition of the forward (red) and reversed (blue) D_{KL}'s between the fit and true distributions. Note the main contributions to each D_{KL}: the missed low-energy states for the forward and the high-energy states for the reverse. (d) Rescaling the inferred energy gap by τ, while keeping $\hat{\mathbf{L}}$ fixed, affects forward and reversed D_{KL}'s. Note that temperature must be changed in opposite directions to achieve minima for each. The minimum of the reversed (blue) D_{KL} corresponds to τ^* in Fig. 3.1. (e-f) D_{KL}'s decomposed into contributions from missed low-energy states, found low-energy states, and high-energy states. Note that raising the temperature, τ, leads to mitigation of the contribution from the missed low-energy state for the forward D_{KL} in (f) and lowering τ leads to mitigation of the contribution from the high-energy states to the reversed D_{KL} in (e). (g-h) Scaled color images of mean optimal τ^* and τ'—calculated from 50 replicates on experiments of a ground truth with $n_l = 20$ and $n_h = 80$ for several values of M and Δ. See Appendix 3.B.3, for details regarding calculation of τ^* and τ'.

ground truth from which we draw samples is given by

$$p_i = \frac{\exp\left(-\Delta L_i\right)}{Z(\Delta, \mathbf{L})}, \tag{3.4}$$

where $\mathbf{L}$ is a vector that assigns each state i to a low- or high-energy level, $L_i \in \{0, 1\}$, Δ is the energy gap between levels, and $Z(\mathbf{L}, \Delta) = \ n_l + \ n_h \exp(-\Delta)$ is the partition function, where n_l and n_h are the number of low- and high-energy states. Sampling this model gives the empirical distribution over states, represented as a vector, $\mathbf{p}_{\text{data}}$. The goal of learning is to get best estimates of the level assignment vector, $\hat{\mathbf{L}}$, and the energy gap, $\hat{\Delta}$. Shown in Fig. 3.2(a) is the workflow of defining a ground truth, generating training samples, and fitting.

This toy model has desirable properties analogous to those in real learning settings, especially when considering the case where $\frac{n_h}{n_l} \gg 1$ and $\Delta \to \infty$. In this case, low-energy states may be thought of as the "realistic" samples within a dataset and high-energy as "unrealistic", e.g., the small number of viable proteins relative to the large number of sequences in the support that the distribution (theoretically) allows for. Importantly, fitting $\hat{\mathbf{L}}$ and $\hat{\Delta}$ corresponds to discovering meaningful states and adequately segregating them from meaningless states. The value of $\hat{\Delta}$ controls the degree of this separation and directly controls how often the fit model will generate "realistic" states when sampled.

As is typically done with real data, the parameters that minimize $D_{\text{KL}}(\mathbf{p}||\mathbf{q})$, where $\mathbf{q}$ represents our model ansatz with the same functional form as $\mathbf{p}$, are found via an empirical

approximation[1] to the true distribution using $\mathbf{p}_{\text{data}}$:

$$\begin{aligned} D_{\text{KL}}(\mathbf{p}||\mathbf{q}_{\tilde{\Delta},\tilde{\mathbf{L}}}) \approx\ & D_{\text{KL}}(\mathbf{p}_{\text{data}}||\mathbf{q}_{\tilde{\Delta},\tilde{\mathbf{L}}}) \\ =\ & -\sum_i p_{\text{data},i} \log q_{i|\tilde{\Delta},\tilde{\mathbf{L}}} - H[\mathbf{p}_{\text{data}}] \\ =\ & -\mathcal{L}(\tilde{\Delta},\tilde{\mathbf{L}}) + \text{constant}, \end{aligned} \tag{3.5}$$

defining the loss function used to get best estimate parameters,

$$\hat{\Delta},\hat{\mathbf{L}} = \underset{\tilde{\Delta},\tilde{\mathbf{L}}}{\text{argmax}}\ \mathcal{L}(\tilde{\Delta},\tilde{\mathbf{L}}) \tag{3.6}$$

with

$$-\mathcal{L}(\tilde{\Delta},\tilde{\mathbf{L}}) = \tilde{\Delta}\tilde{\mathbf{L}}\cdot\mathbf{p}_{\text{data}}\ + \log Z(\tilde{\Delta},\tilde{\mathbf{L}}), \tag{3.7}$$

which correspond to the maximum likelihood estimates as well, where $-\mathcal{L}(\tilde{\Delta},\tilde{\mathbf{L}})$ is the negative log-likelihood function.[2]

For a given estimate of $\hat{\mathbf{L}}$, the estimate of the energy gap is given by

$$\hat{\Delta} = \log\left(\frac{\hat{n}_h}{\hat{n}_l}\cdot\frac{1-\hat{\mathbf{L}}\cdot\mathbf{p}_{\text{data}}}{\hat{\mathbf{L}}\cdot\mathbf{p}_{\text{data}}}\right), \tag{3.8}$$

where $\hat{n}_h = \sum_i \hat{L}_i$ and $\hat{n}_l = N_s - \sum_i \hat{L}_i$ are the model estimates for number of high- and low-energy states; N_s is the total number of states and fixed *a priori*.

This minimal model illustrates the interplay between the training sample M and the ground truth energy gap Δ that leads to the generative temperature-tuning effect. We

1. It is worth noting that the reversed $D_{\text{KL}}(\mathbf{q}||\mathbf{p})$ is not easily fit with a similar approximation, $D_{\text{KL}}(\mathbf{q}||\mathbf{p}_{\text{data}}) = \sum_i q_i \log\frac{q_i}{p_{\text{data},i}}$, as we typically do not have a tractable empirical estimate for this quantity.

2. A pseudo-count regularization term is introduced to ensure no divergences while fitting. We emphasize that this regularization facilitates illustration of core concepts for proceeding sections, and that such regularization schemes are not necessary for the temperature-tuning effect to occur in general. See Appendix 3.B for further details of the sampling and fitting procedure.

are interested in the regime where high-energy "meaningless" states outnumber low-energy "meaningful" states, and training data number is low, resolving this difference poorly. A simple instance of this, where $\Delta = 5$, $n_l = 3$, $n_h = 7$, and $M = 10$ samples, giving the empirical distribution, $\mathbf{p}_{\text{data}}$, is shown in Fig. 3.2(b). For under-sampled training sets such as this, states 3 to 10 have not been sampled. Consequently, states 1 and 2 are assigned as low-energy states by the model and the objective function minimum "misses" one of the low-energy states, assigning it as high-energy, as highlighted in Fig. 3.2(a).

In Fig. 3.2(c), the performance of the fit model $\hat{q}$ is measured according to the two Kullback-Leibler divergences $D_{\text{KL}}(\mathbf{p}||\hat{\mathbf{q}})$ and $D_{\text{KL}}(\hat{\mathbf{q}}||\mathbf{p})$, the forward and reversed D_{KL}, respectively. Each of these quantities is decomposed into the sum over the different states, i.e.

$$\begin{aligned} D_{\text{KL}}(f||g) &= \sum_i f_i \log \frac{f_i}{g_i} \\ &= \sum_{\substack{i \in \text{found} \\ \text{low}}} f_i \log \frac{f_i}{g_i} + \sum_{\substack{i \in \text{missed} \\ \text{low}}} f_i \log \frac{f_i}{g_i} + \ ... \end{aligned}$$

where the sum continues over all other states.

It is clear that the forward D_{KL} (red) is severely penalized by the missed low-energy state, whereas the reversed D_{KL} (blue) suffers more from contributions of the high-energy states. Note that while each high-energy state has low probability in the fit model, the larger number of such states, when compared to the number of low-energy states, leads to a large penalty in the reversed D_{KL}. In addition, the reversed D_{KL} highlights the deleterious effects of higher-energy states—the states which are not desirable for generative performance—where the forward D_{KL} cannot. The forward D_{KL} is exactly the function our objective approximated, Eq. (3.5), and its failure to capture the harm to generativity reflects the misalignment of learning objective and desired performance. This will be explored further in Section 3.2.5.

The generative performance of our fit model is interrogated via rescaling $\hat{\Delta}$ by τ while

maintaining the estimate of $\hat{\mathbf{L}}$. Figure 3.2(d) shows the effect on each D_{KL}. The minimum value for the forward D_{KL} is achieved by raising the temperature. This follows the intuition that an under-sampled fit with low entropy ought have its temperature raised in order to increase probability of those states missed. Figure 3.2(f) shows this explicitly, as this missed state's contribution to the forward D_{KL} is the main one mitigated by raising temperature.

Lowering the temperature mitigates the contribution to the reversed D_{KL} from the high-energy states, as shown in Fig. 3.2(e). This corresponds well with the intuition built up from Fig. 3.1: lowering the temperature allows for sampling more believable states at the cost of lower diversity.

Figures 3.2(g) and (h) depict the results of several experiments on a larger system with $n_l = 20$, $n_h = 80$ for several values of Δ and number of training data, M. From Fig. 3.2(h), we can see the tendency for low M and high Δ to lead generically to the need to lower τ in order to achieve τ^*. The regime where Δ is small, in which it becomes favorable to raise τ, is addressed in the next section. Note also that when considering optimal τ according to the forward D_{KL}, it is never beneficial to lower τ, as shown in Fig. 3.2(g).

The basic intuition from this simplified setting suggests an under-sampled dataset with a wide energy gap between its low- and high-energy states necessitates lowering the temperature on a fit model. Additionally, the comparatively larger number of high-energy states determines the extent to which generative performance is harmed, and therefore, the extent to which τ must be lowered. We expect this intuition to hold beyond this setting, as we expect many true data-generating distributions to exhibit a similar tendency to have far more "meaningless" high-energy states than "meaningful" low-energy ones and corresponding available datasets to be under-sampled.

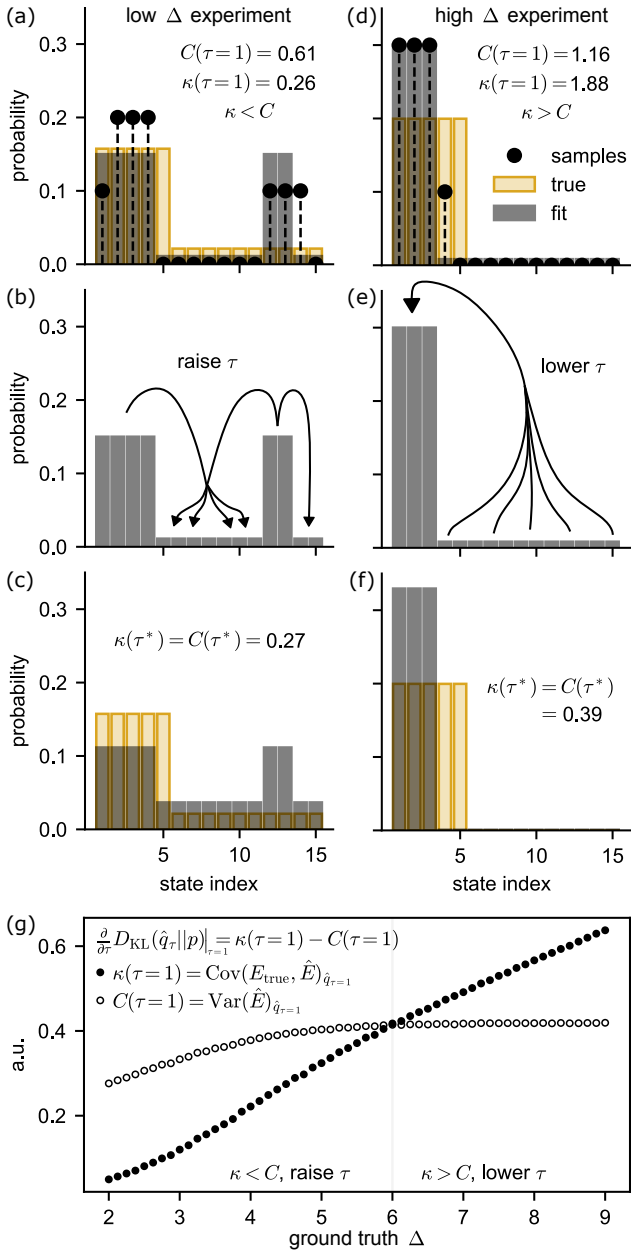

Figure 3.3: κ and C, Eqs. (3.10)-(3.11), determine whether to raise or lower τ in order to improve generative performance. (a-f) Experiments on the illustrative toy model for 5 low-energy states and 10 high-energy states, with models fit to 10 training data. (a-c) One experiment for ground truth $\Delta = 2$. Low training data causes erroneous assignment of excited states as ground states in the model (a), weak correlation of model with true distribution, $\kappa < C$, makes it advantageous to raise τ, (b) and (c). (d-f) One experiment for $\Delta = 7$. Strong correlation of model with true distribution, $\kappa > C$ in (a) makes it advantageous to lower τ, (b) and (c). In (g), each point represents an average of 200 replicates of experiments done for several values of Δ, fixed at 80 training data. The illustrative toy model contains 20 low-energy states and 80 high-energy states as in Fig. 3.2(g) and (h).

3.2.3 When to raise and when to lower sampling temperature τ

To understand the conditions under which τ need be raised or lowered, we must understand the sensitivities of the reversed D_{KL} to changes in τ. To this end we define the following quantities:

$$\frac{\partial D}{\partial \tau} := \frac{\partial}{\partial \tau} D_{\mathrm{KL}}(\hat{q}_\tau || p) = \frac{1}{\tau}(\kappa(\tau) - C(\tau)), \tag{3.9}$$

$$\kappa(\tau) := \tau \frac{H[\hat{q}_\tau, p]}{\partial \tau} = \frac{1}{T_p \tau} \mathrm{Cov}(E_{\mathrm{true}}, \hat{E})_{\hat{q}_\tau}, \tag{3.10}$$

$$C(\tau) := \tau \frac{\partial H[\hat{q}_\tau]}{\partial \tau} = \frac{1}{\tau^2} \mathrm{Var}(\hat{E})_{\hat{q}_\tau}. \tag{3.11}$$

where T_p is a temperature parameter for the ground truth and where we identify the latter quantity as the heat capacity from traditional statistical physics.

Equation (3.9) indicates whether τ should be raised or lowered, as it determines the gradient along which changes in τ lead to decreases in the reversed D_{KL}.

Furthermore, as we also see from Eq. (3.9) the sign of $\partial D/\partial \tau$ is controlled by the relative difference between κ and C. These two quantities may be thought of susceptibilities parameterized by τ. κ measures the propensity for probability mass to come off of true low-energy states if τ is raised and is proportional to the covariance of true and fit energy functions. If $\hat{E}$ is strongly correlated with E_{true}, there is a stronger penalty for raising τ and taking mass off of those true low E states. C measures propensity for probability mass to cover more states as τ is raised.

Figure 3.3 shows the two scenarios of the illustrative toy model for 15 states in which τ is adjusted (5 low energy, 10 high, and $T_p = 1$ without loss of generality). Panels (a-c) contain 1 experiment for a ground truth $\Delta = 2$. In (a) some true excited states are assigned as model ground states, leading to poor correlation of $\hat{E}$ with E_{true} and $\kappa < C$. Panel (b) shows shows τ being raised to mitigate this mismatch and in (c) we can see that at optimal τ^* the two are equal and therefore $\partial D/\partial \tau = 0$. Panels (d-f) show an experiment for ground

truth $\Delta = 7$. Since $\kappa > C$, the penalty for taking mass off of true low energy states is too great relative to the propensity to cover more of state space, and τ should therefore be lowered.

Figure 3.3(g) shows experiments for many values of ground truth Δ, with 200 replicates each, for 20 low-energy states and 80 high-energy states, corresponding to the same system used in Fig. 3.2(g) and (h), with 80 training data for each replicate. We can see that on average, in this low sample regime, low true energy gaps produce the need to raise tau as $\kappa < C$, and high energy gaps necessitate the need to lower tau because $\kappa > C$.

3.2.4 Structured energy landscape

To test whether the intuitions from our simple two-level illustration carry to more general settings, we conduct experiments on training samples drawn from a nearest-neighbor Ising distribution as the ground truth.

The ground truth distribution from which we draw our training data is defined as:

$$p_T(v) = \frac{1}{Z(T)} \exp\left(\frac{1}{T} \sum_{\langle ij \rangle} J^0 v_i v_j\right), \tag{3.12}$$

where $\langle ij \rangle$ indicates the sum is taken over all nearest neighbors on the two-dimensional 4×4 lattice with periodic boundary conditions and $J^0 = 1$. We take M samples from this at temperature T to make the data-set $\mathcal{D}_T = \{v^{(1)}, v^{(2)}, ..., v^{(M)}\}$. We then fit to these samples by minimizing the negative log-likelihood objective function via gradient descent (see Appendix 3.C.1) such that $\hat{\mathbf{J}} = \underset{\tilde{\mathbf{J}}}{\text{argmin}} \left(-\mathcal{L}\left(\mathcal{D}_T | \tilde{\mathbf{J}}\right)\right)$ given by:

$$-\mathcal{L}\left(\mathcal{D}_T | \tilde{\mathbf{J}}\right) = -\frac{1}{M} \sum_{m=1}^{M} \log q(v^{(m)} | \tilde{\mathbf{J}}), \tag{3.13}$$

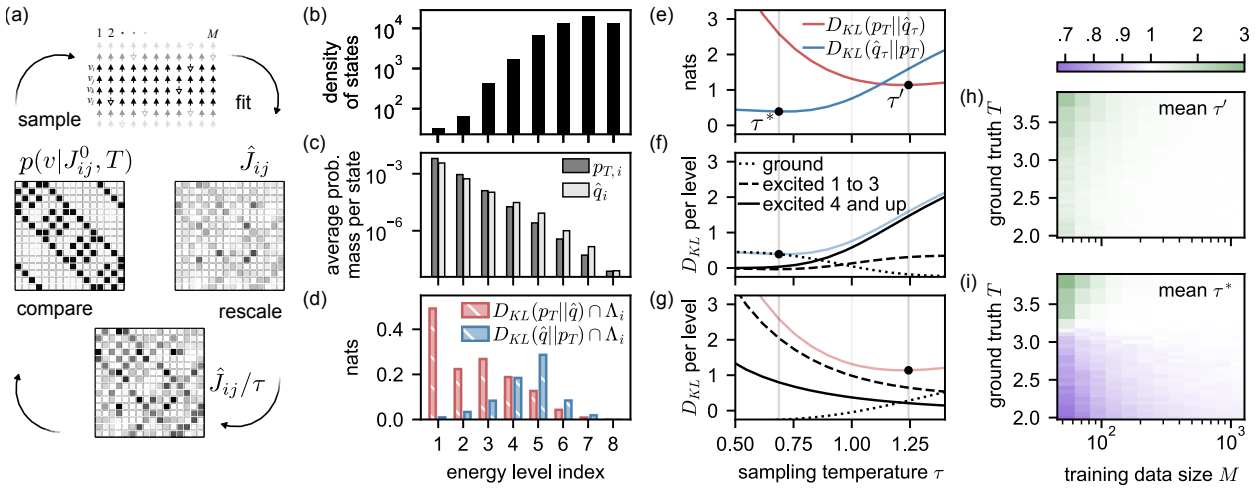

Figure 3.4: Experiments on a 4×4 nearest neighbor Ising model. (a) Starting at left and going clockwise, the ground truth model is defined at a given temperature T by Eq. (3.12). M samples are taken to form the training set, $\mathcal{D}_T$. These data are fit via minimization of Eq. (3.13) to give the parameters $\hat{\mathbf{J}}$. The generative properties are measured via $D_{\mathrm{KL}}(p_T||\hat{q}_\tau)$ and $D_{\mathrm{KL}}(\hat{q}_\tau||p_T)$ (see text). (b-d) Breakdown of contributions to each D_{KL} by energy level, Eqs. (3.17)-(3.18). (b) Density of states for the first 9 excited energy levels of a 4×4 nearest neighbor Ising model. (c-g) Results from one experiment where $M = 93$ and $T = 2.3$. (c) The average amount of probability per state within each energy level, as described by Eqs. (3.19) and (3.20). The amount of probability per state is underestimated by $\hat{q}$ for lower excited states and overestimated for higher excited states. (d) The contribution to each D_{KL} per energy level. The lower excited are more deleterious to the forward D_{KL} (red) and the higher excited states are more deleterious to the reversed D_{KL} (blue). (e-g) Raising versus lowering sampling temperature τ and its dependence on contributions to each D_{KL} from states at different energy levels. (e) D_{KL}'s as a function of sampling temperature τ. Note the minima of each are located on opposite sides of $\tau = 1$. (f) The reversed D_{KL} per energy level. Lowering τ to τ^* mainly mitigates contributions from higher excited states. (g) The forward D_{KL} broken down by contributions from different energy levels. Raising τ to τ' decreases the major contribution from excited states 1-3. (h-i) Ten replicates of an experiment at each T and M are conducted and the corresponding optimal τ^* and τ' are found for each. The scaled color image depicts the average over replicates.

where the model ansatz is given by

$$q(v \mid \tilde{\mathbf{J}}) \propto \exp[-E(v \mid \tilde{\mathbf{J}})], \tag{3.14}$$

and the energy function is defined as:

$$E(v \mid \tilde{\mathbf{J}}) = -\sum_{i<j} \tilde{J}_{ij} v_i v_j. \tag{3.15}$$

We then take the fit model energy function and re-scale it by a post-training temperature, τ,

$$\hat{q}(v|\hat{\mathbf{J}}, \tau) = \exp\left[-E(v|\hat{\mathbf{J}})/\tau\right]/\hat{Z}(\tau). \tag{3.16}$$

The performance of samples drawn from $\hat{q}$ may then be measured by $D_{\mathrm{KL}}(p_T||\hat{q}_\tau)$ and $D_{\mathrm{KL}}(\hat{q}_\tau||p_T)$. This workflow is depicted in Fig. 3.4(a). Note that the ansatz for our fit model does not know *a priori* which J_{ij} are finite; both the spatial structure and coupling strength are inferred from data.

The contributions to each D_{KL} may be broken down to its contributions from each state. A natural way to do this is to consider those states, v, that are on the same energy level in the ground truth distribution, that is,

$$\Lambda_i = \{v_k : E(v_k|\mathbf{J}^0) = E_i\}, \tag{3.17}$$

where E_i is the energy of the ground truth distribution's i^{th} level, with E_0 the ground state

energy and increasing i are higher and higher energies. We then consider $D_{\mathrm{KL}}(f||g)$ as

$$\begin{aligned} D_{\mathrm{KL}}(f||g) &= \sum_{v\in\Lambda_0} f(v)\log\frac{f(v)}{g(v)} + \sum_{v\in\Lambda_1} f(v)\log\frac{f(v)}{g(v)} \\ &\qquad + \sum_{v\in\Lambda_2} f(v)\log\frac{f(v)}{g(v)} + \ ... \\ &= \sum_i \sum_{v\in\Lambda_i} f(v)\log\frac{f(v)}{g(v)} \\ &= \sum_i \left(D_{\mathrm{KL}}(f||g)\ \cap\Lambda_i\right), \end{aligned} \tag{3.18}$$

where we have defined $D_{\mathrm{KL}}(f||g)\ \cap\Lambda_i$ to be the per-energy-level D_{KL}.

Furthermore, to compare the total amount of probability mass on each Λ_i assigned by each model, we consider the quantities

$$\hat{q}_i = \frac{1}{g_i}\sum_{v\in\Lambda_i}\hat{q}(v) \tag{3.19}$$

$$p_{T,i} = \frac{1}{g_i}\sum_{v\in\Lambda_i} p_T(v) = \frac{1}{Z(T)}\exp\left(-E_i/T\right), \tag{3.20}$$

where $g_i = |\Lambda_i|$ is the density of states per energy level.

In Fig. 3.4(b) we see the values of g_i for the first eight excited states in the 4×4 nearest neighbor Ising model. In Fig. 3.4(c), the associated probabilities per level of $\hat{q}$ and p_T are shown for one experiment at low T and M. Note that lower excited states are, on average, underestimated, $p_{T,i} > \hat{q}_i$ for $i = 1, 2$ and 3, while higher excited states are overestimated, $p_{T,i} < \hat{q}_i$, for $i = 4$ and up.

Figure 3.4(d) shows the breakdown of each D_{KL} in accordance with Eqs. (3.17) and (3.18). In analogy to the "missed" low-energy states of Fig. 3.2(c), lower energy states in the nearest-neighbor Ising model (excited states 1 through 3, which are underestimated by $\hat{q}$) form the main contribution to $D_{\mathrm{KL}}(p_T||\hat{q})$. Overestimation of higher excited states by $\hat{q}$ gives the

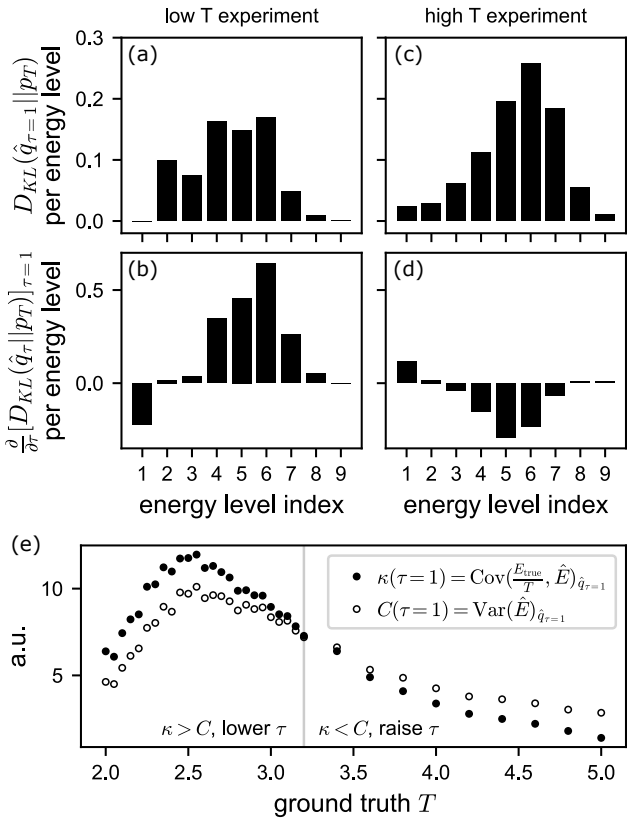

Figure 3.5: Per energy-level breakdown of of reversed D_{KL} and its derivative with respect to τ reveals what dictates the need to raise and lower τ. (a-b) Show results of an experiment done on the 4×4 nearest-neighbor Ising distribution at $T = 2.3$ and $\hat{q}$ trained on $M = 54$ samples. In (a) contributions to the reversed D_{KL} are shown for the first 9 excited energy levels (b) and contributions to its derivative w.r.t. τ evaluated at $\tau = 1$ are positive, indiciative of a need to lower τ. (c-d) Results of an experiment done at $T = 4$ for $M = 54$. Strong contributions to reversed D_{KL} also come from excited states (c), however negative contributions dominate the derivative (d), revealing a need to raise τ. (e) Many experiments done on the 4×4 Ising distribution at various ground truth T. Each point represents an average over 10 experiments done with 54 training data each. For low T, the need to lower τ is necessitated by a strong correlation to the true energy function relative to the model's energy variance; $\kappa > C$, corresponding to a positive value of $\frac{\partial}{\partial\tau} D_{\mathrm{KL}}(\hat{q}_\tau||p_T)\Big|_{\tau=1}$. For high T, the intra-model variance dominates, and probability mass can spread out over state space faster than it comes off of true low-energy states, i.e. $\kappa < C$ and τ should be raised.

main contribution to $D_{\mathrm{KL}}(\hat{q}||p_T)$—and this in spite of the fact that $\hat{q}$ is small in absolute terms. The larger number of states in higher energy levels greatly multiply the deleterious effects of each $\hat{q}(v)$ to $D_{\mathrm{KL}}(\hat{q}||p_T)$, as can be seen in Figs. 3.2(b) and (c) as well.

In Fig. 3.4(e-g), we see how optimal sampling temperature mitigates different pathologies captured by the different D_{KL}'s (cf. Fig 3.2(d-f)). Sampling temperature τ must be changed in opposite directions to improve performance, depending on choice of D_{KL} (Fig. 3.4(e)); the contributions from higher overestimated excited states must be mitigated by lowering *tau* to lower the reversed D_{KL}, (f); and raising τ mitigates the contributions from "missed" lower energy states to the forward D_{KL}, (g).

Figure 3.4(i) shows the regularity with which τ must be lowered to τ^* in order to improve $D_{\mathrm{KL}}(\hat{q}_\tau||p_T)$. Ten replicates of each experiment were done for a ground truth distribution of Eq. (3.12) at several different values of T and M. The average value of τ^* over each set of replicates is reported, generically showing the requirement to lower τ at low training sample number, M, and low ground truth T. Interestingly, τ', which minimizes the forward D_{KL}, must always be raised from $\tau = 1$, if at all (Fig. 3.4(h)), distinguishing τ^*'s and the reversed D_{KL}'s ability to capture the necessity for lowering sampling temperature.

Figure 3.5 depicts the difference in conditions under which τ should be raised or lowered as dictated by the reversed D_{KL}. For an experiment done at low M and *low* T in (a) and (b), we see the typical strong contribution from the excited states. The negative value of the derivative with respect to τ of the reversed D_{KL} indicates it is advantageous to lower τ. For an experiment done at low M and *high* T in (c) and (d), a similar strong contribution from excited states to the reversed D_{KL} dominates. However, the derivative is negative, and therefore implies τ should raised.

The difference between κ and C—which track the strength of the covariance of $\hat{E}$ with E_{true} (the model's susceptibility to move probability mass on/off true low-energy states) and the model's energy variance (the model's susceptibility to spread out probability mass over

state space in general)—control the sign of $\partial D/\partial\tau$, and therefore the direction in which τ should be changed (cf. Fig. 3.3(g)). This is clearly seen in Fig. 3.5(e). At low temperatures $\kappa > C$, and τ must be lowered, while at high T, $\kappa < C$ and τ should be raised.

3.2.5 Bias introduced from empirical approximation of objective function

Without the true distribution, the objective function must be approximated as in Eq. (3.5), $D_{\mathrm{KL}}(p||q) \approx D_{\mathrm{KL}}(p_{\mathrm{data}}||q)$. What kind of bias might this introduce to learning?

Figure 3.6(a) depicts the probabilities of the first $10,000$ states of 4×4 nearest-neighbor Ising model, the associated empirical distribution $p_{\mathcal{D}}$ from training data consisting of $M = 54$ samples from a ground truth distribution at $T = 2.3$, and the associated fit model $\hat{q}$. Note that the minimum value $p_{\mathcal{D}}(v)$ can take is $1/M$, which is well above the probability of higher excited states. Concomitantly, $\hat{q}$, representing our best guess of p_T, overestimates many of these excited states, harming the reversed D_{KL} (Fig. 3.6(b)).

Figure 3.6(c) shows values of $D_{\mathrm{KL}}(p_{\mathcal{D}}||p_T)$ and $D_{\mathrm{KL}}(\hat{q}||p_T)$, each broken down according to their contributions from the first 7 excited energy levels, as done according to Eq. (3.18).

Reported are averages over 50 replicates of an experiment with $T = 2.3$, $M = 54$ for a 4×4 Ising distribution. To ensure adequate comparison across the $P = 50$ replicates, we calculate means reported above as $\langle D_{\mathrm{KL}}(f||g)\rangle \cdot \frac{1}{P}\sum_{k=1}^{P} \frac{[D_{\mathrm{KL}}(f||g)\cap\Lambda_i]_{(k)}}{[D_{\mathrm{KL}}(f||g)]_{(k)}}$, where the first term is the mean of the total D_{KL} and the second term is the mean fractional contribution from each energy level to the total D_{KL}.

We see that, on average, for these excited energy levels we have

$$D_{\mathrm{KL}}(p_{\mathcal{D}}||p_T) \cap \Lambda_i \gtrsim D_{\mathrm{KL}}(\hat{q}||p_T) \cap \Lambda_i,$$

with excited states 3 to 5 contributing most of this systematic overestimation, and therefore, adversely affecting generative performance. See Appendix 3.D for further discussion on this

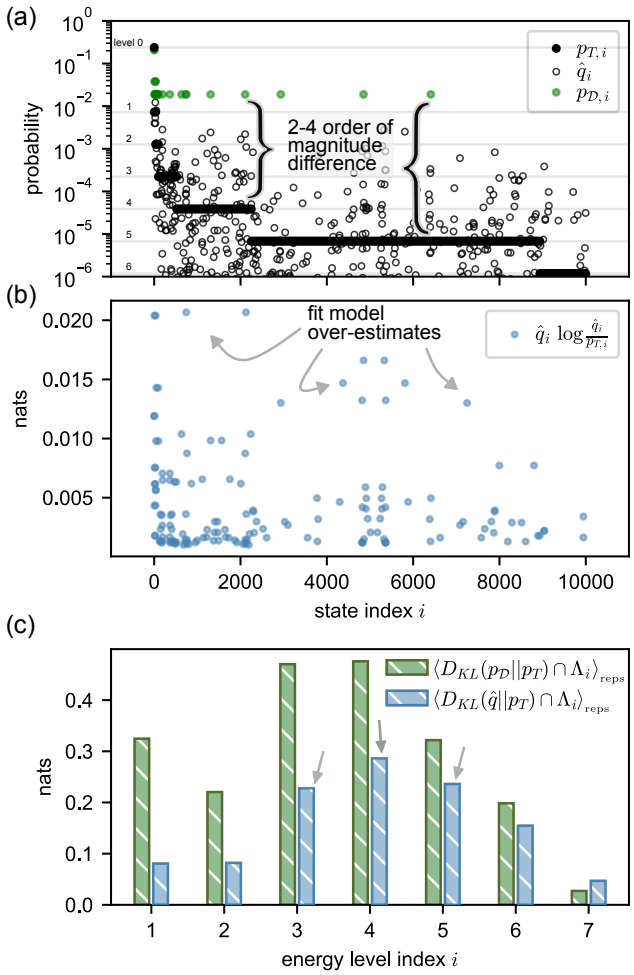

Figure 3.6: Bias introduced by the empirical approximation to $D_{\mathrm{KL}}(p||q)$ with $D_{\mathrm{KL}}(p_{\mathcal{D}}||q)$. (A) Enumerated probability masses for the first 10,000 states of the first 6 energy levels of a 4×4 nearest neighbor Ising model, for a ground truth p_T at $T = 2.3$. $p_{\mathcal{D}}$ is the empirical distribution formed by $M = 54$ samples from p_T. $\hat{q}$ is the maximum likelihood estimate found via Eq. (3.13). Note that for excited energy states we have $p_{\mathcal{D}} \gg p_T$ ($\hat{q} \gg p_T$ for many, though not all, states, as well). $\hat{q}$ is down-sampled to 1000 randomly chosen states for clarity. (B) $\hat{q}_i \log \frac{\hat{q}_i}{p_{T,i}}$ is the contribution to the reversed D_{KL} per state (values below 10^{-3} omitted for clarity). (C) D_{KL}'s decomposed according to Eq. (3.18) and averaged over 50 replicates. $D_{\mathrm{KL}}(p_{\mathcal{D}}||p_T)$ is generically large for higher excited states, and consequently so is $D_{\mathrm{KL}}(\hat{q}||p_T)$.

bound.

The excited energy levels in the empirical distribution, $p_{\mathcal{D},e}$, are systematically over-represented with respect to $p_{T,e}$, and the $\hat{q}_e$ are overestimated accordingly.

3.3 Discussion

We have established a strong and quantitative connection between the need to lower sampling temperature, the amount of available training data, and important properties of the true data-generating distribution—namely the size of meaningful state space versus total state space and the energy landscape that segregates these states, captured by a ground truth density of states g_i and ground truth temperature T. These quantities are especially interesting because learning them is representative of the fundamental goals of learning a generative model: to learn the underlying structure of the data and be able to generalize to states not seen and determine their relevance for desired performance.

In fact, in many learning contexts the high dimensional state space is vastly under-sampled, and only some small fraction of this state space has meaningful states. The fact that a small amount of the support has most of the probability mass is analogous to a system with a wide energy gap between its ground and excited states (or low *true* temperature that effectively increases the gap). The need to lower *sampling* temperature in fit models arises from the fact that the model inaccurately overestimated the probability mass on excited states.

The temperature tuning phenomenon in energy-based models may have implications beyond applications in protein science, as these models are used in various fields across the life sciences, where similar conditions (low sample number and small meaningful sub-space) hold. The maximum entropy principle [Jaynes, 1957]—a form of energy-based modeling whereby a distribution is fit to data such that it reproduces observed statistics but is otherwise as uncertain as possible—has seen success in discovering underlying principles across several

biological domains: (i) within the context of data-driven protein modeling [Weigt et al., 2009; Morcos et al., 2011b; Kleeorin et al., 2023b; Mora et al., 2010], (ii) ecology [Harte, 2011], (iii) collective behavior in animal groups [Bialek et al., 2012; Mora et al., 2016], (iv) cell regulatory [Lezon et al., 2006] and signaling networks [Locasale and Wolf-Yadlin, 2009], and (v) theoretical neuroscience [Schneidman et al., 2006; Tkačik et al., 2014; Berry II et al., 2013; Lynn et al., 2025].

Energy-based models themselves belong to a wider class of machine learning techniques, known as generative modeling [Kingma and Welling, 2019; Goodfellow et al., 2014; Rezende and Mohamed, 2015; Sohl-Dickstein et al., 2015; Vaswani et al., 2017; LeCun et al., 2006; Gu and Dao, 2024], which aims to learn a probability distribution of a data-generating process given samples and has been used in scientific research as tools for discovery of principles underlying complex, high-dimensional systems [Bialek and Ranganathan, 2007; Cocco et al., 2018b; Rives et al., 2021b; Bialek, 2012; Mora and Bialek, 2011; De Martino and De Martino, 2018; Papamakarios et al., 2021; Yang et al., 2023].

Furthermore, our framework opens up the possibility to investigate scaling relationships between training sample number, optimal sampling temperature, and the ground truth energy landscape as captured by density of states and the true temperature. The analysis can extend to larger system size and ground truth energy landscapes beyond the nearest-neighbor Ising model. Generic relationships that hold across system size and data-generating distributions could improve training and performance across a wide variety of systems; such relationships could constrain optimal sampling temperatures in lieu of knowledge of the ground truth.

We have shown that temperature tuning for improved generative performance is not only a heuristic sampling technique but also a potential diagnostic tool. Our framework gestures towards a normative framework within which one can understand temperature tuning, finite-sample bias, generative performance, and their relation to the true data-generating process,

allowing for a tool that may further probe learned energy landscapes and their implications for biological systems.

APPENDICES FOR CHAPTER 3

All code for the simulations reported in this article may be found at

`https://github.com/sepalmer/temp-tune`

3.A Derivation of $K_{\mathbf{detect}}$

Let $\{v_i\}$ be i.i.d. random variables with an expectation defined by:

$$\mu = \mathbb{E}[v_i]$$

Let K be a stopping time taken with respect to some stopping criteria. Then by Wald's identity [Grimmett and Stirzaker, 2001] we have

$$\mathbb{E}[\sum_{i=1}^{K} v_i] = \mu \mathbb{E}[K], \tag{3.21}$$

which can be interpreted as meaning the expected cumulative sum of increments up to a stopping time is equal to the expected number of increments times the mean increment.

For an experimenter, person A, being given samples by person B, where B knows the samples come from a distribution $\hat{q}$, if A is testing between two hypotheses H_0 (null—samples come from p) and H_1 (alternative—samples come from $\hat{q}$), then by the Neyman-Pearson lemma [Cover and Thomas, 2006; Csiszár and Shields, 2004] the optimal criteria for determining between the two for a given sample is the log ratio test.

$$\delta L_i = \log \frac{\hat{q}(v_i)}{p(v_i)}. \tag{3.22}$$

We may consider δL_i as the increment in the amount of evidence in favor of a given

hypothesis and also as a random variable itself. Further we note that

$$\mathbb{E}_{\hat{q}}[\delta L_i] = D_{\mathrm{KL}}(\hat{q}||p)$$

Taking $L_K = \sum_i^K \delta L_i$, then by Eq. (3.21) we have

$$\mathbb{E}_{\hat{q}}[L_K] = \mathbb{E}_{\hat{q}}[K] \cdot \mathbb{E}_{\hat{q}}[\delta L_i] = \mathbb{E}_{\hat{q}}[K] \cdot D_{\mathrm{KL}}(\hat{q}||p) \tag{3.23}$$

It then follows that the expected value (according to person B) of the number of samples that person A will take before L_K passes some threshold γ is

$$\mathbb{E}[K_{\mathrm{detect}}] \approx \frac{\gamma}{D_{\mathrm{KL}}(\hat{q}||p)} \tag{3.24}$$

Considering K_{detect} as a function of τ, we can define a ratio representing the increase in number of samples person A can take before they realize the samples are not from p as

$$R = \frac{D_{\mathrm{KL}}(\hat{q}_{\tau=1}||p)}{D_{\mathrm{KL}}(\hat{q}_{\tau^*}||p)}$$

where we note that R is invariant to choice of γ. R is an interesting quantity as it may be considered in experiments where p is not known, but a proxy for it is known, i.e. as in [Russ et al., 2020b] where an experiment is done on synthesized proteins *in vivo* and histograms of functionality of new sequences can be directly compared to histograms of functionality for the training data.

3.B Simple toy model

3.B.1 Description

As described in Section 3.2.2 of the main text, the probability distribution of the simple toy model is given by Eq. (3.4), reproduced here

$$p_i = \frac{\exp\left(-\Delta L_i\right)}{Z(\Delta, \mathbf{L})}, \tag{3.25}$$

$$E_i = \Delta L_i$$

where $\mathbf{L}$ is a vector that assigns each state i to a low- or high-energy level, $L_i \in \{0, 1\}$, Δ is the energy gap between levels, and $Z(\mathbf{L}, \Delta) = \ n_l + \ n_h \exp(-\Delta)$ is the partition function, where n_l and n_h are the number of low- and high-energy states.

The goal of inference is to find best estimates of the level assignment vector, $\hat{\mathbf{L}}$ and energy gap between levels, $\hat{\Delta}$. The objective function to be fit, corresponding to maximum likelihood estimation is

$$\mathcal{L}(\tilde{\Delta}, \tilde{\mathbf{L}}) = \tilde{\Delta}\tilde{\mathbf{L}} \cdot \mathbf{p}_{\mathcal{D}} + \log Z(\tilde{\Delta}, \tilde{\mathbf{L}}), \tag{3.26}$$

where the best estimates of model parameters are

$$\hat{\Delta}, \hat{\mathbf{L}} = \underset{\tilde{\Delta}, \tilde{\mathbf{L}}}{\operatorname{argmin}} \ \mathcal{L}(\tilde{\Delta}, \tilde{\mathbf{L}}), \tag{3.27}$$

and for a given $\tilde{\mathbf{L}}$ we have

$$\tilde{\Delta} = \log \frac{\tilde{n}_h(1 - \tilde{\mathbf{L}} \cdot \mathbf{p}_{\mathcal{D}})}{\tilde{n}_l(\tilde{\mathbf{L}} \cdot \mathbf{p}_{\mathcal{D}})}, \tag{3.28}$$

given by Eqs. (3.7) and (3.8), respectively, in the main text.

The samples over states from Eq. (3.25) gives the empirical distribution, $\mathbf{p}_{\mathcal{D}}$, where $p_{\mathcal{D},i} \in [0, 1]$ is the measured frequency of state i from M samples.

Note that the objective function, Eq. (3.26) also corresponds to a data-approximation of $D_{\mathrm{KL}}(p_i||q_{i|\tilde{\Delta},\tilde{\mathbf{L}}})$ (up to an additive constant that does not affect inference) and is simply $D_{\mathrm{KL}}(\mathbf{p}_{\mathcal{D}}||\mathbf{q}_{\tilde{\Delta},\tilde{\mathbf{L}}})$.

Some properties of $\hat{\mathbf{L}} \cdot \mathbf{p}_{\mathcal{D}}$ are worth noting. If we denote the set of states labeled excited by the fit model as $\hat{e} = \{i_{\hat{e}}\}$ and recall that $\hat{L}_i = 0$ for all ground states, then we can see that

$$\hat{\mathbf{L}} \cdot \mathbf{p}_{\mathcal{D}} = \sum_i \hat{L}_i p_{i,\mathcal{D}} \ = \sum_{i \in \hat{e}} p_{i,\mathcal{D}}. \tag{3.29}$$

The quantity $\tilde{\Delta}\tilde{\mathbf{L}} \cdot \mathbf{p}_{\mathcal{D}}$ may be thought of as the energy of states averaged over the data distribution.

$$\tilde{\Delta}\tilde{\mathbf{L}} \cdot \mathbf{p}_{\mathcal{D}} = \sum_i \tilde{\Delta}\tilde{L}_i p_{i,\mathcal{D}} = \sum_i \tilde{E}_i p_{i,\mathcal{D}} = \langle \tilde{E}_i \rangle_{\mathcal{D}} \tag{3.30}$$

Furthermore, for a well sampled distribution we expect to fit such that true parameters are recovered: $\hat{\mathbf{L}} = \mathbf{L}$, $p_{i,\mathcal{D}} = p_i$ and therefore

$$\lim_{M \to \infty} \hat{\mathbf{L}} \cdot \mathbf{p}_{\mathcal{D}} = \sum_i L_i p_i \ = \sum_{i \in \{i_e\}} p_i = \frac{n_h \exp(-\Delta)}{n_l + n_h \exp(-\Delta)}. \tag{3.31}$$

3.B.2 Fitting procedure

We briefly note here that for under-sampled datasets in this setting it is possible for $\hat{\mathbf{L}} \cdot \mathbf{p}_{\mathrm{data}} = 0$, and therefore, $\mathcal{L}(\hat{\Delta}, \hat{\mathbf{L}}) = \log \frac{\hat{n}_l}{N_s}$ and $\hat{\Delta} \to \infty$. To avoid such a divergence, and to ensure a fit model with support on all states, we introduce a hard-constraint regularization that assumes the probability of seeing any high-energy state under the model is $\approx 1/(M+1)$ if and only if $\hat{\mathbf{L}} \cdot \mathbf{p}_{\mathrm{data}} = 0$.

$$\begin{aligned} \frac{1}{M+1} &= \frac{\hat{n}_h \exp(-\hat{\Delta})}{\hat{n}_l + \hat{n}_h \exp(-\hat{\Delta})} \\ \Rightarrow \hat{\Delta} &= \log(\frac{n_h}{n_l} M) \end{aligned}$$

Algorithm 1 below gives the fitting procedure for finding $\hat{\Delta}$ and $\hat{\mathbf{L}}$ given $\mathbf{p}_{\mathcal{D}}$.

Algorithm 1 Find $\hat{\Delta}, \hat{\mathbf{L}}$

given $\mathbf{p}_{\mathcal{D}} = (p_{\mathcal{D},1}, p_{\mathcal{D},2}, \ldots, p_{\mathcal{D},N})$, init containers:

`losses`=LIST[LENGTH=N] ▷ *ordered list for losses at each training iteration*
`deltas`=LIST[LENGTH=N] ▷ *for energy gap estimates at each iter*
`g`={ } ▷ *collection of indices of states assigned ground energy level. init as empty*
`e`=$\{1, 2, \ldots, N\}$ ▷ *indices of states assigned excited energy level. init all states as excited*
`R`=$\{p_{\mathcal{D},1}, p_{\mathcal{D},2}, \ldots, p_{\mathcal{D},N}\}$ ▷ *collection of empirical frequencies of states*
`all_e`=LIST[LENGTH=N]
`all_g`=LIST[LENGTH=N] ▷ *to store state-assignment collections at each learning step*
`L`=LIST[LENGTH=N]; `L[i]`=1 for `i`=$1, 2, \ldots, N$ ▷ *state assignment vector. init all states as excited*

for `t` in 1 to N **do**
    find current largest $p_{\mathcal{D},i}$ in `R` and assign corresponding state to ground energy level
    $\bar{\mathtt{p}}$ = LARGEST(`R`)
    $\bar{\mathtt{i}}$ = INDEXOF($\bar{\mathtt{p}}$)
    `L`[$\bar{\mathtt{i}}$] = 0
    calculate estimate of energy gap and associated loss
    `deltas[t]`=GETENERGYGAP(`L`, $\mathbf{p}_{\mathcal{D}}$) ▷ *using Eq. (3.28).*
    `losses[t]`=GETLOSS(`deltas[t]`, `L`) ▷ *using Eq. (3.26).*
    record current set of states at each energy level
    `all_e[t]`=`e`\$\bar{\mathtt{i}}$
    `all_g[t]`=`g`$\cup\{\bar{\mathtt{i}}\}$
    remove largest $p_{\mathcal{D},i}$ from `R`
    `R`=`R`\$\bar{\mathtt{p}}$

return `L` and `deltas[t]` corresponding to minimum of `losses`
$\hat{\mathtt{t}}$ = INDEXOFMIN(`losses`)
`L[i]`=1 $\forall$ `i` $\in$ `all_e`[$\hat{\mathtt{t}}$]
`L[i]`=0 $\forall$ `i` $\in$ `all_g`[$\hat{\mathtt{t}}$]
RETURN(`deltas`[$\hat{\mathtt{t}}$],`L`)

We may further simplify fitting the model by reordering the states in $\mathbf{p}_{\text{data}}$, $i \to k$, such

that $p_{\text{data},k} \geq p_{\text{data},k+1}$ for $1 \leq k \leq N_s - 1$, and defining the quantity

$$\begin{aligned}\mathcal{G}(\ell) :&= \sum_{k=1}^{\ell} p_{\text{data},k} \\ &= 1 - \mathbf{L} \cdot \mathbf{p}_{\text{data}},\end{aligned}$$

where we note that $k = 1, 2, ..., \ell$ index the ground states and therefore $\ell = n_g$. Substituting Eq. (3.28) into (3.26) and taking $\tilde{\mathbf{L}} \cdot \mathbf{p}_{\text{data}} \to 1 - \mathcal{G}(\tilde{\ell})$, we may define the objective function

$$\begin{aligned}\mathcal{L}(\tilde{\ell}) = &- (1 - \mathcal{G}(\tilde{\ell})) \log(1 - \mathcal{G}(\tilde{\ell})) - \mathcal{G}(\tilde{\ell}) \log \mathcal{G}(\tilde{\ell}) \\ &+ (1 - \mathcal{G}(\tilde{\ell})) \log(1 - \frac{\tilde{\ell}}{N_s}) + \mathcal{G}(\tilde{\ell}) \log \frac{\tilde{\ell}}{N_s},\end{aligned} \tag{3.32}$$

which is the same as Eq. (3.26) up to a constant that does not depend on $\tilde{\ell}$.

Equation (3.32) is optimized over $\tilde{\ell}$, and $\hat{\ell}$ is used to find $\hat{\Delta}$ and $\hat{\mathbf{L}}$.

3.B.3 Exact expressions for τ^* and τ'

The toy model is tractable such that we may take derivatives with respect to τ of $D_{\text{KL}}(\hat{\mathbf{q}}_\tau || \mathbf{p})$ and $D_{\text{KL}}(\mathbf{p} || \hat{\mathbf{q}}_\tau)$ directly and using this to find $\tau*$ and τ'. The key is break up the sum over all states into 4 separate contributions from each of the possible combinations level assignments per state: (i) $\mathbf{L} = 0, \hat{\mathbf{L}} = 0$ ("found low-energy states"), (ii) $\mathbf{L} = 0, \hat{\mathbf{L}} = 1$ ("missed low"), (iii)$\mathbf{L} = 1, \hat{\mathbf{L}} = 1$ ("found high"), (iv) $\mathbf{L} = 1, \hat{\mathbf{L}} = 1$ ("missed high").

Noting that the number of found high-energy states is $\mathbf{L} \cdot \hat{\mathbf{L}}$ and taking care to break up

the sum over states accordingly we can write

$$\begin{aligned} D_{\mathrm{KL}}(\hat{\mathbf{q}}_\tau||\mathbf{p}) &= \sum_i \hat{q}_{\tau,i} \log \frac{\hat{q}_{\tau,i}}{p_i} \\ &= \log\left(\frac{Z}{\hat{Z}(\tau)}\right) - \sum_i \frac{\exp(-\hat{\Delta}\hat{L}_i/\tau)}{\hat{Z}(\tau)} \left(\frac{\hat{\Delta}\hat{L}_i}{\tau} - \Delta L_i\right) \\ &= \log\left(\frac{Z}{\hat{Z}(\tau)}\right) - \frac{\exp(-\hat{\Delta}\hat{L}_i/\tau)}{\hat{Z}(\tau)} \left[\left(\hat{n}_h - \hat{\mathbf{L}}\cdot\mathbf{L}\right)\frac{\hat{\Delta}}{\tau} + \hat{\mathbf{L}}\cdot\mathbf{L}\left(\frac{\hat{\Delta}}{\tau} - \Delta\right)\right] \\ &\qquad + \frac{1}{\hat{Z}(\tau)}(n_h - \hat{\mathbf{L}}\cdot\mathbf{L})\Delta \end{aligned} \tag{3.33}$$

Finding the stationary point w.r.t. τ yields

$$\tau^* = \frac{\hat{n}_h}{\hat{\mathbf{L}}\cdot\mathbf{L} + \frac{\hat{n}_h}{\hat{n}_l}(\hat{\mathbf{L}}\cdot\mathbf{L} - n_h)} \frac{\hat{\Delta}}{\Delta}. \tag{3.34}$$

Similarly we have

$$\begin{aligned} D_{\mathrm{KL}}(\mathbf{p}||\hat{\mathbf{q}}_\tau) = \log\left(\frac{\hat{Z}(\tau)}{Z}\right) &+ \frac{1}{Z}(\hat{n}_h - \hat{\mathbf{L}}\cdot\mathbf{L})\frac{\hat{\Delta}}{\tau} \\ &- \frac{\exp(-\Delta)}{Z}\left[(n_h - \hat{\mathbf{L}}\cdot\mathbf{L})\Delta - \hat{\mathbf{L}}\cdot\mathbf{L}\left(\frac{\hat{\Delta}}{\tau} - \Delta\right)\right] \end{aligned} \tag{3.35}$$

and

$$\begin{aligned} \tau' &= \frac{\hat{\Delta}}{\Delta - \log x} \\ x &= \frac{\hat{n}_l(\hat{\mathbf{L}}\cdot\mathbf{L} + e^{\Delta}(n_h - \hat{\mathbf{L}}\cdot\mathbf{L} + n_l + \hat{n}_l))}{\hat{n}_h(\hat{n}_l + (n_h - \hat{\mathbf{L}}\cdot\mathbf{L})(e^{-\Delta} - 1))} \end{aligned} \tag{3.36}$$

Equations (3.34) and (3.36) were used to calculate τ^* and τ' each experiment in Fig. 3.2(d-h) and Fig. 3.3(c) and (f).

3.C Ising model distribution

3.C.1 Training from samples

The model is trained until the relative change in the negative log-likelihood goes below 10^{-5}, that is:

$$\frac{|\mathcal{L}^{\{t\}} - \mathcal{L}^{\{t-1\}}|}{\mathcal{L}^{\{t-1\}}} < 10^{-5}$$

where t indexes the training iteration of the gradient descent. Note that gradient updates, which are defined as

$$\tilde{J}_{ij}^{(t+1)} = \tilde{J}_{ij}^{(t)} + \eta \left(\langle v_i v_j \rangle_{\mathcal{D}_T} - \langle v_i v_j \rangle_{\mathcal{M}_{(t)}} \right)$$

with learning rate η, can be calculated exactly since expectation values with respect to the learned model at iteration t, $\langle v_i v_j \rangle_{\mathcal{M}_{(t)}}$, can be taken over all 2^{16} states stored in memory, and does not require a sampling approximation.

3.C.2 Calculating τ^* and τ'

For the values τ^* and τ' reported in Fig. 3.4(e-i), exact expressions are not available as in the case of the illustrative toy model. For each model fit, $\hat{q}$, a vector of all probabilities for all 2^{16} is generated for each value of τ (swept in the interval $[0.2, 5]$) and compared to the corresponding p_T (which generated the training data) via the forward and reversed D_{KL}. Resulting D_{KL} vs. τ plots are fit with a spline function of degree 4, with exact interpolation (0 smoothing), and zero boundary conditions. τ^* and τ' are then calculated from the resulting curves.

In the Fig. 3.4(h) and (i), 10 replicates of experiments are done for each value of M and T, and τ^*, τ' are calculated as above, then averaged. The scaled color image represents the mean over these 10 replicates.

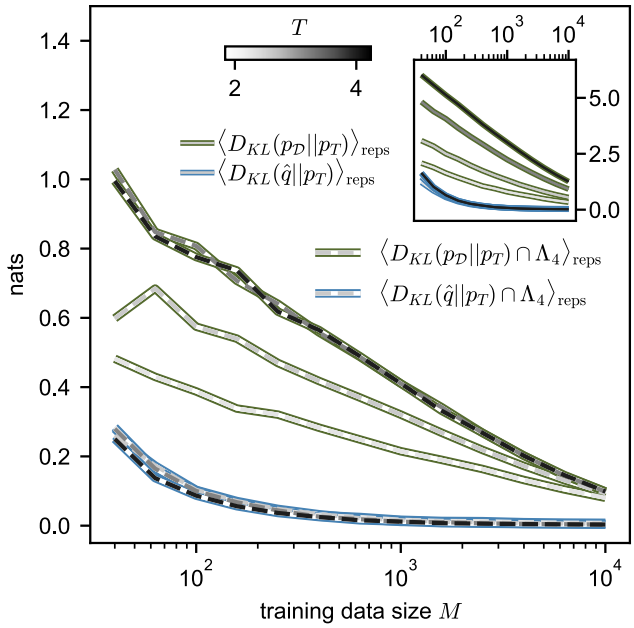

Figure 3.7: Mean values of $D_{\text{KL}}(p_\mathcal{D}||p_T) \cap \Lambda_4$ and $D_{\text{KL}}(\hat{q}||p_T) \cap \Lambda_4$ over 50 replicates of experiments for several values of M and T. The value of T is denoted by the shade of black of each line. Both D_{KL}'s decrease as M increases, though $D_{\text{KL}}(p_\mathcal{D}||p_T)\cap\Lambda_4 > D_{\text{KL}}(\hat{q}||p_T)\cap\Lambda_4$ always. (Inset) This bound is obeyed when considering the total value of each D_{KL}. Means with respect to replicates are calculated as in Fig. 3.6.

3.D Upper bounds on $D_{\text{KL}}(\hat{q}||p)$

$D_{\text{KL}}(\hat{q}||p_T)$ is bounded from above by $D_{\text{KL}}(p_\mathcal{D}||p_T)$ per-energy-level, which makes intuitive sense—we would expect are best estimate of the distribution to be no worse than the data itself.

In information geometry, a standard identity (Pythagorean relation for the KL divergence) relates the maximum likelihood estimate of a model to data taken from a distribution, (that is, of $\hat{q}$ to $p_\mathcal{D}$ taken from p) [Csiszár and Shields, 2004].

$$D_{\text{KL}}(p_\mathcal{D}||p) = D_{\text{KL}}(p_\mathcal{D}||\hat{q}) + D_{\text{KL}}(\hat{q}||p). \tag{3.37}$$

This is true when $\hat{q}$ is in the exponential family of distributions and the support of $p_\mathcal{D}$ and $\hat{q}$ are the same. The same support is shared by both distributions only when $p_\mathcal{D}$ is

well-sampled, that is, every state is sampled at least once. This is especially rare for low M and low T, as many states are not sampled. However, Eq. (3.37), implies the following bound,

$$D_{\mathrm{KL}}(p_{\mathcal{D}}||p) > D_{\mathrm{KL}}(\hat{q}||p), \tag{3.38}$$

and furthermore, we expect this bound to hold in the limiting behavior, $M \to \infty$. We find empirically that this bound is mostly obeyed on a per-energy-level basis, on average, for low M experiments. In Fig. 3.7, many such experiments are conducted for various values of T and M on the 4×4 Ising model. We see that for energy level 4, $D_{\mathrm{KL}}(p_{\mathcal{D}}||p_T) \cap \Lambda_4 > D_{\mathrm{KL}}(\hat{q}||p_T) \cap \Lambda_4$, even as M is quite low, and for various values of T.

CHAPTER 4

FUTURE DIRECTIONS

As we saw in Chapter 3, the temperature tuning phenomena is not only a heuristic technique for improving the performance of generative models; it arises from specific properties of the ground truth distribution and its interplay with the limited amount of training data and choice of fitting protocol. Since all these considerations are intermingled, might our understanding of the temperature tuning phenomenology be leveraged for practical application for fitting real datasets? Furthermore, to what extent may this phenomenology hold (and alter) across different systems, i.e. ones of higher dimensions, different energy landscapes (ground truth T and density of states g), and so on?

This chapter will propose two future directions of research that explore these aforementioned questions. The first direction will explore the scaling relation of optimal sampling temperature τ^* when considered as a function of the ground truth temperature T, training data amount M, and density of states g, that is, $\tau^* = f(M, T, g)$. Some preliminary observations will be made regarding the simpler case of the $\tau^* = f(M, T)$ for the 4×4 nearest-neighbor Ising model.

The second will leverage an interesting relationship between the heat capacity of the fit model and the sampling temperature τ—specifically, that whether τ need be raised or lowered to improve performance (under certain conditions) the resulting fit model heat capacity has decreased (almost always). This is interesting because without the ground truth, one may not know which direction to tune temperature *a priori*, but choosing that which decreases heat capacity (and even incorporating this into the fitting procedure) may increase generative performance, all without knowing the true nature of the data-generating process.

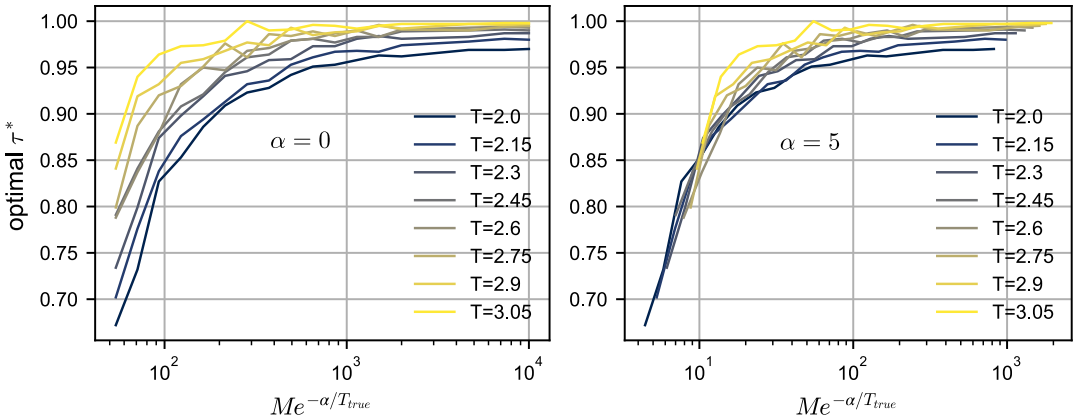

Figure 4.1: Mean optimal sampling temperature τ^* versus rescaled training sample number, M. (Means are the same as depicted in Fig. 3.4—an average taken over many replicates of experiments at a given T and M).

4.1 Scaling relations between sampling temperature, number of training data, and ground truth T

In Fig. 4.1 the mean optimal sampling temperature τ^* (the same as depicted in Fig. 3.4—an average taken over many replicates of experiments at a given T and M) is plotted versus a rescaled number of training samples, $Me^{-\alpha/T}$. Going from the left panel to the right, we see that changing α from 0 to 5 causes near collapse of curves onto each other. This is indicative of a possible scaling relation between optimal sampling temperature τ^*, training data amount M, and ground truth T. Such a scaling relation would not only be interesting in and of itself, but could also be leveraged for training purposes, as knowing M and the proper scaling relation constrains the possible values of τ^* and T—both quantities of interest for learning and generating.

How may we proceed finding such a scaling relation? Some preliminary observations shall be made here. The first is taken from Appendix 3.D of Ch. 3—Eq. (3.37) reproduced below:

$$D_{\mathrm{KL}}(p_{\mathcal{D}}||p) = D_{\mathrm{KL}}(p_{\mathcal{D}}||\hat{q}) + D_{\mathrm{KL}}(\hat{q}||p), \tag{4.1}$$

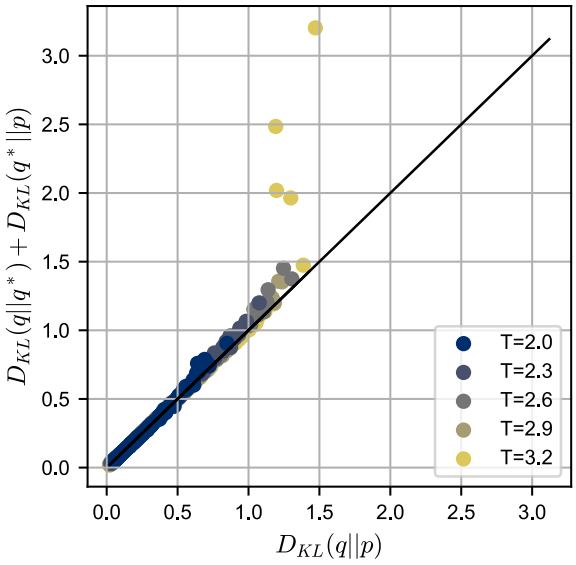

Figure 4.2: Depiction of the bound of Eq. (4.2) obeyed by several experiments on the 4×4 nearest-neighbor Ising model. Each dot is an experiment done at a given T (denoted by color) and M ranging from 54 to 10,000 (not indicated on graph).

where we recall $p_{\mathcal{D}}$ is the empirical distribution defined by the training data taken from p and used to train the maximum likelihood estimate $\hat{q}$. Note that the left hand side of (4.1) depends implicitly on M via $p_{\mathcal{D}}$; for example, the minimal value a state from $p_{\mathcal{D}}$ can take is $1/M$.

Secondly, we note that from Fig. 4.2, for experiments done on the 4×4 nearest-neighbor Ising distribution at several values of T and M, that empirically the following bound is nearly always held:

$$D_{KL}(\hat{q}_{\tau=1}||p_T) \lesssim D_{KL}(\hat{q}_{\tau=1}||\hat{q}_{\tau^*}) + D_{KL}(\hat{q}_{\tau^*}||p_T). \tag{4.2}$$

Many of the models fit to data taken at low T from the Ising distribution saturate the bound. Combining (4.2) and (4.1) and assuming training samples taken from a low-T distribution we obtain

$$D_{KL}(\hat{q}_{\tau=1}||\hat{q}_{\tau^*}) \approx D_{\mathrm{KL}}(p_{\mathcal{D}}||p_T) - D_{\mathrm{KL}}(p_{\mathcal{D}}||\hat{q}) - D_{KL}(\hat{q}_{\tau^*}||p_T), \tag{4.3}$$

where we note that the LHS of the above equation has no explicit dependence on T or M and the last term on the RHS is exactly the quantity we used throughout Ch. 3 to characterize optimal generative performance.

Additionally, as $\tau^* \to 1$ for experiments where M is large enough that the model is fit well enough to no longer require *post hoc* tuning, $D_{KL}(\hat{q}_{\tau=1}||\hat{q}_{\tau^*})$ in the LHS of Eq. (4.3) can be approximated via the heat capacity:

$$\begin{aligned} D_{KL}(\hat{q}_{\tau=1}||\hat{q}_{\tau^*}) &\approx D_{KL}(\hat{q}_{\tau=1}||\hat{q}_{\bar{\tau}=1}) + \frac{\partial}{\partial \bar{\tau}} \left[D_{KL}(\hat{q}_{\tau=1}||\hat{q}_{\bar{\tau}})\right]_{\bar{\tau}=1} (\tau^* - 1)^2 \\ &= \frac{1}{2} C_{\tau=1} (\tau^* - 1)^2 \end{aligned} \tag{4.4}$$

Equations (4.3) and (4.4) may serve as a starting point that may yield the desired relation $\tau^* = f(T, M)$. Furthermore, such study may be extended to system sizes beyond the 4×4 model and to ground truth distributions with structure different than the nearest-neighbor Ising model, which would have inherently different density of states for their energy landscapes $g(E)$.

4.2 Relationship between model heat capacity and sampling temperature

In Figure 4.3, each panel is the change in heat capacity from the trained model at $\tau = 1$ to τ^* plotted against the corresponding change in C, for experiments done on the 4×4 Ising model of Chapter 3. Each point in the scatter plot corresponds to one experiment, as done according to Fig. 3.4, with T indicated by the color and M indicated at the top of each panel.

Note that whether or not $\tau^* - 1$ is positive or negative, the resulting change in heat capacity is nearly always negative. We note, however, that this effect holds only in the low and high T limits, as only experiments on the interval $[2, 2.4] \cup [3.6, 5]$ are shown. Intermediate

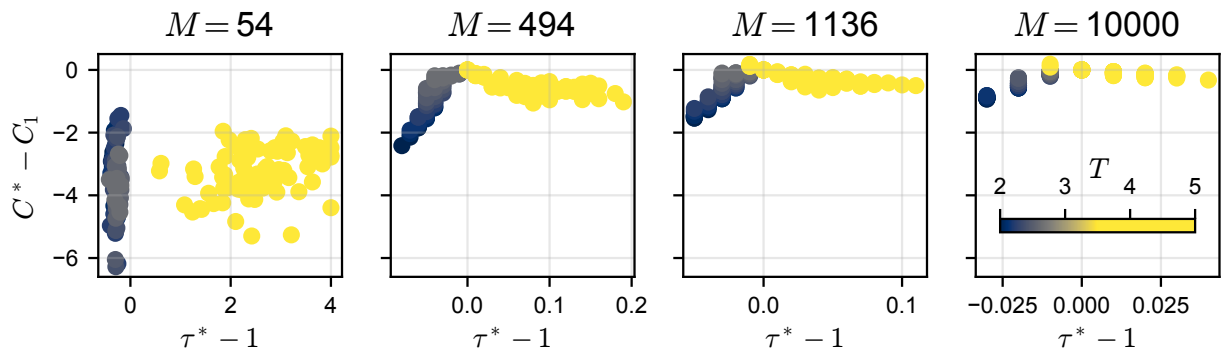

Figure 4.3: The change in heat capacity, $C^* - C_{\tau=1}$, of fit models from $\tau = 1$ to optimal sampling temperature τ^*, as defined in Ch. 3. Each dot in each panel represents 1 experiment done at temperature indicated by the color and training sample number by the panel heading. Experiments are shown for samples taken from the 4×4 nearest-neighbor Ising distribution as used in Ch. 3. Only experiments for $T \in [2, 2.4] \cup [3.6, 5]$ are shown. Note that for these low- and high-T experiments the heat capacity always decreases, whether or not τ is raised or lowered.

temperatures do not exhibit this always-lower-C effect.

We recall that we are particularly interested in the low-T regime, because this regime coupled with under-sampling leads to over-estimation of excited states. And furthermore we expect such regimes to be exemplary of many real-world data sets where we are both under-sampled and we know that meaningful state space is far smaller than meaningless state space. And so, given the observation that τ ought be lowered in this regime, and with regularity, heat capacity along with it, might we be able to leverage this fact to improve generative performance?

An interesting direction to take would be to regularize the variance of the learned energy function $\mathrm{Var}(E_{\mathbf{J}})$ and control the strength of this regularization via a hyperparameter $1/\mathscr{T}^2$, where we identify $\mathscr{T}$ as a *training* temperature. The full objective function to be minimized would be given by:

$$\begin{aligned} -\mathcal{L}'(\mathcal{D}|\mathbf{J}) &= -\mathcal{L}(\mathcal{D}|\mathbf{J}) + \frac{1}{\mathscr{T}^2}\mathrm{Var}(E_{\mathbf{J}}) \\ &= \langle E(v|\mathbf{J})\rangle_{v\sim\mathcal{D}} + \log Z(\mathbf{J}) + \frac{1}{\mathscr{T}^2}\Big\langle \big(E(v|\mathbf{J}) - \langle E(v|\mathbf{J})\rangle_{\mathcal{M}}\big)^2 \Big\rangle_{\mathcal{M}} \end{aligned} \tag{4.5}$$

where $\langle\cdot\rangle_{\mathcal{M}}$ represents an expectation taken with respect to the model defined by $E(v|\mathbf{J})$, and $\mathcal{D}$ is the training data, and furthermore, the regularization term has the same functional form as the heat capacity.

Gradients with respect to the model expectations may not, in general, be easy to take, but a useful simplification may be to introduce a data approximation to the variance:

$$\begin{aligned} -\mathcal{L}' \approx -\mathcal{L}'' &= \langle E(v|\mathbf{J})\rangle_{v\sim\mathcal{D}} + \log Z(\mathbf{J}) + \frac{1}{\mathscr{T}^2}\Big\langle \big(E(v|\mathbf{J}) - \langle E(v|\mathbf{J})\rangle_{v\sim\mathcal{D}}\big)^2 \Big\rangle_{v\sim\mathcal{D}} \\ &= -\mathcal{L} + \mathrm{Reg}(\mathcal{D}, \mathbf{J}, \mathscr{T}), \end{aligned} \tag{4.6}$$

where we identify the first two terms of the RHS of the first line with $-\mathcal{L}$ and the last term with $\mathrm{Reg}(\mathcal{D}, \mathbf{J}, \mathscr{T})$.

4.2.1 Investigating the relationship of $\mathscr{T}$-regularized energy variance and optimal sampling temperature τ^*

The introduction of $\mathscr{T}$ into the learning objective motivates an investigation into its relationship with our previous use of the *post-training* temperature parameter, τ. Can a proper choice of $\mathscr{T}$ give the desired reduction in $D_{\mathrm{KL}}(\hat{q}||p_T)$ without requiring the post-hoc tuning? (Note that $\mathscr{T}$ is not introduced into the model itself; $\mathbf{J}$ does *not* become $\mathbf{J}/\mathscr{T}$ in the fit distribution itself).

A typical workflow may be:

1. Take M training data from ground truth distribution at temperature T.
2. Fit the data using Eq. (4.6) for several choices of $\mathscr{T}$. Call each model fit at training temperature $\mathscr{T}$ as $\hat{\mathbf{J}}_{\mathscr{T}}$.
3. For each value of $\hat{\mathbf{J}}_{\mathscr{T}}$ find the corresponding value of τ^* and $D_{\mathrm{KL}}(\hat{q}_{\tau^*}||p_T)$ and generate plots of each versus $\mathscr{T}$.

4. Repeat for several replicates of training data for each ground truth T.

It would also be useful to check other metrics of performance, such as the error in the fit parameters, defined as

$$\varepsilon = \sum_{i<j} \frac{\left| (\hat{J}_{\mathscr{T},ij} - \frac{J_{\text{true},ij}}{T}) \right|}{\frac{J_{\text{true},ij}}{T}}. \tag{4.7}$$

One may then see if there are optimal training temperature values, $\mathscr{T}^*$, under metrics such as this and $D_{\text{KL}}(\hat{q}||p_T)$.

It would be particularly interesting to see if any behaviors of the objective function itself, Eq. (4.6), correlate with $\mathscr{T}$. A first idea may be to sweep over $\mathscr{T}$ for many different training runs of one dataset, and track the final values of the objective function, $-\hat{\mathcal{L}}''$, as well as the unaugmented negative-log-likelihood, $-\hat{\mathcal{L}}$ and the regularization term, $\text{Reg}(\mathcal{D}, \hat{\mathbf{J}}_{\mathscr{T}}, \mathscr{T})$. If any minima in these quantities exist, could they correlate with $\mathscr{T}^*$? Unfortunately, most likely no; in the limit of $\mathscr{T} \to \infty$, we have $\text{Reg}(\mathcal{D}, \hat{\mathbf{J}}_{\mathscr{T}}, \mathscr{T}) \to 0$ and $-\hat{\mathcal{L}}'' \to -\hat{\mathcal{L}}$. That is, learning will simply return the maximum likelihood values for $\mathbf{J}$. As $\mathscr{T}$ decreases, each relearned value of $-\hat{\mathcal{L}}$ will be bounded from below by $-\hat{\mathcal{L}}_{\text{MLE}}$. One should find that $-\hat{\mathcal{L}}$ should be non-decreasing as $\mathscr{T}$ decreases as well, as smaller $\mathscr{T}$ suppresses the magnitudes of each $\hat{J}_{ij}$ leading to underfitting of the training set and larger $-\hat{\mathcal{L}}$.

Similarly, as $\mathscr{T}$ decreases one would also expect $\text{Var}(\text{E}_{\mathbf{J}})_{\mathcal{D}}$ to decrease monotonically as the regularization on the variance strengthens. One should check if $\text{Reg}(\mathcal{D}, \hat{\mathbf{J}}_{\mathscr{T}}, \mathscr{T})$ decreases with decreasing $\mathscr{T}$ as well, though the factor of $1/\mathscr{T}^2$ would slow this decrease.

Though the behaviors of $\text{Reg}(\mathcal{D}, \hat{\mathbf{J}}_{\mathscr{T}}, \mathscr{T})$, $-\hat{\mathcal{L}}$, and $-\hat{\mathcal{L}}''$ may be uncorrelated with any interesting properties of $\hat{\mathbf{J}}_{\mathscr{T}}$ and its relations to τ^* and generative performance, it may be the case that these quantities give interesting proxies for generative performance when evaluated on a held-out test set of the data. It is well-known in learning theory that the test-loss is typically U-shaped as hyper-parameters for regularization are tuned [Abu-Mostafa et al., 2012]. It would be especially interesting to see if the full objective, $-\hat{\mathcal{L}}''_{\text{test}} = -\hat{\mathcal{L}}_{\text{test}} +$

$\mathrm{Reg}(\mathcal{D}_{\mathrm{test}}, \hat{\mathbf{J}}_{\mathscr{T}}, \mathscr{T})$, when evaluated for models trained for different $\mathscr{T}$, has a minimum at a value of $\mathscr{T}$ that corresponds well with generative performance as measured by $D_{\mathrm{KL}}(\hat{q}||p_T)$, as each term in $-\hat{\mathcal{L}}''$ captures the trade-off that the aforementioned D_{KL} tries to capture; that is, $-\hat{\mathcal{L}}_{\mathrm{test}}$ captures fidelity to the ground truth distribution and $\mathrm{Reg}(\mathcal{D}_{\mathrm{test}}, \hat{\mathbf{J}}_{\mathscr{T}}, \mathscr{T})$ is proportional to the variance of the learned energies, representing a notion of the spread (or diversity) of the states representative of the learned model.

One should also check how learned energy gaps behave. (One may track this by tracking the difference in the maximum and minimum of energies on the training data or generated samples). On the one hand, one may expect energy gaps to increase for models learned on training data taken from "cold" or highly ordered ground truths. Decreasing the variance, and heat capacity with it, should push the model into the ordered phase. On the other hand, one could expect that energy gaps decrease as the spread in energies is decreased, leading to a flatter landscape.

It is worth noting that other approaches to utilizing a training temperature $\mathscr{T}$ may be employed. The heat capacity of a model is proportional to the Fisher information of the temperature parameter. Including $\mathscr{T}$ as an explicitly learned parameter may be done via Firth's regularization scheme [Firth, 1993] or via empirical Bayesian techniques with a hyper-prior on the temperature itself, such as a Jeffreys prior on $\mathscr{T}$ that gives it a uniform distribution on $\log \mathscr{T}$. (In these cases, it may be advantageous to include $\mathscr{T}$ as a parameter in the learned distribution itself, given that one is learning an overall scale parameter on the energy function itself). The Fisher information of the temperature parameter may also be utilized to create an optimal noise distribution from which one may draw samples if they are fitting with noise contrastive estimation; see [Chehab et al., 2022, 2023] for more details.

4.2.2 Applications for protein-like distributions

We present in this section arguments that may indicate the favorability of this regularization term to promote learning a J_{ij} structure indicative of low-fluctuation and uncorrelated motifs of subsets of amino acids within the larger protein sequence.

Let us define P patterns as ξ_μ, of length equal to that of the sequences we have been considering, v, and define $J = WW^\top$, where the columns of W are each of the P patterns. We can further simplify our energy function:

$$-E(v) = \frac{1}{2}\sum_{i,j} J_{ij} v_i v_j = \frac{1}{2}\sum_{\mu} (\xi_\mu^\top v)^2 = \frac{1}{2} v^\top W W^\top v. \tag{4.8}$$

We then define

$$y_m = W^\top v^{(m)} \tag{4.9}$$

for the m^{th} training sample. If we assume that all components of v are zero-mean and Gaussian distributed, then so must all components of y. (This assumption is reasonable if v are generated from a ground truth at high T and 0 field strength). Furthermore, y will have a covariance defined as $\sigma = W^\top \widehat{C} W$, where $\widehat{C} = \frac{1}{M}\sum_m v^{(m)} {v^{(m)}}^\top$ is the data covariance matrix.

We can then write our regularization term as:

$$\text{Reg}(\mathcal{D}, \mathbf{J}, \mathscr{T}) = \frac{1}{4\mathscr{T}^2}\left[\frac{1}{M}\sum_m (y_m^\top y_m)^2 - \left(\frac{1}{M}\sum_m y_m^\top y_m\right)^2\right]. \tag{4.10}$$

Since we have assumed y to be Gaussian distributed we can approximate the sample expectation with an expectation over its respective Gaussian:

$$\sum_m (y_m^\top y_m)^2 \approx M\ \mathbb{E}[(y^\top y)^2]. \tag{4.11}$$

By Wick's probability theorem we can write

$$\begin{aligned}\mathbb{E}[(y^\top y)^2] &= \mathbb{E}\left[\sum_{\mu,\nu} y_\mu y_\mu y_\nu y_\nu\right] = \sum_{\mu,\nu} \mathbb{E}\left[y_\mu y_\mu y_\nu y_\nu\right] \\ &= \sum_{\mu,\nu} (\sigma_{\mu\mu}\sigma_{\nu\nu} + 2\sigma_{\mu\nu}^2) \\ &= \left(\sum_{\mu} \sigma_{\mu\mu}\right)^2 + 2\sum_{\mu,\nu} \sigma_{\mu\nu}^2.\end{aligned} \tag{4.12}$$

Similarly we have:

$$\begin{aligned}\left(\frac{1}{M}\sum_m y_m^\top y_m\right)^2 \approx\ (\mathbb{E}[y^\top y])^2 =\ \left(\sum_\mu \mathbb{E}[y_\mu y_\mu]\right)^2 \\ = \left(\sum_\mu \sigma_{\mu\mu}\right)^2.\end{aligned} \tag{4.13}$$

Plugging the above into (4.10) yields:

$$\mathrm{Reg}(\mathcal{D}, \mathbf{J}, \mathscr{T}) \approx \frac{1}{2\mathscr{T}^2}\sum_{\mu,\nu}(W^\top \widehat{C} W)_{\mu\nu}^2 \tag{4.14}$$

One may think of this regularization term as penalizing the norms of each pattern (and dot products between patterns, $\widehat{C}_{\mu\nu}\xi^\mu\xi^\nu$) in an inner product space defined by the training data-covariance matrix $\widehat{C}$. That is, the regularization term favors patterns that are uncorrelated and low variance when viewed from a space defined by the data covariance.

This may be interesting for applications in protein distributions where the function-determining set of amino acids are known to be comprised of a small subset of the total amino acids in the sequence, known as the sector; and when multiple sectors exist in a protein, they tend to be mutually exclusive from each other (as discussed in Chapter 2 and [Halabi et al., 2009b; Rivoire et al., 2016b; Salinas and Ranganathan, 2018b]).

This regularization term may help with fitting energy-based models and discovering these

low-fluctuating motifs within protein sequence datasets. It should be noted that W is ill-defined up to an orthogonal rotation matrix U. That is, if we have $V = WU$, we will have $J = VV^\top = WUU^\top W^\top = WW^\top$. To determine the best choice of rotation matrix for interpretability of the patterns of W, one may consider using ICA techniques similar to those used in [Rivoire et al., 2016b].

Kleeorin et al. [2023b] have noted that typical $L2$ regularization schemes employed for pairwise models in proteins introduces an inductive bias (when undersampled) that causes the fit model to highlight strong couplings for isolated pairwise sites, and attenuates couplings of larger interaction networks among positions. This biases the model away from learning the interaction structure of the sector—a potential pathology for generative performance.

The regularization of Eq. (4.14) may side step this issue; so long as it is able to learn orthogonal patterns, one may be able to strengthen and attenuate each pattern as one chooses, and then compare states from these pattern-selective rescaled models to *in vivo* experiments or to previous sector analysis. One may also consider using the covariance matrix used in sequence coupling analysis (SCA) in Eq. (4.14)—a reweighted version of the data-covariance matrix used to discover sectors [Rivoire et al., 2016b].

Furthermore, we have seen the post-hoc temperature tuning for sampling performance is a *homogeneous* rescaling of all interaction terms, which we have noted above as being *inhomogeneously* inferred in low sample number regimes. What our experiments on post-hoc rescaling of temperature have indicated is the potential use of heat capacity as a proxy for generative performance; and regularizing the variance of the energy function with a temperature hyperparameter $\mathscr{T}$ leads to a regularization term that could possibly re-weight the learned patterns defined by J_{ij}, and do so *inhomogeneously* via a metric defined by the data-covariance $\widehat{C}$. Proof-of-concept experiments may be performed on the protein-like minimal ground truth models of [Kleeorin et al., 2023a] and Ch. 2.

REFERENCES

Y.S. Abu-Mostafa, M. Magdon-Ismail, and H.T. Lin. *Learning from Data: A Short Course*. AMLBook.com, 2012. ISBN 978-1-60049-006-4. URL https://books.google.com/books?id=iZUzMwEACAAJ.

Pierre Barrat-Charlaix, Anna Paola Muntoni, Kai Shimagaki, Martin Weigt, and Francesco Zamponi. Sparse generative modeling via parameter reduction of boltzmann machines: application to protein-sequence families. *Physical Review E*, 104(2):024407, 2021.

Shane Barratt and Rishi Sharma. A Note on the Inception Score, June 2018. URL http://arxiv.org/abs/1801.01973. arXiv:1801.01973 [stat].

Michael J Berry II, Gašper Tkačik, Julien Dubuis, Olivier Marre, and Rava Azeredo da Silveira. A simple method for estimating the entropy of neural activity. *Journal of Statistical Mechanics: Theory and Experiment*, 2013(03):P03015, March 2013. ISSN 1742-5468. doi:10.1088/1742-5468/2013/03/P03015. URL https://doi.org/10.1088/1742-5468/2013/03/P03015. Publisher: IOP Publishing and SISSA.

William Bialek. *Biophysics: Searching for Principles*. Princeton University Press, Princeton, NJ, 2012.

William Bialek and Rama Ranganathan. Rediscovering the power of pairwise interactions, December 2007. URL http://arxiv.org/abs/0712.4397. arXiv:0712.4397 [q-bio].

William Bialek, Andrea Cavagna, Irene Giardina, Thierry Mora, Edmondo Silvestri, Massimiliano Viale, and Aleksandra M. Walczak. Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, 109(13):4786–4791, March 2012. doi:10.1073/pnas.1118633109. URL https://www.pnas.org/doi/10.1073/pnas.1118633109. Publisher: Proceedings of the National Academy of Sciences.

Surojit Biswas, Grigory Khimulya, Ethan C. Alley, Kevin M. Esvelt, and George M. Church. Low-n protein engineering with data-efficient deep learning. *Nature Methods*, 18(4):389–396, 2021. doi:10.1038/s41592-021-01100-y.

Sam Bond-Taylor, Adam Leach, Yang Long, and Chris G. Willcocks. Deep Generative Modelling: A Comparative Review of VAEs, GANs, Normalizing Flows, Energy-Based and Autoregressive Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44:7327–7347, 2021. URL https://api.semanticscholar.org/CorpusID:232147810.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum Dependent Learning Curves in Kernel Regression and Wide Neural Networks. In *Proceedings of the 37th International Conference on Machine Learning*, pages 1024–1034. PMLR, November 2020. URL https://proceedings.mlr.press/v119/bordelon20a.html. ISSN: 2640-3498.

Barbara Bravi, Riccardo Ravasio, Carolina Brito, and Matthieu Wyart. Direct coupling analysis of epistasis in allosteric materials. *PLoS computational biology*, 16(3):e1007630, 2020.

Omar Chehab, Alexandre Gramfort, and Aapo Hyvärinen. The optimal noise in noise-contrastive learning is not what you think. In *Proceedings of the Thirty-Eighth Conference on Uncertainty in Artificial Intelligence*, pages 307–316. PMLR, August 2022. URL https://proceedings.mlr.press/v180/chehab22a.html. ISSN: 2640-3498.

Omar Chehab, Alexandre Gramfort, and Aapo Hyvarinen. Optimizing the Noise in Self-Supervised Learning: from Importance Sampling to Noise-Contrastive Estimation, January 2023. URL http://arxiv.org/abs/2301.09696. arXiv:2301.09696 [stat].

Simona Cocco, Remi Monasson, and Vitor Sessak. High-dimensional inference with the generalized hopfield model: Principal component analysis and corrections. *Physical Review E*, 83(5):051123, 2011.

Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Rémi Monasson, and Martin Weigt. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics*, 81(3):032601, 2018a.

Simona Cocco, Christoph Feinauer, Matteo Figliuzzi, Rémi Monasson, and Martin Weigt. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics. Physical Society (Great Britain)*, 81(3):032601, March 2018b. ISSN 1361-6633. doi:10.1088/1361-6633/aa9965.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, USA, 2006. ISBN 0-471-24195-4.

Imre Csiszár and Paul Shields. *Information Theory and Statistics: A Tutorial*. Foundations and Trends in Communications and Information Theory, 2004. doi:10.1561/0100000004.

Andrea De Martino and Daniele De Martino. An introduction to the maximum entropy approach and its application to inference problems in biology. *Heliyon*, 4(4):e00596, April 2018. ISSN 2405-8440. doi:10.1016/j.heliyon.2018.e00596.

Peter William Fields, Vudtiwat Ngampruetikorn, Rama Ranganathan, David J. Schwab, and Stephanie Palmer. Understanding Energy-Based Modeling of Proteins via an Empirically Motivated Minimal Ground Truth Model. July 2023. URL https://openreview.net/forum?id=vxn5QGPFyi.

David Firth. Bias reduction of maximum likelihood estimates. *Biometrika*, 80(1):27–38, March 1993. ISSN 0006-3444. doi:10.1093/biomet/80.1.27. URL https://doi.org/10.1093/biomet/80.1.27.

Douglas M Fowler and Stanley Fields. Deep mutational scanning: a new style of protein science. *Nature Methods*, 11(8):801–807, 2014. doi:10.1038/nmeth.3027.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. In *Advances in*

Neural Information Processing Systems, volume 27. Curran Associates, Inc., 2014. URL https://papers.nips.cc/paper_files/paper/2014/hash/f033ed80deb0234979a61f95710dbe25-Abstract.html.

G. Grimmett and D. Stirzaker. *Probability and Random Processes*. Probability and Random Processes. OUP Oxford, 2001. ISBN 978-0-19-857222-0. URL https://books.google.com/books?id=G3ig-0M4wSIC.

Albert Gu and Tri Dao. Mamba: Linear-Time Sequence Modeling with Selective State Spaces, May 2024. URL http://arxiv.org/abs/2312.00752. arXiv:2312.00752 [cs].

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On Calibration of Modern Neural Networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 1321–1330. PMLR, July 2017. URL https://proceedings.mlr.press/v70/guo17a.html. ISSN: 2640-3498.

Ulrike Göbel, Chris Sander, Reinhard Schneider, and Alfonso Valencia. Correlated mutations and residue contacts in proteins. *Proteins: Structure, Function, and Bioinformatics*, 18 (4):309–317, 1994. doi:https://doi.org/10.1002/prot.340180402.

Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, 2009a.

Najeeb Halabi, Olivier Rivoire, Stanislas Leibler, and Rama Ranganathan. Protein sectors: evolutionary units of three-dimensional structure. *Cell*, 138(4):774–786, 2009b. Publisher: Elsevier.

John Harte. *Maximum Entropy and Ecology: A Theory of Abundance, Distribution, and Energetics*. Oxford University Press, Oxford, UK, 2011.

Alex Hawkins-Hooker, Florence Depardieu, Sebastien Baur, Guillaume Couairon, Arthur Chen, and David Bikard. Generating functional protein variants with variational autoencoders. *PLOS Computational Biology*, 17(2):e1008736, 02 2021. doi:10.1371/journal.pcbi.1008736.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network, March 2015. URL http://arxiv.org/abs/1503.02531. arXiv:1503.02531 [stat].

Ferenc Huszár. How (not) to Train your Generative Model: Scheduled Sampling, Likelihood, Adversary?, November 2015. URL http://arxiv.org/abs/1511.05101. arXiv:1511.05101 [stat].

Aapo Hyvärinen. Some extensions of score matching. *Computational statistics & data analysis*, 51(5):2499–2512, 2007.

Yuichi Ishida, Yuma Ichikawa, Aki Dote, Toshiyuki Miyazawa, and Koji Hukushima. Ratio divergence learning using target energy in restricted Boltzmann machines: Beyond

Kullback-Leibler divergence learning. *Physical Review E*, 112(4):045306, October 2025. doi:10.1103/fxnm-y5pd. URL https://link.aps.org/doi/10.1103/fxnm-y5pd. Publisher: American Physical Society.

E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review*, 106(4):620–630, May 1957. doi:10.1103/PhysRev.106.620. URL https://link.aps.org/doi/10.1103/PhysRev.106.620. Publisher: American Physical Society.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

Diederik P. Kingma and Max Welling. An Introduction to Variational Autoencoders. *Foundations and Trends® in Machine Learning*, 12(4):307–392, November 2019. ISSN 1935-8237, 1935-8245. doi:10.1561/2200000056. URL https://www.nowpublishers.com/article/Details/MAL-056. Publisher: Now Publishers, Inc.

Yaakov Kleeorin, William P Russ, Olivier Rivoire, and Rama Ranganathan. Undersampling and the inference of coevolution in proteins. *Cell Systems*, 14(3):210–219, 2023a.

Yaakov Kleeorin, William P. Russ, Olivier Rivoire, and Rama Ranganathan. Undersampling and the inference of coevolution in proteins. *Cell Systems*, 14(3):210–219.e7, March 2023b. ISSN 2405-4720. doi:10.1016/j.cels.2022.12.013.

Maximilian B. Kloucek, Thomas Machon, Shogo Kajimura, C. Patrick Royall, Naoki Masuda, and Francesco Turci. Biases in inverse Ising estimates of near-critical behavior. *Phys. Rev. E*, 108(1):014109, July 2023. doi:10.1103/PhysRevE.108.014109. URL https://link.aps.org/doi/10.1103/PhysRevE.108.014109. Publisher: American Physical Society.

Yann Lecun, Sumit Chopra, Raia Hadsell, Marc Aurelio Ranzato, and Fu Jie Huang. *A tutorial on energy-based learning*. MIT Press, 2006.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fu Jie Huang. A Tutorial on Energy-Based Learning. *Predicting structured data*, 2006.

Timothy R. Lezon, Jayanth R. Banavar, Marek Cieplak, Amos Maritan, and Nina V. Fedoroff. Using the principle of entropy maximization to infer genetic interaction networks from gene expression patterns. *Proceedings of the National Academy of Sciences*, 103(50):19033–19038, December 2006. doi:10.1073/pnas.0609152103. URL https://www.pnas.org/doi/10.1073/pnas.0609152103. Publisher: Proceedings of the National Academy of Sciences.

Xinran Lian, Niksa Praljak, Andrew L Ferguson, and Rama Ranganathan. Deep-learning generative models enable design of synthetic orthologs of a signaling protein. *Biophysical Journal*, 122(3):311a, 2023.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637): 1123–1130, 2023. doi:10.1126/science.ade2574.

R. Lipsh-Sokolik, O. Khersonsky, S. P. Schröder, C. de Boer, S.-Y. Hoch, G. J. Davies, H. S. Overkleeft, and S. J. Fleishman. Combinatorial assembly and design of enzymes. *Science*, 379(6628):195–201, 2023. doi:10.1126/science.ade9434.

Jason W. Locasale and Alejandro Wolf-Yadlin. Maximum Entropy Reconstructions of Dynamic Signaling Networks from Quantitative Proteomics Data. *PLOS ONE*, 4(8): e6522, August 2009. ISSN 1932-6203. doi:10.1371/journal.pone.0006522. URL `https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0006522`. Publisher: Public Library of Science.

Steve W Lockless and Rama Ranganathan. Evolutionarily conserved pathways of energetic connectivity in protein families. *Science*, 286(5438):295–299, 1999.

Christopher W. Lynn, Qiwei Yu, Rich Pang, Stephanie E. Palmer, and William Bialek. Exact minimax entropy models of large-scale neuronal activity. *Physical Review E*, 111 (5):054411, May 2025. doi:10.1103/PhysRevE.111.054411. URL `https://link.aps.org/doi/10.1103/PhysRevE.111.054411`. Publisher: American Physical Society.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos, Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 2023. doi:10.1038/s41587-022-01618-2.

Debora S. Marks, Lucy J. Colwell, Robert Sheridan, Thomas A. Hopf, Andrea Pagnani, Riccardo Zecchina, and Chris Sander. Protein 3d structure computed from evolutionary sequence variation. *PLOS ONE*, 6(12):e28766, 12 2011. doi:10.1371/journal.pone.0028766. URL `https://doi.org/10.1371/journal.pone.0028766`.

Richard N McLaughlin Jr, Frank J Poelwijk, Arjun Raman, Walraj S Gosal, and Rama Ranganathan. The spatial architecture of protein function and adaptation. *Nature*, 491 (7422):138–142, 2012.

Pankaj Mehta, Marin Bukov, Ching-Hao Wang, Alexandre G. R. Day, Clint Richardson, Charles K. Fisher, and David J. Schwab. A high-bias, low-variance introduction to Machine Learning for physicists. *Physics Reports*, 810:1–124, May 2019. ISSN 0370-1573. doi:10.1016/j.physrep.2019.03.001. URL `https://www.sciencedirect.com/science/article/pii/S0370157319300766`.

Thierry Mora and William Bialek. Are Biological Systems Poised at Criticality? *Journal of Statistical Physics*, 144(2):268–302, July 2011. ISSN 1572-9613. doi:10.1007/s10955-011-0229-4. URL `https://doi.org/10.1007/s10955-011-0229-4`.

Thierry Mora, Aleksandra M. Walczak, William Bialek, and Curtis G. Callan. Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences*, 107(12):5405–5410, March 2010. doi:10.1073/pnas.1001705107. URL https://www.pnas.org/doi/10.1073/pnas.1001705107. Publisher: Proceedings of the National Academy of Sciences.

Thierry Mora, Aleksandra M. Walczak, Lorenzo Del Castello, Francesco Ginelli, Stefania Melillo, Leonardo Parisi, Massimiliano Viale, Andrea Cavagna, and Irene Giardina. Local equilibrium in bird flocks. *Nature physics*, 12(12):1153–1157, December 2016. ISSN 1745-2473. doi:10.1038/nphys3846. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5131848/.

Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S Marks, Chris Sander, Riccardo Zecchina, José N Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, 2011a.

Faruck Morcos, Andrea Pagnani, Bryan Lunt, Arianna Bertolino, Debora S. Marks, Chris Sander, Riccardo Zecchina, José N. Onuchic, Terence Hwa, and Martin Weigt. Direct-coupling analysis of residue coevolution captures native contacts across many protein families. *Proceedings of the National Academy of Sciences*, 108(49):E1293–E1301, December 2011b. doi:10.1073/pnas.1111471108. URL https://www.pnas.org/doi/10.1073/pnas.1111471108. Publisher: Proceedings of the National Academy of Sciences.

Radford M Neal. Annealed importance sampling. *Statistics and computing*, 11:125–139, 2001.

E Neher. How frequent are correlated changes in families of protein sequences? *Proceedings of the National Academy of Sciences*, 91(1):98–102, 1994. doi:10.1073/pnas.91.1.98.

H Chau Nguyen, Riccardo Zecchina, and Johannes Berg. Inverse statistical problems: from the inverse ising problem to data science. *Advances in Physics*, 66(3):197–261, 2017a.

H Chau Nguyen, Riccardo Zecchina, and Johannes Berg. Inverse statistical problems: from the inverse Ising problem to data science. *Advances in Physics*, 66(3):197–261, 2017b. Publisher: Taylor & Francis.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *J. Mach. Learn. Res.*, 22(1):57:2617–57:2680, January 2021. ISSN 1532-4435.

Frank J Poelwijk, Michael Socolich, and Rama Ranganathan. Learning the pattern of epistasis linking genotype and phenotype in a protein. *Nature communications*, 10(1):4213, 2019.

Qi Qi, Jiameng Lyu, Kung-Sik Chan, Er-Wei Bai, and Tianbao Yang. Stochastic Constrained DRO with a Complexity Independent of Sample Size. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=VpaXrBFYZ9.

Zi-Hao Qiu, Quanqi Hu, Zhuoning Yuan, Denny Zhou, Lijun Zhang, and Tianbao Yang. Not all semantics are created equal: contrastive self-supervised learning with automatic temperature individualization. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML'23*, pages 28389–28421, Honolulu, Hawaii, USA, July 2023. JMLR.org.

Arjun S Raman, K Ian White, and Rama Ranganathan. Origins of allostery and evolvability in proteins: a case study. *Cell*, 166(2):468–480, 2016.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K. Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield Networks is All You Need. In *International Conference on Learning Representations*, 2021. URL `https://openreview.net/forum?id=tL89RnzIiCd`.

Roshan Rao, Joshua Meier, Tom Sercu, Sergey Ovchinnikov, and Alexander Rives. Transformer protein language models are unsupervised structure learners. In *International Conference on Learning Representations*, 2021a. URL `https://openreview.net/forum?id=fylclEqgvgd`.

Roshan M Rao, Jason Liu, Robert Verkuil, Joshua Meier, John Canny, Pieter Abbeel, Tom Sercu, and Alexander Rives. Msa transformer. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8844–8856. PMLR, 2021b. URL `https://proceedings.mlr.press/v139/rao21a.html`.

Kimberly A Reynolds, Richard N McLaughlin, and Rama Ranganathan. Hot spots for allosteric regulation on protein surfaces. *Cell*, 147(7):1564–1575, 2011.

Kimberly A Reynolds, William P Russ, Michael Socolich, and Rama Ranganathan. Evolution-based design of proteins. In *Methods in enzymology*, volume 523, pages 213–235. Elsevier, 2013.

Danilo Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 1530–1538. PMLR, June 2015. URL `https://proceedings.mlr.press/v37/rezende15.html`. ISSN: 1938-7228.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021a. doi:10.1073/pnas.2016239118.

Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C. Lawrence Zitnick, Jerry Ma, and Rob Fergus. Biological structure

and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences of the United States of America*, 118 (15):e2016239118, April 2021b. ISSN 0027-8424. doi:10.1073/pnas.2016239118. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC8053943/.

Olivier Rivoire, Kimberly A Reynolds, and Rama Ranganathan. Evolution-based functional decomposition of proteins. *PLoS computational biology*, 12(6):e1004817, 2016a.

Olivier Rivoire, Kimberly A Reynolds, and Rama Ranganathan. Evolution-based functional decomposition of proteins. *PLoS computational biology*, 12(6):e1004817, 2016b. Publisher: Public Library of Science San Francisco, CA USA.

William P Russ, Drew M Lowery, Prashant Mishra, Michael B Yaffe, and Rama Ranganathan. Natural-like function in artificial ww domains. *Nature*, 437(7058):579–583, 2005a.

William P Russ, Drew M Lowery, Prashant Mishra, Michael B Yaffe, and Rama Ranganathan. Natural-like function in artificial WW domains. *Nature*, 437(7058):579–583, 2005b. Publisher: Nature Publishing Group UK London.

William P Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, and others. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369 (6502):440–445, 2020a. Publisher: American Association for the Advancement of Science.

William P. Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, and Rama Ranganathan. An evolution-based model for designing chorismate mutase enzymes. *Science (New York, N.Y.)*, 369(6502):440–445, July 2020b. ISSN 1095-9203. doi:10.1126/science.aba3304.

William P Russ, Matteo Figliuzzi, Christian Stocker, Pierre Barrat-Charlaix, Michael Socolich, Peter Kast, Donald Hilvert, Remi Monasson, Simona Cocco, Martin Weigt, et al. An evolution-based model for designing chorismate mutase enzymes. *Science*, 369(6502): 440–445, 2020c.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved Techniques for Training GANs. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/file/8a3363abe792db2d8761d6403605aeb7-Paper.pdf.

Victor H Salinas and Rama Ranganathan. Coevolution-based inference of amino acid interactions underlying protein function. *elife*, 7:e34300, 2018a.

Victor H Salinas and Rama Ranganathan. Coevolution-based inference of amino acid interactions underlying protein function. *elife*, 7:e34300, 2018b. Publisher: eLife Sciences Publications, Ltd.

Elad Schneidman, Michael J. Berry, Ronen Segev, and William Bialek. Weak pairwise correlations imply strongly correlated network states in a neural population. *Nature*, 440 (7087):1007–1012, April 2006. ISSN 1476-4687. doi:10.1038/nature04701.

Alexander Schug, Martin Weigt, José N Onuchic, Terence Hwa, and Hendrik Szurmant. High-resolution protein complexes from integrating genomic information with molecular simulation. *Proceedings of the National Academy of Sciences*, 106(52):22124–22129, 2009.

Damiano Sgarbossa, Umberto Lupo, and Anne-Florence Bitbol. Generative power of a protein language model trained on multiple sequence alignments. *eLife*, 12:e79854, 2023. doi:10.7554/eLife.79854.

Michael Socolich, Steve W Lockless, William P Russ, Heather Lee, Kevin H Gardner, and Rama Ranganathan. Evolutionary information for specifying a protein fold. *Nature*, 437 (7058):512–518, 2005.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *Proceedings of the 32nd International Conference on Machine Learning*, pages 2256–2265. PMLR, June 2015. URL `https://proceedings.mlr.press/v37/sohl-dickstein15.html`. ISSN: 1938-7228.

Yang Song and Diederik P. Kingma. How to train your energy-based models, 2021.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, December 2022. URL `https://proceedings.neurips.cc/paper_files/paper/2022/hash/7b75da9b61eda40fa35453ee5d077df6-Abstract-Conference.html`.

Tyler N Starr and Joseph W Thornton. Epistasis in protein evolution. *Protein science*, 25 (7):1204–1218, 2016.

Natasa Tagasovska, Nathan C. Frey, Andreas Loukas, Isidro Hotzel, Julien Lafrance-Vanasse, Ryan Lewis Kelly, Yan Wu, Arvind Rajpal, Richard Bonneau, Kyunghyun Cho, Stephen Ra, and Vladimir Gligorijevic. A pareto-optimal compositional energy-based model for sampling and optimization of protein sequences. In *NeurIPS 2022 AI for Science: Progress and Promises*, 2022. URL `https://openreview.net/forum?id=U2rNXaTTXPQ`.

Tiberiu Teşileanu, Lucy J. Colwell, and Stanislas Leibler. Protein sectors: Statistical coupling analysis versus conservation. *PLOS Computational Biology*, 11(2):e1004091, 02 2015. doi:10.1371/journal.pcbi.1004091.

L. Theis, A. van den Oord, and M. Bethge. A note on the evaluation of generative models. In *International Conference on Learning Representations*, April 2016. URL `http://arxiv.org/abs/1511.01844`.

Gašper Tkačik, Olivier Marre, Dario Amodei, Elad Schneidman, William Bialek, and Michael J. Berry Ii. Searching for Collective Behavior in a Large Network of Sensory Neurons. *PLOS Computational Biology*, 10(1):e1003408, January 2014. ISSN 1553-7358. doi:10.1371/journal.pcbi.1003408. URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1003408. Publisher: Public Library of Science.

Jeanne Trinquier, Guido Uguzzoni, Andrea Pagnani, Francesco Zamponi, and Martin Weigt. Efficient generative modeling of protein sequences using simple autoregressive models. *Nature Communications*, 12(1):5800, 2021. doi:10.1038/s41467-021-25756-4.

Jérôme Tubiana, Simona Cocco, and Rémi Monasson. Learning compositional representations of interacting systems with restricted boltzmann machines: Comparative study of lattice proteins. *Neural computation*, 31(8):1671–1717, 2019.

Guido Uguzzoni, Shalini John Lovis, Francesco Oteri, Alexander Schug, Hendrik Szurmant, and Martin Weigt. Large-scale identification of coevolution signals across homo-oligomeric protein interfaces by direct coupling analysis. *Proceedings of the National Academy of Sciences*, 114(13):E2662–E2671, 2017.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://papers.nips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html.

Martin Weigt, Robert A. White, Hendrik Szurmant, James A. Hoch, and Terence Hwa. Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America*, 106(1): 67–72, January 2009. ISSN 1091-6490. doi:10.1073/pnas.0805923106.

Florian Wenzel, Kevin Roth, Bastiaan Veeling, Jakub Swiatkowski, Linh Tran, Stephan Mandt, Jasper Snoek, Tim Salimans, Rodolphe Jenatton, and Sebastian Nowozin. How Good is the Bayes Posterior in Deep Neural Networks Really? In *Proceedings of the 37th International Conference on Machine Learning*, pages 10248–10259. PMLR, November 2020. URL https://proceedings.mlr.press/v119/wenzel20a.html. ISSN: 2640-3498.

Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion Models: A Comprehensive Survey of Methods and Applications. *ACM Comput. Surv.*, 56(4):105:1–105:39, November 2023. ISSN 0360-0300. doi:10.1145/3626235. URL https://dl.acm.org/doi/10.1145/3626235.

Edwin Zhang, Vincent Zhu, Naomi Saphra, Anat Kleiman, Benjamin L. Edelman, Milind Tambe, Sham Kakade, and Eran Malach. Transcendence: Generative Models Can Outperform The Experts That Train Them. *Advances in Neural Information Processing Systems*, 37:86985–87012, December 2024a. URL https://proceedings.neurips.cc/paper_fil

es/paper/2024/hash/9e3bba153aa362f961dc43de5cababac-Abstract-Conference.html.

Yijie Zhang, Yi-Shan Wu, Luis A. Ortega, and Andres R. Masegosa. If there is no underfitting, there is no Cold Posterior Effect, 2024b. URL https://openreview.net/forum?id=zamGHHs2u8.

Cheyenne Ziegler, Jonathan Martin, Claude Sinner, and Faruck Morcos. Latent generative landscapes as maps of functional diversity in protein sequence space. *Nature Communications*, 14(1):2222, 2023. doi:10.1038/s41467-023-37958-z.

APPENDIX

Biographical note on my interest in this research

The circumstances through which I came to be interested in the topics concerning this thesis are interesting (and strange) enough that they merit recording.

My parents raised me within an esoteric flavor of Roman Catholicism heavily influenced by the teachings of one Monsignor Luigi Giussani, to whom the first epigraph of this dissertation is attributed. Among the many riches of this education I received was Giussani's esteem for (and somewhat-mystical understanding of) the faculty of human reason.

As the aforementioned epigraphs make clear, Giussani shunned narrow conceptions of reason—ones that put intellectual schema above observation. For him, reason is the demand for total explanation of any and all things, guided by a subordination to the objects of its interest (lest over-analysis and presumption reign). "The object defines the method," I was often told, and was meant to understand it as advice-for-all-of-life, from "Big" questions and relationships down to homework and chores.

As an undergraduate, I found in physics a compelling attempt to "explain everything." But the restlessness of my own reason, a restlessness I was taught to foster, felt constrained by purely reductive theorems. What's more, the certitude offered by these beautiful theorems—of electrodynamics, mechanics, and quantum mechanics—felt remote in many regards, often making little contact with reality as I knew and saw it. Two experiences addressed these concerns.

In statistical mechanics, I found the first inklings of a constructionist's physics, that is, an attempt at understanding not isolated but inter-related parts—both to each other and the whole. This felt closer to "reason as openness to the totality of reality" than any other field in physics I had studied.[1]

1. At the time, words like "constructionist" and "emergence" were not in my vocabulary. Rather, a more

In my experimental course work, I encountered John R. Taylor's *Introduction to Error Analysis: The Study of Uncertainties in Physical Measurements*. With incisive clarity, he outlines what it means to understand the quantifiable *extent* of one's knowledge. My professor at the time would say, "If someone quotes a number and does not give you error bars, then it is not science." Where (and how much) theory actually meets reality became more well-defined. "We don't know" became a quantifiable statement, allowing for new principles to live inside the error bars. I understood the "how" of theory's contact with and subordination to observation, to reality as reality.

In my first year of graduate school, when asked by my would-be advisor about my interests, I mentioned my tendency to reread Taylor's *Error Analysis* for fun (and my broader interest in scientific uncertainty). I was then introduced to maximum entropy modeling. In this technique, I found a principled approach to studying collective phenomenon that adequately respects observation. It has been a delight to study an answer to a twofold desire—my reason's demand for an explanation of the whole and my belief in respecting observation over schema—in one idea.

---

inarticulate wonder predominated; understanding entropy (from arguments of multiplicity of microstates) as an undergraduate remains a most cherished memory.