

Scaling Monte-Carlo-Based Inference on Antibody and TCR Repertoires

Josiah Couch¹, Rohit Arora^{1,*}, Jasper Braun¹, Joseph Kaplinsky^{1,†},
Elliot Hill^{1,‡}, Anthony Li^{1,§}, Brett Altschul² and Ramy Arnaout^{1,3,¶}

¹*Department of Pathology, Beth Israel Deaconess Medical Center, Boston, MA 02215*

²*Department of Physics and Astronomy, University of South Carolina, Columbia, SC 29208*

³*Harvard Medical School, Boston, MA 02115*

(Dated: December 21, 2023)

Previously, it has been shown that maximum-entropy models of immune-repertoire sequence can be used to determine a person’s vaccination status. However, this approach has the drawback of requiring a computationally intensive method to compute each model’s partition function (Z), the normalization constant required for calculating the probability that the model will generate a given sequence. Specifically, the method required generating approximately 10^{10} sequences via Monte-Carlo simulations for each model. This is impractical for large numbers of models. Here we propose an alternative method that requires estimating Z this way for only a few models: it then uses these expensive estimates to estimate Z more efficiently for the remaining models. We demonstrate that this new method enables the generation of accurate estimates for 27 models using only three expensive estimates, thereby reducing the computational cost by an order of magnitude. Importantly, this gain in efficiency is achieved with only minimal impact on classification accuracy. Thus, this new method enables larger-scale investigations in computational immunology and represents a useful contribution to energy-based modeling more generally.

CONTENTS

I. Introduction	1
A. Energy-Based Models	2
B. Estimating Partition Functions	2
II. Methods	3
A. Data	3
B. Models	4
C. Partition Function Estimates Using Non-Repertoire Teammates (Previous Method)	4
D. Partition Function Estimates with Immune-Repertoire Teammates (New Method)	5
E. Quality Testing Via Inference on Sequences	7
III. Results	8
IV. Discussion	8
V. Acknowledgements	10
References	10

I. INTRODUCTION

Energy-based models (EBMs) generally—and maximum entropy (MaxEnt) models particularly—have a wide range of applications, including in statistical physics [1–5] (where the major statistical ensembles all take the form of EBMs), natural language processing [6, 7], finance [8], RNA [9] and protein [10] sequence motifs, ecology [11–13], modeling flocking behavior in birds [14], modeling voting behavior [15], describing patterns of activity in neurons [16, 17], modeling disease outbreaks [18], modeling the environmental preferences of plant pathogens [19], and modeling immune repertoires [20, 21], among many others [22]. As probabilistic models, EBMs can be used as generative models when coupled with Monte Carlo sampling methods. When properly normalized, they can, unlike most types of discriminative models, also be used for Bayesian inference. In nontrivial settings, however, normalizing EBMs (or indeed any probabilistic models based on initially unnormalized probabilities) can be computationally quite expensive.

In previous work, MaxEnt models were trained on antibody heavy-chain (IGH) and T-cell receptor β -chain (TRB) repertoires’ third complementary-determining regions (CDR3s), using features based on the physicochemical properties of their constituent amino acids [21]. It was demonstrated that these models allowed for the classification of influenza vaccination status among 31 samples from 14 individuals. However, this classification required the estimation of partition functions: the normalization constants of the individual probability distributions. This was done using in-house Monte-Carlo (MC) based estimation software, which was computationally expensive. The partition function for a given model is usually abbreviated as Z .

* Current address: Novo Nordisk, Cambridge, MA 02142

† Current address: Genomics England, One Canada Square, London, E14 5AB, UK

‡ Current address: Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC 27710

§ Current address: Argenx US Inc. 33 Arch Street, 32nd Floor Boston, MA 02110

¶ rarnaout@bidmc.harvard.edu

To understand the computational difficulty of the problem, consider a model that represents the distribution of amino-acid sequences comprising some collection of proteins, such as TCRs or B-cell receptors (BCRs) (e.g. IGH). Consider specifically the set of all possible polypeptide chains 100 amino acids in length, which is approximately the length of a TRB or IGH variable region. There are $20^{100} \approx 10^{130}$ such sequences. Thus an exact computation of the partition function would involve a sum of 10^{130} terms. Such a sum is infeasible with present-day computational resources, even before considering that we will likely want to normalize many such models (e.g. one per person per timepoint). In a few special instances, such as the one-dimensional and two-dimensional local Ising or Potts models, there are shortcuts to computing this sum. It is very unlikely such simplifications will exist in general, however, as the problem of computing partition functions has been shown to be #P-hard [23, 24]. Of course in practice, we do not need an exact result, and methods such as bridge sampling [25–27] exist precisely in order to approximate these kinds of sums more efficiently. Yet even in these cases, it may be necessary to generate an enormous Monte Carlo sample from the model in question. We asked whether we could improve the efficiency of estimating Z for each model without sacrificing classification accuracy, using immune repertoires as a test case.

A. Energy-Based Models

An EBM is a model that assigns an unnormalized probability $U_{\vec{\theta}}(x)$ to every potential state x (meaning, in the context of CDR3 repertoires, every possible amino acid sequence up to some maximum length) based on a parameterized energy function $E(x, \vec{\theta})$ according to

$$U_{\vec{\theta}}(x) = e^{-E(x, \vec{\theta})}. \quad (1)$$

The models used in this paper are MaxEnt models, in which the energy takes the form

$$E(x, \vec{\theta}) := E_{\vec{\theta}}(x) = \sum_i \theta_i f_i(x) \quad (2)$$

for a set of features $\{f_i\}$. Such models were introduced in Refs. [4, 5] and are based on distributions long studied in statistical physics. The name “maximum entropy” comes from the fact that these models maximize the entropy of the resulting distribution subject only to constraints on the moments of the features. The parameters $\vec{\theta}$ fix these moments and determine how the distribution is allowed to vary from the uniform distribution (which corresponds to $\vec{\theta} = \vec{0}$) [9, 20, 28].

MaxEnt models can be trained using, for example, gradient descent to maximize the likelihood of a training sample as estimated by the model. Gradients of the log

likelihood turn out to depend only on the feature moments for the current model, which can be estimated using Monte Carlo methods, and the sample moments of the training sample.

B. Estimating Partition Functions

The problem of normalizing an initially unnormalized probability distribution shows up in a number of contexts and has an extensive literature going back several decades. A review of some of this work may be found in section 6 of Ref. [29]. In statistical mechanics, such a normalization constant shows up for the various (microcanonical, canonical, etc.) thermodynamic ensembles and is known as the *partition function*¹, a term we shall use in most of this work, and is usually written Z . Knowing the partition function (as a function of the distribution parameters) allows one to compute all the macroscopic physical quantities that characterize the distribution, such as the mean values of the entropy, internal energy, and magnetization, as well as each of their fluctuations. In the context of Bayesian inference, the posterior distribution takes the form of an unnormalized distribution when the distribution of the evidence is unknown. In a few cases (for example, a multitude of models in one spatial dimension, or the Ising model with nearest-neighbor interactions in two dimensions), the partition functions may be computed analytically, but in the typical case an exact solution is intractable. Indeed, the general case has been shown to be #P-hard [24]. As a result, approximation schemes—either analytical or computational—typically need to be employed.

A large class of computational approximation schemes rely on Monte-Carlo methods for generating model samples; these schemes only estimate the ratio of the partition functions of two models. Alternatively, one can view them as estimating the partition function of one model based on the already known partition function of a second model. In the present work, we will refer to these as the *target* and the *teammate*, respectively. Such methods include bridge sampling [25] and the free energy perturbation method [30]—also known as simple importance sampling (SIS)—among others. The method used previously for immune repertoires by Arora et al. [21] is also in this class. Here we focus on estimating the partition functions themselves (though it should be noted that for the task of maximum likelihood inference, strictly speaking all that is needed is the ratio of the partition function of every model to some fixed reference model).

¹ Strictly speaking, the partition function in statistical mechanics should be understood as the function which, for a parameterized family of distributions, maps parameter values to the corresponding normalization constant, but that is not a distinction we will make here.

Given two unnormalized probability distributions whose densities are given by

$$\mathcal{P}_{\vec{\theta}_0}(x) \propto e^{-E_{\vec{\theta}_0}(x)} \quad (3)$$

and

$$\mathcal{P}_{\vec{\theta}}(x) \propto e^{-E_{\vec{\theta}}(x)} \quad (4)$$

the normalization constants of these (unnormalized) distributions—i.e. the partition functions of these models—are defined to be

$$Z(\vec{\theta}) = \sum_x e^{-E_{\vec{\theta}}(x)} \quad (5)$$

and

$$Z(\vec{\theta}_0) = \sum_x e^{-E_{\vec{\theta}_0}(x)}. \quad (6)$$

In the following we consider $\mathcal{P}_{\vec{\theta}}(x)$ as the target distribution and $\mathcal{P}_{\vec{\theta}_0}$ as the teammate.

The free energy perturbation method [29, 31] estimates $Z(\vec{\theta})$ from $Z(\vec{\theta}_0)$ (or alternatively, estimates their ratio) according to

$$\frac{Z(\vec{\theta})}{Z(\vec{\theta}_0)} = \frac{\sum_x e^{-E_{\vec{\theta}}(x)}}{\sum_x e^{-E_{\vec{\theta}_0}(x)}} \quad (7)$$

$$= \sum_x \frac{e^{-E_{\vec{\theta}}(x)}}{e^{-E_{\vec{\theta}_0}(x)}} \frac{e^{-E_{\vec{\theta}_0}(x)}}{\sum_{x'} e^{-E_{\vec{\theta}_0}(x')}} \quad (8)$$

$$= \left\langle e^{-(E_{\vec{\theta}} - E_{\vec{\theta}_0})} \right\rangle_{\vec{\theta}_0} \quad (9)$$

$$\approx \sum_i e^{-[E_{\vec{\theta}}(x_i) - E_{\vec{\theta}_0}(x_i)]}, \quad (10)$$

where $\{x_i\}$ is a Monte Carlo sample drawn from $\mathcal{P}_{\vec{\theta}_0}$. This Monte Carlo procedure will typically provide a good estimate if every region with non-negligible probability under $\mathcal{P}_{\vec{\theta}}$ also has non-negligible probability under $\mathcal{P}_{\vec{\theta}_0}$. Otherwise, it will tend to do poorly [32]. One way around this is to consider these distributions as part of a set of models $\mathcal{P}_{\vec{\theta}_\lambda}$, $\lambda \in [0, 1]$, such that the $\vec{\theta}_\lambda$ interpolate between $\vec{\theta}_1 := \vec{\theta}$ and $\vec{\theta}_0$. One may then estimate [32]

$$\frac{Z(\vec{\theta})}{Z(\vec{\theta}_0)} = \prod_{i=0}^N \frac{Z(\vec{\theta}_{\lambda_{i+1}})}{Z(\vec{\theta}_{\lambda_i})} \quad (11)$$

$$= \prod_{i=0}^N \left\langle e^{-(E_{\vec{\theta}_{\lambda_{i+1}}} - E_{\vec{\theta}_{\lambda_i}})} \right\rangle_{\vec{\theta}_{\lambda_i}}, \quad (12)$$

with $\lambda_i = \frac{i}{N}$ for some $N > 0$. Bridge sampling was introduced in Ref. [25] as the “acceptance ratio method,” before being rediscovered in Ref. [26], whose authors coined the term “bridge sampling” [32]. Bridge sampling seeks to cure the weaknesses of the free energy perturbation method by using a single intermediate model

$\mathcal{P}_{\text{bridge}}(x) \propto e^{-E_{\text{bridge}}(x)}$. One then estimates

$$\frac{Z(\vec{\theta})}{Z(\vec{\theta}_0)} = \frac{\left\langle e^{-(E_{\text{bridge}} - E_{\vec{\theta}_0})} \right\rangle_{\vec{\theta}_0}}{\left\langle e^{-(E_{\text{bridge}} - E_{\vec{\theta}})} \right\rangle_{\vec{\theta}}}. \quad (13)$$

Both of these methods (and others in this general family) may not perform well (or alternatively may perform well only when the samples used to compute moments are taken to be very large) if the target and teammate distributions are very different (i.e. there is a large distance between them, for an appropriate choice of metric).

In the immune-repertoire example in Ref. [21], computing the partition functions required sampling $\geq 10^{10}$ amino acid sequences via Markov-chain Monte Carlo (MCMC) methods. Even on a highly-parallelized high-performance computing cluster, this requires a day or more of running time. This severely limits the practical feasibility of using this technique directly for Bayesian inference, especially in the case where many such models need to be normalized. We believe that the main reason this method is so high-cost is that this teammate distribution is very far from the model distribution. In fact, it is essentially a uniformly random distribution at each length, combined with a distribution on lengths that depends on the target. (Immune repertoires include amino-acid sequences of multiple lengths.) The advantage of this distribution is that its normalization factor can easily be calculated exactly. The disadvantage is, it has a much higher entropy than any of the target models. As such, it takes an extremely large sample to encounter most of the states which have a high-probability in the target model. In fact, many of the target models are much closer to one-another than they are to their respective teammate models.

The key observation of this paper is that we can use this proximity to our advantage. Once the first few models have been normalized using the expensive but proven method in [21], those few models can be used as alternative teammates for the remaining targets. This strategy can even be used iteratively, with the high-entropy teammates used to normalize the first targets, these targets used as teammates for a second round of targets, this second round of targets used as teammates for a third round, and so on. In the rest of this paper we will show empirically that this method works well, achieving high classification accuracy while significantly speeding up the process of (approximately) normalizing a fairly sizeable batch of IGH and TRB immune-repertoire models.

II. METHODS

A. Data

A total of 19 unique repertoires were studied, summarized in table I. Seventeen of these were IGH and

Repertoire Type	Status	Number of Repertoires
TRB	baseline	4
	infected	3
	cancer	2
IGH	baseline	4
	vaccinated	2
	cancer	2
Random	-	2

TABLE I: Summary of repertoires. For TRB repertoires, *baseline* and *infected* indicate imputed CMV infection status, whereas for IGH repertoires, *baseline* and *vaccinated* indicate influenza vaccination status. Additionally, the TRB cancer repertoires are from breast cancer, whereas the IGH cancer repertoires are from chronic lymphocytic leukemia.

TRB CDR3 repertoires representing diverse physiological and pathophysiological states, including infection, vaccination, and cancer, as well as repertoires from subjects that lacked these conditions. Two were artificial “repertoires” of randomized sequences created from TRB repertoires (see below). The TRB repertoires included three from subjects imputed to be positive for cytomegalovirus (CMV) (“infected”), four imputed to be CMV-negative (“baseline”) [33], and two from subjects with breast cancer (“cancer”) [34]. CMV infection status was imputed as in [35]. The IGH repertoires similarly include two repertoires from subjects who had received an influenza vaccine (“vaccinated”) and four from subjects that had not (“baseline”) [36], as well as two repertoires from subjects with chronic lymphocytic leukemia (“cancer”—although in this case the repertoire includes sequences from the cancerous clone itself, not just sequences elaborated in response to/in the context of the cancer) [37]. The random repertoires were created from TRB repertoires [33] by preserving the lengths of each sequence but randomizing the amino acids among all sequences. As such, these have the same length and single-amino-acid distributions as their source TRB repertoires, but with all correlations between different amino acids, or between amino acids and position in the sequence, randomized away.

B. Models

A total of 29 maximum entropy (MaxEnt) models were trained as described previously [21]. For reference, each model was named according to its cell type, disease state, feature set, and a number that along with this other information uniquely identifies the model. Twenty of these models (one per repertoire, plus a replicate model for one of the random repertoires, as a control) were trained using a set of features consisting of lengths, frequencies of single amino acids in both an entire sequence and in the first and last four amino acids

of a sequence (the canonical stems; IGH and TRB proteins adopt stem-loop structures), and sums of pairwise products of physio-chemical descriptors of amino acids between different locations (including nearest neighbors, next-to-nearest neighbors, and opposites (i.e. first with last, second with second from last, etc.)), and summed over both the entire sequence and just the first and last four amino acids, as well as products of physio-chemical descriptors of four consecutive amino acids. This was feature set 1. To test a second set of features, nine additional MaxEnt models were trained on a subset of the repertoires: two each on IGH baseline, IGH vaccinated, and TRB baseline repertoires, as well as three on TRB infected repertoires. These were trained using a different set of features that did not include products of four physiochemical descriptors, but which did include products between 3rd nearest neighbors. This was feature set 2. Of these 29 models, two fits failed to converge and were thus excluded from the remainder of the study. These were the models trained on the two IGH cancer repertoires, which as described above come from subjects with chronic lymphocytic leukemia. As such, the failure of these models to converge is perhaps unsurprising, given that these repertoires are dominated by a single large clone. Conversely, because these repertoires can be well described by the sequence clone, there is a diminished need for a compact generative model (e.g. a MaxEnt model) to describe them.

C. Partition Function Estimates Using Non-Repertoire Teammates (Previous Method)

For each model, we first used the previous method [21] to estimate the partition function for each model. This method uses the Metropolis-Hastings algorithm to sample from two distributions, the target distribution (one of the immune repertoire models) and a teammate distribution. The teammate distribution was such that the probability of a sequence depends only on its length.

From each of these samples, we estimated the density of states of the target distribution: the distributions of energies, i.e. (unnormalized) negative log probabilities. We describe the procedure graphically (Fig. 1). For the target sample (green in Fig. 1), this was done by binning the energies and counting the number of unique sequences in each bin. For the teammate distribution (yellow in Fig. 1) each sequence contributed a weight equal to its (unnormalized) probability in the target distribution divided by its (normalized) probability in the teammate distribution. Energies were then binned with the same binning as before, with the weight for each sequence added to the corresponding energy bin. This resulted in two histograms representing the density of states. The first of these was estimated based on a sample of 10^{10} Monte Carlo (MC)-generated sequences from the target distribution and represents an absolute estimate; that is, the entries directly estimate the number of unique se-

Name	Repertoire Type	Disease State	Feature Set	Test/Train	Converged
Random 1-1	randomers	-	1	train	yes
Random 1-2	randomers	-	1	train	yes
Random 1-3	randomers	-	1	test	yes
TRB b.l. 1-2	TRB	baseline	1	train	yes
TRB b.l. 1-1	TRB	baseline	1	train	yes
TRB b.l. 1-3	TRB	baseline	1	test	yes
TRB b.l. 1-4	TRB	baseline	1	test	yes
TRB infex 1-3	TRB	infected	1	train	yes
TRB infex 1-1	TRB	infected	1	train	yes
TRB infex 1-2	TRB	infected	1	test	yes
TRB can. 1-1	TRB	cancer	1	train	yes
TRB can. 1-2	TRB	cancer	1	test	yes
TRB b.l. 2-1	TRB	baseline	2	train	yes
TRB b.l. 2-2	TRB	baseline	2	test	yes
TRB infex 2-3	TRB	infected	2	train	yes
TRB infex 2-1	TRB	infected	2	train	yes
TRB infex 2-2	TRB	infected	2	test	yes
IGH b.l. 1-1	IGH	baseline	1	train	yes
IGH b.l. 1-4	IGH	baseline	1	train	yes
IGH b.l. 1-2	IGH	baseline	1	test	yes
IGH b.l. 1-3	IGH	baseline	1	test	yes
IGH vax 1-2	IGH	vaccinated	1	train	yes
IGH vax 1-1	IGH	vaccinated	1	test	yes
IGH b.l. 2-1	IGH	baseline	2	train	yes
IGH b.l. 2-2	IGH	baseline	2	test	yes
IGH vax 2-1	IGH	vaccinated	2	train	yes
IGH vax 2-2	IGH	vaccinated	2	test	yes
IGH can. 1-1	IGH	cancer	1	train	no
IGH can. 1-2	IGH	cancer	1	test	no

TABLE II: Summary of models. Note that the first and second parts of the name (first only for randomer models) indicates cell type and disease state, the next part indicates the feature set used, and the last number (along with the other information) uniquely identifies that model.

quences per bin. The second of these was estimated from a sample of 10^{11} MC-generated sequences drawn from the teammate distribution and represents only a relative estimate, in that the overall histogram differs from the (estimated) density of states by an overall multiplicative constant: a vertical shift in Fig. 1.

We know that scaling the teammate by (the natural logarithm of) the partition function, $\ln Z$, will make the absolute probabilities in each bin equal: it will shift the teammate distribution up or down until the high-confidence part of this curve coincides with the high-confidence part of the target distribution. The high-confidence part of the teammate distribution begins when bins contain enough sequences; it will fall off to the left of that (the lowest-energy sequences are unlikely to be sampled with sufficient density by the teammate’s random sampling). Meanwhile, the high-confidence part of the target distribution is the leftmost part; the distribu-

tion will fall off to the right (higher-energy sequences are unlikely to be sampled sufficiently densely by sampling from the target). The magnitude of the shift required gives the target’s $\ln Z$.

Note this method requires a substantial amount of computational effort (typically a day or more on an academic supercomputing cluster) due to the large numbers of sequences sampled. This large sizes are necessary to ensure meaningful overlap between the two density-of-states histograms.

D. Partition Function Estimates with Immune-Repertoire Teammates (New Method)

Following the bridge-sampling partition function estimates described above, we performed a second analysis using the new method developed for this paper. For

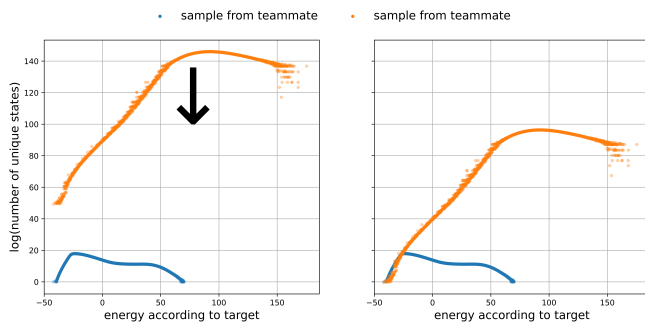


FIG. 1: Left: the density of states estimates from both the target and teammate samples for model DCW4o. Right: the same plot, but with the estimate based on the teammate sample rescaled (downward arrow in left panel). The downward shift required to bring the upper distribution into alignment with the lower distribution is the $\ln Z$.

each of the 27 models, we computed 26 additional estimates using the free energy perturbation method, one using each of the other 26 models as a teammate. These estimates were computed using a 300,000-sequence sample generated from each model using MCMC methods. These samples were independent of those used to compute the bridge-sampling estimates.

For each of these 702 ($= 27$ targets \times 26 teammates) free-energy-perturbation estimates, we found an estimated empirical log error, computed as the absolute value of the difference between the estimate in question for the $\ln Z$ and that found using the non-immune repertoire teammate. These error estimates ranged from 0.000281 to 8.81. We then trained a model to predict when these errors will fall below a threshold, which we chose, somewhat arbitrarily, to be that there should be less than a 30% error in the estimated value of Z . This threshold translated to empirical log errors of up to $\ln(1.3) \approx 0.262$. Since partition functions for different repertoires are observed to differ by many orders of magnitude, a 30% error indicates a quite accurate estimate of the relevant Z .

To this end, we divided the 27 models into two sets: a 15-model training set and a 12-model validation set. The details of this split are given in table II. Since each individual model is designated as either a training model or a validation model, target-teammate pairs divide naturally into three sets: a training set, where both the target and teammate are from the model training set; a validation set, where both are from the model validation set; and a “crossover” set, consisting of the remaining mixed pairs. We trained a random-forest classifier on the training set, achieving a validation accuracy of 89%. The input features for this classifier were simple functions of the model parameters: the root-mean-squared difference between the bias vectors for five different types of biases (including two types of first-order bias, two types of second-order bias, and fourth-order biases), as well as

a binary variable that flagged when one member of the pair was fit on IGH but the other was fit on TRB. On the set of all pairs (including validation pairs, training pairs, and crossover pairs), 75% of pairs classified by the model as “good” had log errors of 0.215 or lower ($< 24\%$).

The methods described above consider every model as both a target and a teammate, and require previously-computed estimates for the partition functions of all models. However, this was only necessary for the purpose of training the classifier. In the remaining analysis, we simulated the scenario where a seed set of only a small number of models were initially identified as likely good teammates for the remaining models. The previous method outlined above was then used to generate estimates for these few models, and the resulting estimates allowed us to use those chosen models as teammates for a second batch of models, which were used as teammates for a third batch of models, and so on, until all partition functions had been estimated. Because all but the initial estimates are computationally inexpensive, this method is overall much more efficient. (See the description of the results in section III below.)

We chose the seed set as follows. For each of our 27 models x , we listed each other model y for which x was a predicted good teammate for y . For each x , and for each y in the list for x , we then added to x ’s list all models z for which y was also a predicted good teammate for z (assuming z was not already on the list). We did this iteratively until we reached a step where no additional models were added to the lists. The result of this is a list of “descendants” for each model. We chose the model with the most descendants as our first model in the seed set, breaking ties arbitrarily. We call this model a . In our data, we found that there was no one model that had every other model as a descendant. Therefore, for all the models x which were not descendants of a , we listed the descendants of x which were not descendants of a , and the one with the most such descendants was chosen as the second model for the seed set, which we refer to as model b . We chose a third model for the seed set using a similar method. The three models chosen for the seed set in this way were those labeled yY7aq, J3AmH, and O8QGE. Collectively, these had all other models as descendants.

The seed-set models were assigned their partition function estimates computed using the previous method. We then iterated through the remaining models, iterating through direct descendants of the seed-set models first. For each model x , we listed the models that had already been assigned a partition function that were predicted by the random forest classifier to be good teammates for x . Using each of those models, we then used free energy perturbation to compute a partition function estimate for x . The median of these estimates was then assigned as the partition function for x , and x was added to the list of models that had been assigned partition functions. Both these assigned partition function estimates and the partition function estimates computed using the non-immune repertoire teammate were then used to do maximum like-

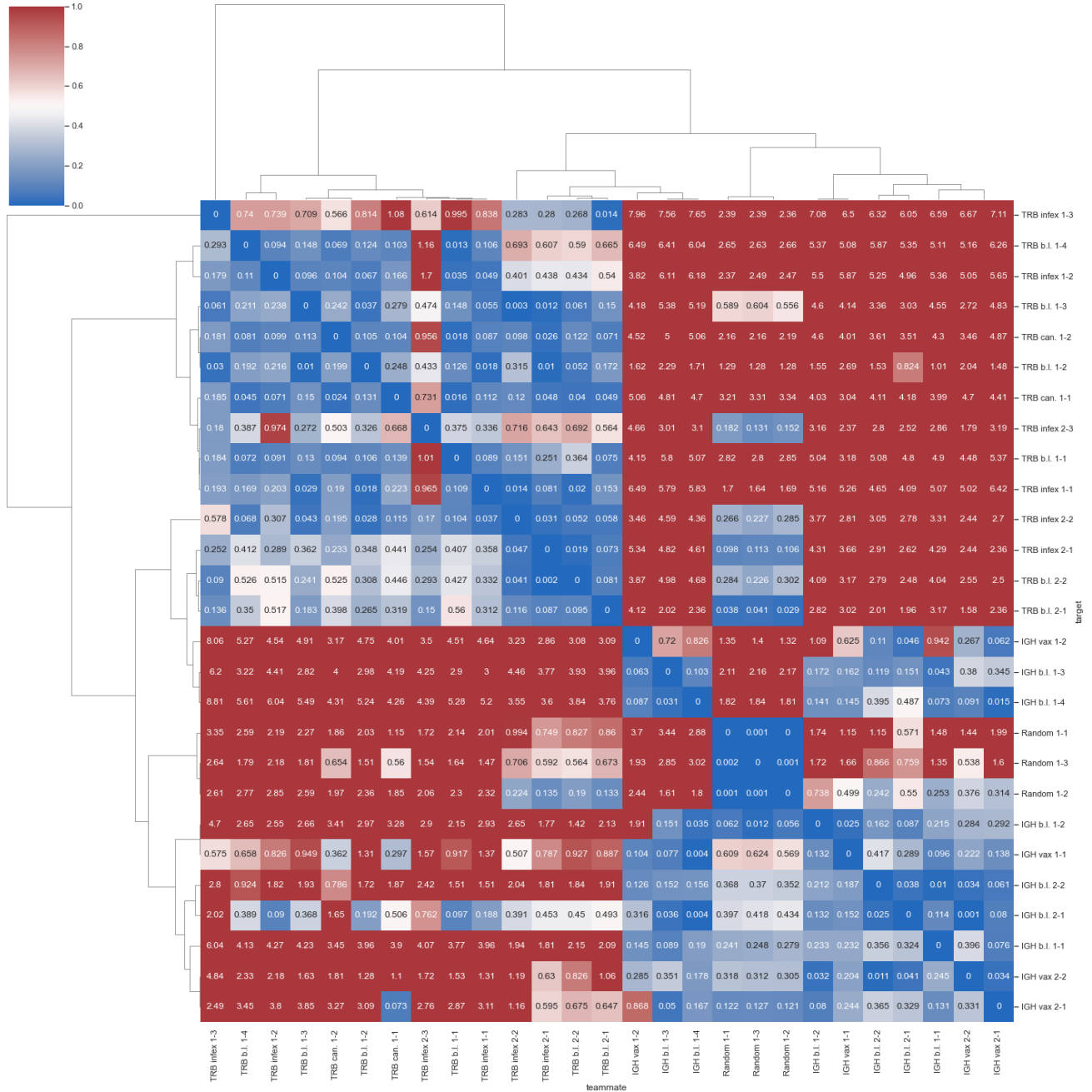


FIG. 2: Differences in log Z estimates using methods (i) and (ii) for each pair (target, teammate) of immune repertoire models.

likelihood inference on sequences.

E. Quality Testing Via Inference on Sequences

For each model, we made an aggregate estimate of the partition function as follows. First we identified all other models which were classified by our random-forest classifier as giving good estimates as teammates for that

model. Our aggregate estimate was then the median of all of these estimates. We refer to these estimates the *median teammate* estimates.

To test if the median teammate estimates were sufficiently accurate, we took 100 samples of 100 sequences from each model, each selected as an independent sub-sample of the 300,000-sequence sample used to compute expectation values in that code. For each of these 2,700 samples (27 models \times 100 samples per model) we used

maximum likelihood to guess which model it originated from. We did this using both the likelihood estimates from the bridge-sampled partition functions and the median teammate partition functions. We then compared the accuracies of both for identifying the correct source models, to see if there was any significant diminution in accuracy resulting from using the median teammate estimate partition functions rather than the bridge sampling partition functions.

III. RESULTS

We compared the previous method[21] and the new method for both computational efficiency and performance of the resulting Bayesian classification. Figures 3 and 4 show the confusion matrices resulting from classifying 100 Monte-Carlo-generated sequences, both by the exact repertoire model they were sampled from (Fig. 3), as well as by sequence type (IGH vs. TRB) and disease or immunization state (Fig. 4). Comparisons of Figs. 3a- 3b to Figs. 4a- 4b show that the new method gives comparable classification performance.

In addition, the new method had a much lower computational cost. Most of the cost came from the MC generation of sequence samples. Each estimate based on the non-immune repertoire teammates required about 10^{10} sequences. In contrast, the estimates using another immune repertoire model as a teammate only used 3×10^5 sequences per model, a savings of 99.997%. This translates to about 10^7 sequences in total across the 27 models—negligible compared to the previous method. As such, the computational cost essentially comes down to the number of more expensive estimates that need to be generated. In the previous method, each of the 27 estimates were of the expensive type. For the new method, we required only 3 of the expensive estimates, from which partition functions for the other 24 models were estimated much more efficiently. As a result, overall the new method leads to about an order of magnitude savings in computational cost.

IV. DISCUSSION

EBMs provide a way to summarize complex systems such as immune repertoires compactly and efficiently, based on aggregate features that are often human-interpretable. They also have the advantage of being generative models, which for immune repertoires means they can quickly and easily produce arbitrarily many *de novo* sequences that are representative of a given repertoire. Partition function estimation is important in EBMs because it allows calculation of the absolute probability of a given state—for example determining which of several immunological states, such as infection or cancer (each described by one or more models), a set of sequences is diagnostically most consistent with [38]. Together with

methods for measuring immunological diversity, EBMs could become an important part of the diagnostic toolkit in next-generation immunology [35, 39], provided partition functions can be estimated efficiently. Here we have demonstrated a highly-efficient new method for estimating partition functions that performs as well as a previous method but much more efficiently, as assessed by correct classification of immune-repertoire sequences using models of real-world repertoires.

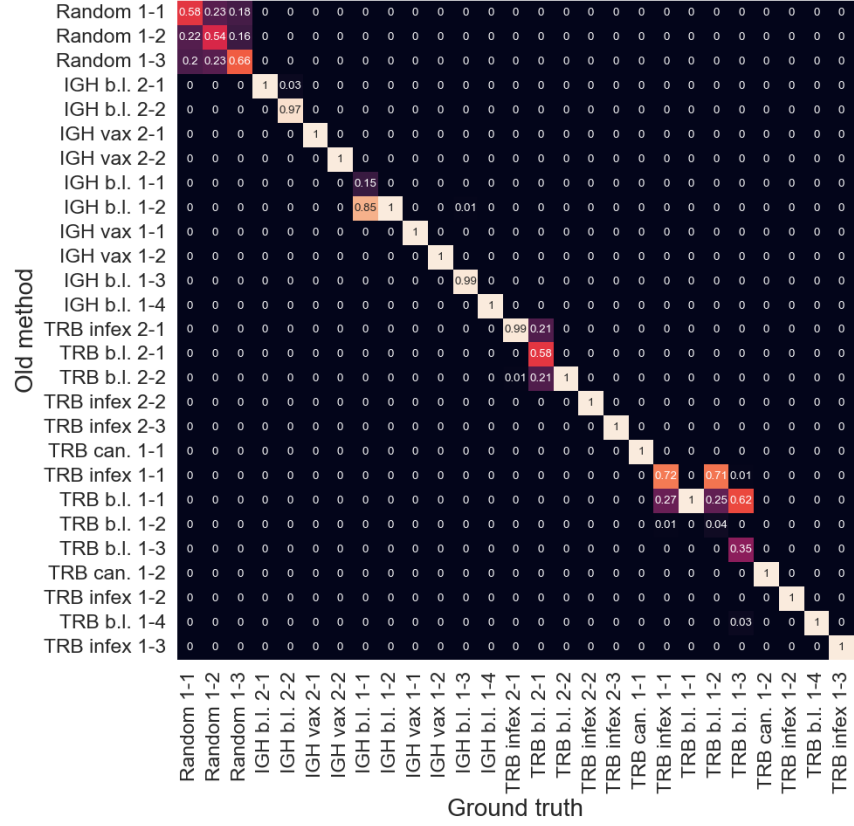
Although we have demonstrated substantial computational savings on this set of diverse IGH and TRB repertoires from a variety of states of health and disease, it should be noted that further work is necessary to precisely define how the computational cost of estimating partition functions will scale as the number of models requiring partition-function estimation increases into the hundreds or thousands. The answer will likely depend in part on how closely related the models are, since we find closely related models tend to make good teammates. If the new method were to scale linearly, as the previous method does [21], then the advantage would be merely a (substantial) multiplicative factor. Based on our results, the new method likely scales sub-linearly, significantly improving the utility of this method in situations where many repertoires are modeled, e.g. representing precisely defined or multifaceted disease states across large clinical cohorts. Curating a database of previously fitted models with partition functions computed would maximize the cost savings of this method, by creating a bank of potential teammates for use in normalizing new models.

In the new method presented here, after partition functions for the seed models are estimated, additional estimates are found using free energy perturbation, arguably the simplest of the MCMC-based methods. In the future, it may be interesting to implement this idea using other MCMC-based methods, to test more broadly how those estimates compare in terms of accuracy and computational cost. It would also be interesting to explore replacing the initial teammate used in the old method, for which probabilities depended only on sequence length, with a better choice of teammate. A model trained on the same repertoire as the target of interest, but with features restricted to contain only couplings between nearest-neighbor pairs of amino acids², would be formally the same as a 20-state 1D Potts model for each length, and as such the partition function for such a model could be found exactly using standard methods. This class of models could provide an improved teammate for the initial estimate, further reducing the cost to normalize an entire batch of models.

We conclude with a general note regarding obstacles to interdisciplinary adoption of EBMs. While the liter-

² Longer range interactions up to k th nearest neighbors can be included by grouping k amino acids into a single variable with 20^k states, though the cost of the exact computation scales exponentially in k .

(a)



(b)

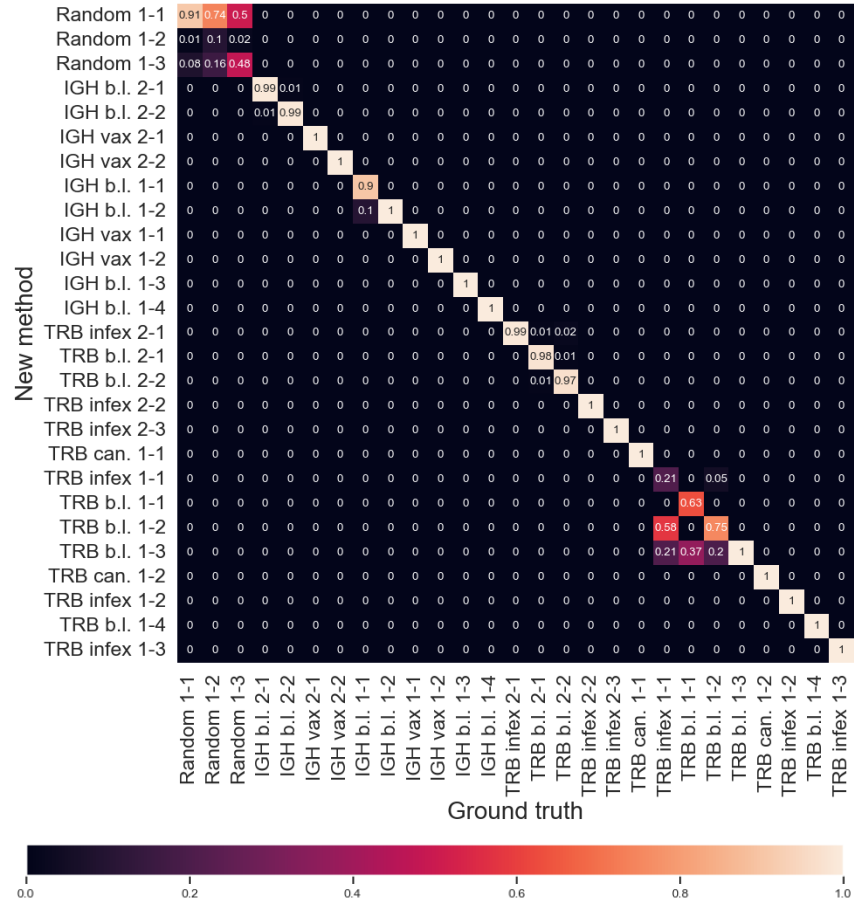


FIG. 3: (a) Confusion matrices between models using method (i). (b) Confusion matrices between models using method (ii).

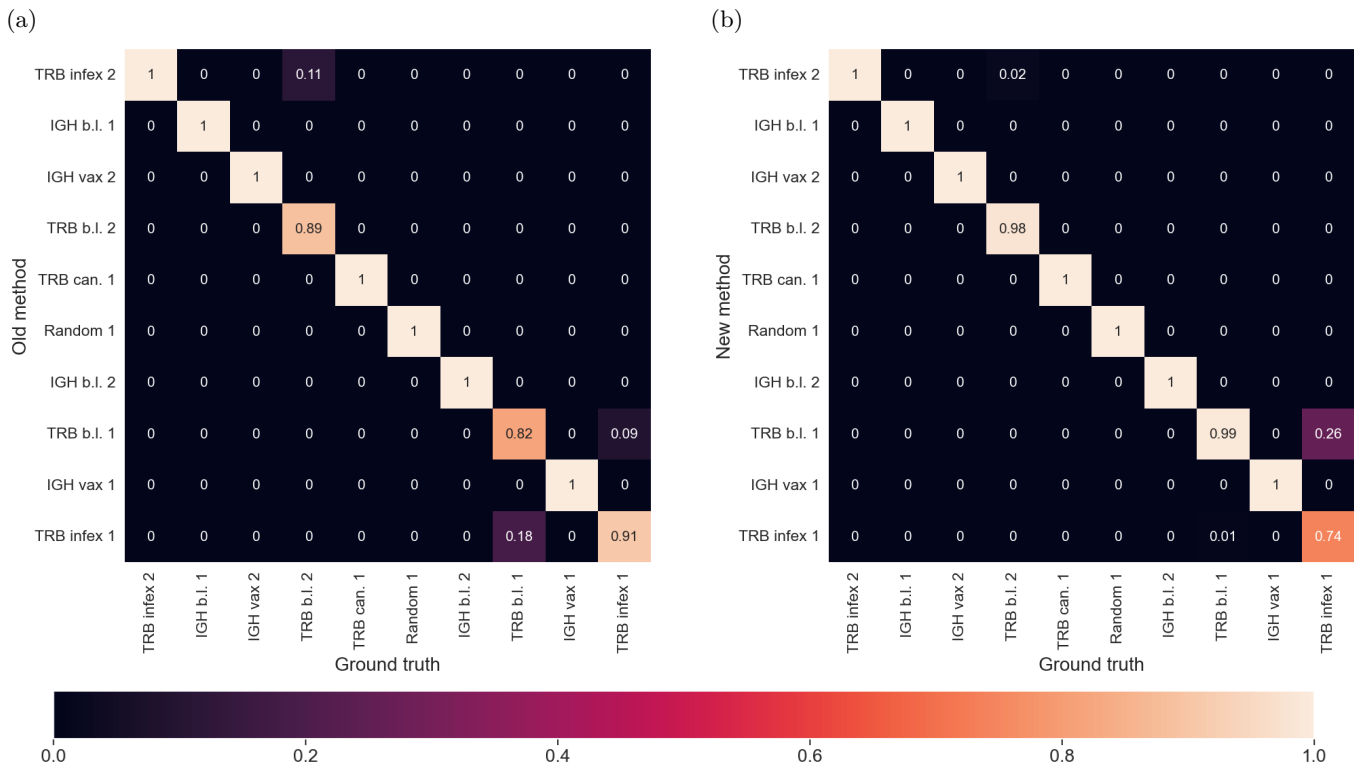


FIG. 4: (a) Confusion matrices between model parameter set/cell type/disease state labels using method (i). (b) Confusion matrices between model parameter set/cell type/disease state labels using method (ii). b.l. = baseline, vax = vaccinated, infex = infected, can. = cancer, and 1,2 refers to the feature set.

ature on MC methods for computing partition functions is extensive and goes back decades, it primarily traces its origins to statistics and statistical physics. Consequently, terminologies and concepts may not be readily accessible to researchers from diverse fields, including the biomedical sciences, whom they could otherwise benefit. Therefore, it may be useful to have a review that introduces these concepts specifically to biomedical researchers. Such a resource would facilitate the dissemination of knowledge, help avoid delays due to reinvention, and encourage the adoption of these powerful computational tools in various research domains. This is especially as the amount of data available in biology and

related fields continues to increase, bringing ever-more-complex systems more fully into the realm of scientific study.

V. ACKNOWLEDGEMENTS

The authors would like to acknowledge that the Research Computing group in the Division of Information Technology at the University of South Carolina contributed to the results of this research by providing High Performance Computing resources and expertise. This work was supported by the NIH (R01AI148747-01).

- [1] L. Boltzmann, Studien über das gleichgewicht der lebendigen kraft zwischen bewegten materiellen punkten, Wiener Berichte **58**, 517 (1868).
- [2] J. W. Gibbs, On the equilibrium of heterogeneous substances, Transactions of the Connecticut Academy of Arts and Sciences **3**, 108–248 (October 1875 – May 1876).
- [3] J. W. Gibbs, On the equilibrium of heterogeneous substances, Transactions of the Connecticut Academy of Arts and Sciences **3**, 343–524 (May 1877 – July 1878).

- [4] E. T. Jaynes, Information Theory and Statistical Mechanics, Physical Review **106**, 620 (1957).
- [5] E. T. Jaynes, Information Theory and Statistical Mechanics. II, Physical Review **108**, 171 (1957).
- [6] J. D. Lafferty and B. Suhm, Cluster expansions and iterative scaling for maximum entropy language models, in *Maximum Entropy and Bayesian Methods*, arXiv:cmp-lg/9509003, edited by K. M. Hanson and R. N. Silver (Springer Netherlands, Dordrecht, 1996) pp. 195–202, cmp-lg/9509003.

- [7] A. L. Berger, S. A. Della Pietra, and V. J. Della Pietra, A maximum entropy approach to natural language processing, *Computational Linguistics* **22**, 39 (1996).
- [8] J. Molins and E. Vives, Long range ising model for credit risk modeling in homogeneous portfolios (2004), [cond-mat/0401378](https://arxiv.org/abs/cond-mat/0401378).
- [9] G. Yeo and C. B. Burge, Maximum entropy modeling of short sequence motifs with applications to RNA splicing signals, *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology* **11**, 377 (2004).
- [10] K. Shimagaki and M. Weigt, Selection of sequence motifs and generative hopfield-potts models for protein families, *Phys. Rev. E* **100**, 032128 (2019), [arXiv:1905.11848 \[q-bio.BM\]](https://arxiv.org/abs/1905.11848).
- [11] B. Shipley, D. Vile, and E. Garnier, From plant traits to plant communities: A statistical mechanistic approach to biodiversity, *Science* **314**, 812 (2006).
- [12] S. J. Phillips, R. P. Anderson, and R. E. Schapire, Maximum entropy modeling of species geographic distributions, *Ecological Modelling* **190**, 231 (2006).
- [13] R. J. Williams, Simple MaxEnt models for food web degree distributions (2009), [0901.0976 \[q-bio\]](https://arxiv.org/abs/0901.0976).
- [14] A. Cavagna, I. Giardina, F. Ginelli, T. Mora, D. Piovani, R. Tavarone, and A. M. Walczak, Dynamical maximum entropy approach to flocking, *Physical Review E* **89**, 042707 (2014), [1310.3810 \[cond-mat, physics:physics, q-bio\]](https://arxiv.org/abs/1310.3810).
- [15] E. D. Lee, C. P. Broedersz, and W. Bialek, Statistical mechanics of the US supreme court, *Journal of Statistical Physics* **160**, 275 (2015), [1306.5004](https://arxiv.org/abs/1306.5004).
- [16] U. Ferrari, T. Obuchi, and T. Mora, Random versus maximum entropy models of neural population activity, *Physical Review E* **95**, 042321 (2017), [1612.02807 \[cond-mat, q-bio\]](https://arxiv.org/abs/1612.02807).
- [17] T.-A. Nghiem, B. Telenczuk, O. Marre, A. Destexhe, and U. Ferrari, Maximum-entropy models reveal the excitatory and inhibitory correlation structures in cortical neuronal activity, *Phys. Rev. E* **98**, 012402 (2018), [arXiv:1801.01853 \[q-bio.NC\]](https://arxiv.org/abs/1801.01853).
- [18] M. Ansari, D. Soriano-Paños, G. Ghoshal, and A. D. White, Inferring spatial source of disease outbreaks using maximum entropy, *Phys. Rev. E* **106**, 014306 (2022), [arXiv:2110.03846 \[physics.soc-ph\]](https://arxiv.org/abs/2110.03846).
- [19] S. D. Cohen, Estimating the climate niche of sclerotinia sclerotiorum using maximum entropy modeling, *Journal of Fungi (Basel, Switzerland)* **9**, 892 (2023).
- [20] T. Mora, A. M. Walczak, W. Bialek, and C. G. Callan, Maximum entropy models for antibody diversity, *Proceedings of the National Academy of Sciences of the United States of America* **107**, 5405 (2010).
- [21] R. Arora, J. Kaplinsky, A. Li, and R. Arnaout, Repertoire-based diagnostics using statistical biophysics, *bioRxiv* [10.1101/519108](https://doi.org/10.1101/519108) (2019), <https://www.biorxiv.org/content/early/2019/01/13/519108>.
- [22] A. De Martino and D. De Martino, An introduction to the maximum entropy approach and its application to inference problems in biology, *Heliyon* **4**, e00596 (2018).
- [23] D. Agrawal, Y. Pote, and K. S. Meel, Partition Function Estimation: A Quantitative Study (2021), [arXiv:2105.11132 \[cs.AI\]](https://arxiv.org/abs/2105.11132).
- [24] D. Roth, On the hardness of approximate reasoning, *Artificial Intelligence* **82**, 273 (1996).
- [25] C. H. Bennett, Efficient estimation of free energy differences from monte carlo data, *Journal of Computational Physics* **22**, 245 (1976).
- [26] W. W. H. MENG X. L., Simulating ratios of normalizing constants via a simple identity: a theoretical exploration, *Statistica Sinica* **6**, 831 (1996).
- [27] Q. F. Gronau, A. Sarafoglou, D. Matzke, A. Ly, U. Boehm, M. Marsman, D. S. Leslie, J. J. Forster, E.-J. Wagenmakers, and H. Steingroever, A Tutorial on Bridge Sampling (2017), [arXiv:1703.05984 \[stat.CO\]](https://arxiv.org/abs/1703.05984).
- [28] W. P. Russ, D. M. Lowery, P. Mishra, M. B. Yaffe, and R. Ranganathan, Natural-like function in artificial WW domains, *Nature* **437**, 579 (2005).
- [29] R. M. Neal, *Probabilistic Inference Using Markov Chain Monte Carlo Methods*, Tech. Rep. (Dept. of Computer Science, University of Toronto., 1993).
- [30] R. W. Zwanzig, High-Temperature Equation of State by a Perturbation Method. I. Nonpolar Gases, *The Journal of Chemical Physics* **22**, 1420 (1954).
- [31] C. J. Geyer and E. A. Thompson, Constrained Monte Carlo Maximum Likelihood for Dependent Data, *Journal of the Royal Statistical Society: Series B (Methodological)* **54**, 657 (1992).
- [32] R. M. Neal, Estimating Ratios of Normalizing Constants Using Linked Importance Sampling (2005), [arXiv:math/0511216 \[math.ST\]](https://arxiv.org/abs/math/0511216).
- [33] O. V. Britanova, E. V. Putintseva, M. Shugay, E. M. Merzlyak, M. A. Turchaninova, D. B. Staroverov, D. A. Bolotin, S. Lukyanov, E. A. Bogdanova, I. Z. Mamedov, Y. B. Lebedev, and D. M. Chudakov, Age-related decrease in TCR repertoire diversity measured with deep and normalized sequence profiling, *Journal of Immunology (Baltimore, Md.: 1950)* **192**, 2689 (2014).
- [34] J. F. Beausang, A. J. Wheeler, N. H. Chan, V. R. Hanft, F. M. Dirbas, S. S. Jeffrey, and S. R. Quake, T cell receptor sequencing of early-stage breast cancer tumors identifies altered clonal structure of the t cell repertoire, *Proceedings of the National Academy of Sciences* **114**, 10.1073/pnas.1713863114 (2017).
- [35] R. Arora and R. Arnaout, Repertoire-scale measures of antigen binding, *Proceedings of the National Academy of Sciences of the United States of America* **119**, e2203505119 (2022).
- [36] C. Vollmers, R. V. Sit, J. A. Weinstein, C. L. Dekker, and S. R. Quake, Genetic measurement of memory b-cell recall using antibody repertoire sequencing, *Proceedings of the National Academy of Sciences of the United States of America* **110**, 13463 (2013).
- [37] R. J. Bashford-Rogers, A. L. Palser, B. J. Huntly, R. Rance, G. S. Vassiliou, G. A. Follows, and P. Kellam, Network properties derived from deep sequencing of human b-cell receptor repertoires delineate b-cell populations, *Genome Research* **23**, 1874 (2013).
- [38] R. A. Arnaout, E. T. L. Prak, N. Schwab, F. Rubelt, and the Adaptive Immune Receptor Repertoire Community, The future of blood testing is the immunome, *Frontiers in Immunology* **12**, 626793 (2021).
- [39] J. Kaplinsky and R. Arnaout, Robust estimates of overall immune-repertoire diversity from high-throughput measurements on samples, *Nature Communications* **7**, 11881 (2016).