Generative Capacity of Probabilistic Protein Sequence Models

Francisco McGee
Temple University
Quentin Novinger
Temple University
Ronald Levy
Temple University
Vincenzo Carnevale
Temple University https://orcid.org/0000-0002-1918-8280
Allan Haldane (tuf33565@temple.edu)
Temple University

Article

Keywords: generative protein sequence models (GPSMs), Potts models, variational autoencoders (VAEs)

Posted Date: April 5th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-145189/v1

Version of Record: A version of this preprint was published at Nature Communications on November 2nd, 2021. See the published version at https://doi.org/10.1038/s41467-021-26529-9.

Generative Capacity of Probabilistic Protein Sequence Models

Francisco McGee[1,2,4]**, Quentin Novinger**[2,5]**, Ronald M Levy**[1,3,4,6]**, Vincenzo Carnevale**[2,3,*]**, and Allan Haldane**[1,6,*]

[1]Center for Biophysics and Computational Biology, Temple University, Philadelphia, 19122, USA
[2]Institute for Computational Molecular Science, Temple University, Philadelphia, 19122, USA
[3]Department of Biology, Temple University, Philadelphia, 19122, USA
[4]Department of Chemistry, Temple University, Philadelphia, 19122, USA
[5]Department of Computer & Information Sciences, Temple University, Philadelphia, 19122, USA
[6]Department of Physics, Temple University, Philadelphia, 19122, USA
[*]Corresponding authors: vincenzo.carnevale@temple.edu, allan.haldane@temple.edu

ABSTRACT

Potts models and variational autoencoders (VAEs) have recently gained popularity as generative protein sequence models (GPSMs) to explore fitness landscapes and predict the effect of mutations. Despite encouraging results, quantitative characterization and comparison of GPSM-generated probability distributions is still lacking. It is currently unclear whether GPSMs can faithfully reproduce the complex multi-residue mutation patterns observed in natural sequences arising due to epistasis. We develop a set of sequence statistics to comparatively assess the accuracy, or "generative capacity", of three GPSMs: a pairwise Potts Hamiltonian, a vanilla VAE, and a site-independent model, using natural and synthetic datasets. We show that the generative capacity of the Potts Hamiltonian model is the largest; the higher order mutational statistics generated by the model agree with those observed for natural sequences. In contrast, we show that the vanilla VAE's generative capacity lies between the pairwise Potts and site-independent models. Importantly, our work measures GPSM generative capacity in terms of higher-order sequence covariation and provides a new framework for evaluating and interpreting GPSM accuracy that emphasizes the role of epistasis.

Introduction

Recent progress in decoding the patterns of mutations in protein multiple sequence alignments (MSAs) has highlighted the importance of mutational covariation in determining protein function, conformation and evolution, and has found practical applications in protein design, drug design, drug resistance prediction, and classification.[1–3] These developments were sparked by the recognition that the pairwise covariation of mutations observed in large MSAs of evolutionarily diverged sequences belonging to a common protein family can be used to fit maximum entropy "Potts" statistical models.[4–6] These contain pairwise statistical interaction parameters reflecting epistasis[7] between pairs of positions. Such models have been shown to accurately predict physical contacts in protein structure,[6,8–10] and have been used to significantly improve the prediction of the fitness effect of mutations to a sequence compared to site-independent sequence variation models which do not account for covariation.[11,12] They are "generative" in the sense that they define the probability, $p(S)$, that a protein sequence S results from the evolutionary process. Intriguingly, the probability distribution $p(S)$ can be used to sample unobserved, and yet viable, artificial sequences. In practice, the model distribution $p(S)$ depends on parameters that are found by maximizing a suitably defined likelihood function on observations provided by the MSA of a target protein family. As long as the model is well specified and generalizes well from the training MSA, it can then be used to generate new sequences, and thus a new MSA whose statistics should match those of the original target protein family. We refer to probabilistic models that create new protein sequences in this way as generative protein sequence models (GPSMs).

The fact that Potts maximum entropy models are limited to pairwise epistatic interaction terms and have a simple functional form for $p(S)$ raises the possibility that their functional form is not flexible enough to describe the data, i.e. that the model is not well specified. While a model with only pairwise interaction terms can predict complex patterns of covariation involving three or more positions through chains of pairwise interactions, it cannot model certain triplet and higher patterns of covariation that require a model with more than pairwise interaction terms.

For example, a Potts model cannot predict patterns described by an XOR or boolean parity function in which the n-th residue is determined by whether an odd number of the $n-1$ previous residues have a certain value (see Supplementary Information). While some evidence has suggested that in the case of protein sequence data the pairwise model is sufficient and necessary to model sequence variation,[13–15] some of this evidence is based on averaged properties, and there appears to be some weak evidence for the possibility of rare "higher-order epistasis" affecting protein evolution,[7,16–18] by which we mean the possibility that subsequence frequencies of three or more positions cannot be reproduced by a model with only pairwise interactions. Fitting maximum entropy models with all triplet interactions is not feasible without significant algorithmic innovation, since for a protein of length 100 it would require approximately 10B parameters and enormous MSA datasets to overcome finite sampling error (see Supplementary Information). However, recent developments in powerful machine learning techniques applied to images, language, and other data have shown how complex distributions $p(S)$ can be fit with models using more manageable parameter set sizes. Building on the demonstrated power of incorporating pairwise epistasis into protein sequence models, this has motivated investigation of machine learning strategies for generative modelling of protein sequence variation which can go beyond pairwise interactions, including Restricted Boltzmann Machines (RBMs),[3] variational autoencoders (VAEs),[19–22] General Adversarial Networks (GANs),[23] transformers,[24–27] and others.

One technique in particular, the VAE,[28,29] has been cited as being well suited for modelling protein sequence covariation, with the potential to detect higher order epistasis.[19,20] The VAE also potentially gives additional insights into the topology of protein sequence space through examination of the "latent" (hidden) parameters of the model, which have been suggested to be related to protein sequence phylogenetic relationships.[19–21,26] One implementation of a VAE-GPSM, "DeepSequence", found that the VAE model was better able to predict experimental measurements of the effect of mutations in deep mutational scans than a pairwise Potts model, which was attributed to the VAE's ability to model higher-order epistasis.[11,19] However, it has also been suggested by others that the improvement shown by DeepSequence could be attributed to the use of biologically motivated priors and engineering efforts, rather than because it truly captured higher-order epistasis.[20] Furthermore, while VAE-GPSMs are generative and aim to capture the protein sequence distribution $p(S)$, to our knowledge none of these studies have thoroughly tested what we will call the "generative capacity"[30,31] of the VAE model, meaning the ability of the model to generate new sequences drawn from the model distribution $p(S)$, which are statistically indistinguishable from those of a given "target" protein family. Testing the generative capacity, specifically higher-order covariation, of a GPSM is a fundamental check of whether the model is well specified and generalizes well from the training set, two prerequisites to capturing higher-order epistasis.

Here, we perform a series of numerical experiments that comparatively explore the generative capacity of GPSMs. These GPSMs include a pairwise Potts Hamiltonian model with pairwise interaction terms (Mi3),[32] a vanilla variational autoencoder (vVAE), and a site-independent model which does not model covariation (Indep). We choose the simplest, original[28,29] "vanilla"[33] architecture qualitatively similar to that used in previous VAE-GPSM studies,[21] as opposed to more complex VAE-GPSM architectures used by others (see Methods).[19,22] We evaluate the generative capacity of a model using four MSA statistics: pairwise covariance correlations,[2,13,34,35] higher-order marginals (r_{20}),[13] Hamming distance distributions,[2,13,36,37] and statistical energy correlations.[13] We compute these statistics for MSAs generated by each GPSM, and validate the GPSMs by comparison to expectation. We also test how each GPSM behaves as fewer training sequences are provided, and thus how well the GPSMs generalize, using two training MSA sizes: one representing the estimated size upper bound for a Pfam protein family (10K) (see Supplementary Information),[38] and one being large enough to eliminate out-of-sample error in our generative capacity measurements (see Results).[39] We evaluate GPSM performance both for natural datasets and for synthetic datasets in which the ground truth is known. Our comparative analysis thus consists of four intersecting variables: three GPSMs, a suite of four generative capacity metrics, two training MSA sizes, and two dataset types.

Our results show that while the vVAE models some epistasis, it is unable to reproduce MSA statistics as well as Mi3, but is more accurate than Indep for all metrics tested. We find that vVAE performs more similarly to Mi3 than it does to Indep according to commonly used metrics such as the Hamming distance distribution, but performs more like the Indep model for the r_{20} metric, which is the metric that most directly tests the GPSM's ability to model higher-order sequence covariation. This result suggests r_{20} is a more sensitive measure of GPSM generative capacity than the other standard metrics, representing a novel and powerful metric to discriminate between GPSMs by averaging over higher-order covariation terms up to the 10th order. Our work consolidates several benchmark statistics of generative capacity for GPSMs, offering a novel framework for evaluating and interpreting GPSM accuracy in the context of higher-order covariation. By quantifying and comparing GPSM performance in our innovative epistasis-oriented approach, we hope to better understand the challenges and limitations inherent to

Specification Error	occurs when the functional form $p_\theta(S)$ of a GPSM is not flexible enough to accurately model the target distribution $p^0(S)$ for any choice of parameters θ
Out-of-Sample Error	occurs when a GPSM fit to a finite training dataset fails to correctly model unseen data, and is a consequence of overfitting
Estimation Error	occurs due to statistical error in the MSA evaluation metrics when computed from finite evaluation and target MSAs
Training MSA	drawn from $p^0(S)$, used to train or parameterize a GPSM
Target MSA	drawn from $p^0(S)$ separately from the training MSA, used as a validation dataset
Evaluation MSA	drawn from $p_\theta(S)$ for a parameterized GPSM, used to compare to the target MSA

Table 1. Glossary of error types, and the MSA datasets we use to evaluate these errors.

generative modelling of natural protein sequence datasets, better gauge the state of the art in the field, and provide insight for future efforts in terms of minimizing the confounding effects of data limitations.

Results

Target Distributions

Our goal is to set baseline expectations for the generative capacity of the vVAE and other GPSMs when fit to synthetic or natural protein sequence data of varying training MSA sizes. Generative models of protein MSAs define a distribution $p_\theta(S)$ for the probability of a sequence S appearing in an MSA dataset given model parameters θ. The model parameters are fit by either exact or approximate maximum likelihood inference of the likelihood $\mathscr{L} = \sum_{S\in \text{MSA}} p_\theta(S)$ over a training MSA, using regularization techniques to prevent overfitting. The sequences in the training MSA are assumed to be identical independent samples from a "target" probability distribution $p^0(S)$, which is generally unknown.[28] For a model with high generative capacity, $p_\theta(S)$ will closely approximate $p^0(S)$.[20] In this study, we test the generative capacity of three such GPSMs: a pairwise Potts model (Mi3),[32] a vanilla VAE (vVAE) (see Methods), and a site-independent model (Indep), each with a different functional form of $p_\theta(S)$.

It is not possible to measure the similarity of $p_\theta(S)$ and $p^0(S)$ directly because of the high dimensionality of sequence space, since the number of sequence probabilities to compare is equal to q^L, where L is the sequence length (typically ~ 300) and q is the alphabet size (~ 21). Instead, we measure how well derived statistics computed from evaluation MSAs generated by each GPSM match those of target MSAs drawn from the target distribution. We use four statistics relevant to quantifying protein MSAs: the pairwise Hamming distance distribution, the pairwise covariance scores, higher order marginals (r_{20}), and, when possible, the GPSM's ability to predict $p^0(S)$ for individual sequences (see Methods).

GPSM Error

We divide our tests into two analysis tracks: one synthetic, in which we train the GPSMs on synthetic MSAs generated from a known target distribution; and one natural, in which we train the GPSMs on a representative natural protein family MSA, the kinase super family, sequestered from Uniprot/TrEMBL.[40] These analysis tracks are meant to probe and isolate two distinct forms of error which may cause $p_\theta(S)$ to deviate from $p^0(S)$ (see Table 1). The first is "specification error",[41] which occurs when the functional form of $p_\theta(S)$ of a model is not flexible enough to accurately model the target probability distribution $p^0(S)$ for any choice of parameters. A key motivation for choosing a VAE over a Potts model is its potentially lower specification error when higher-order epistasis is present.[19] Indeed, Potts models are limited to pairwise interaction terms of a particular functional form, while VAEs are not. The second form of error is "out-of-sample error",[42] caused by a paucity of training samples, and is the consequence of overfitting.[43] Even a well-specified model could fail to generalize when fit to a small training dataset, and may mis-predict $p^0(S)$ for test sequences, so it follows that increasing training MSA size reduces out-of-sample error. Beyond specification and out-of-sample error, which each reflect an aspect of GPSM generalization error, there can be "estimation error" in our MSA test statistics due to the finite MSA sizes we use to estimate their values, which sets an upper bound on how well these statistics can match their target values, depending on the metric.[44] Finally, other errors may arise due to implementation limitations of the inference methods, for instance due to finite precision arithmetic or to finite sample effects when Monte Carlo methods are used.

The synthetic analysis allows us to isolate specification error, and minimize both out-of-sample and estimation error, because here the target distribution is known exactly, and we can generate arbitrarily large training, target, and evaluation datasets. It also allows us to quantify out-of-sample error by modulating the training MSA size.

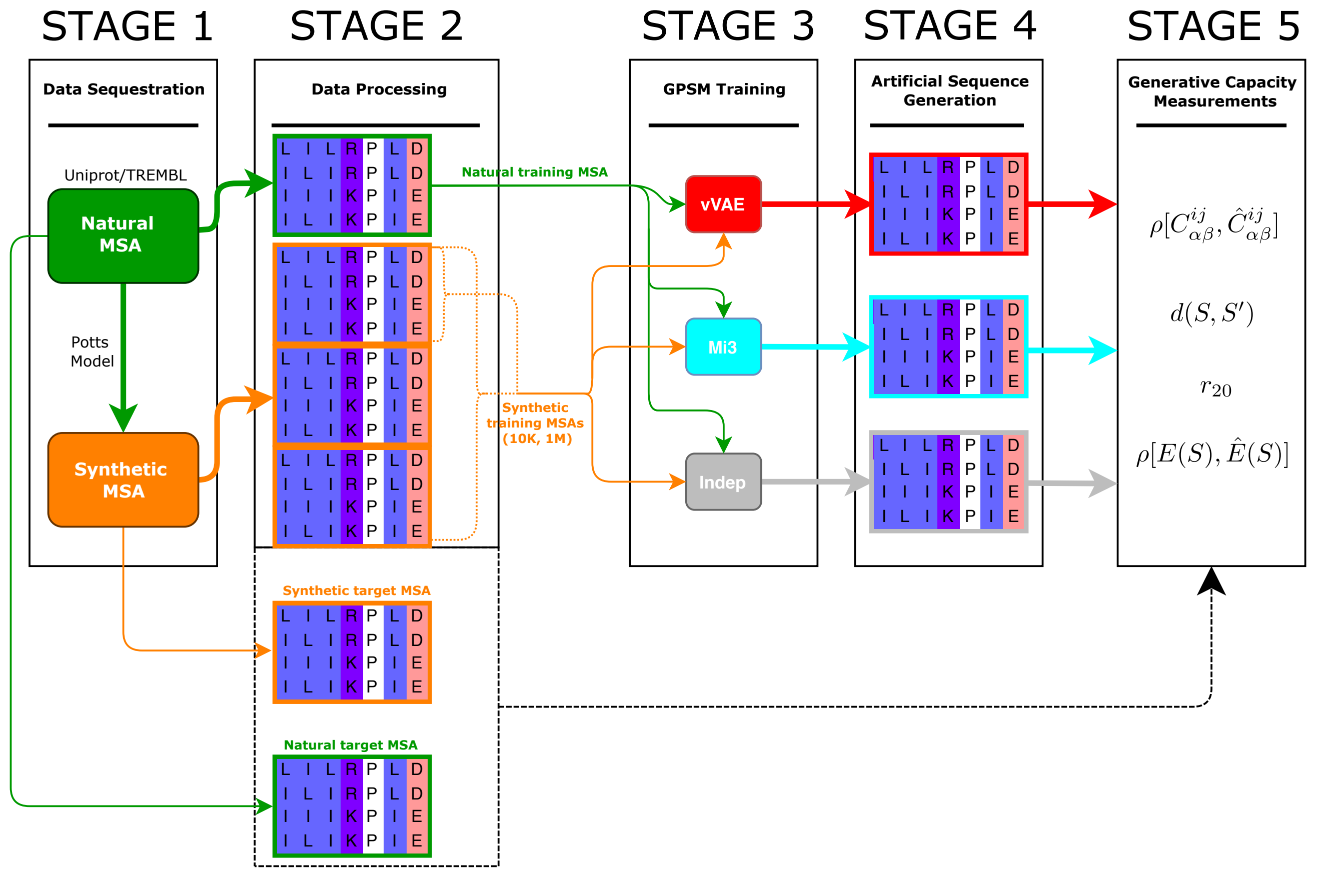

Figure 1. 5-stage analysis pipeline diagram. Stage 1: Data Sequestration. Two different MSAs are sequestered, one natural and the other synthetic, generated by a Potts model fit to the natural MSA. Stage 2: Data Processing. Sequences are indexed and split. Non-overlapping training, target, and evaluation MSAs are shown. Stage 3: GPSM Training. GPSMs are trained. Stage 4: Artificial Sequence Generation. Evaluation MSAs are generated from each GPSM. Stage 5: Generative Capacity Measurements. Computation and visualization of generative capacity metrics is performed.

We specify the synthetic target probability distribution $p^0(S)$ to be exactly the Potts model distribution we inferred based on natural protein-kinase sequence data using Mi3 in our natural analysis (see Methods).[32] The sequences generated from this synthetic target distribution should have statistical properties similar to real, or "natural," protein family MSAs, albeit constrained by the fact that the Hamiltonian model used to generate the synthetic dataset is limited to pairwise epistatic interaction terms only.[6,13,45] Whereas Indep and vVAE may still fail to model this known probability distribution in the synthetic analysis, our expectation is that Mi3 will be unaffected by specification error in the synthetic tests, since the target MSA is sampled from the same probability distribution used to carry out the inference. In Supplementary Information, we also perform an alternate synthetic test which does not favor Mi3 in this way, in which the target distribution is instead specified by a vVAE model, finding that both Mi3 and a vVAE are able to fit this target vVAE distribution accurately. In our synthetic analysis we test two synthetic training MSA sizes: (i) 1M sequences, to minimize overfitting effects and consequently out-of-sample error, thereby isolating GPSM specification error in this experiment; and (ii) 10K sequences, to illustrate the expected GPSM performance on typical datasets, as most protein families in Pfam have less than 10K independent effective sequences (see Supplementary Information).[38]

The natural analysis examines the performance of the models on natural sequence data, which potentially contain higher-order correlations that require a Hamiltonian model with triplet or higher-order interaction terms to capture. On this dataset, vVAE could potentially outperform Mi3, depending on the importance of higher-order epistatic terms, if present.[19] However, unlike in the synthetic analysis, here we do not know *a priori* the target distribution and, most importantly, we have only limited datasets for both training and evaluation. In the natural tests, the training and target MSAs each contain $\sim$10K non-overlapping kinase sequences from the Uniprot/TREMBL database.

Our overall testing procedure is outlined in Fig.1 (see Methods), and the terms discussed are summarized in Table 1. Our training datasets are either a natural protein sequence dataset obtained from Uniprot/TREMBL, or a synthetic training dataset. We then fit the GPSMs to the training datasets, and generate evaluation MSAs from each model. Finally, using our suite of four generative capacity metrics, we compare statistics of the evaluation MSAs to those of "target" MSAs, which contain sequences drawn from the target distributions and therefore represent our expectation.

Pairwise covariance correlations

We first examine the pairwise covariance scores for pairs of amino acids of an MSA defined as $C^{ij}_{\alpha\beta} = f^{ij}_{\alpha\beta} - f^i_\alpha f^j_\beta$. Here, $f^{ij}_{\alpha\beta}$ are the MSA bivariate marginals, meaning the frequency of amino acid combination α, β at positions i, j in the MSA. f^i_α and f^j_β are the univariate marginals, or individual amino acid frequencies at positions i and j. Each covariance term measures the difference between the joint frequency for pairs of amino acids and the product of the single-site residue frequencies, i.e. the expected counts in the hypothesis of statistical independence. The scores equal 0 if the two positions do not covary. Coevolving amino acids are an important aspect of sequence variation in protein MSAs, and a GPSM's ability to reproduce the pairwise covariance scores of the training dataset has been used in the past as a fundamental, non-trivial measure of the GPSM's ability to model protein sequence covariation.[2,13,34,35]

For each GPSM, we compare pairwise covariance scores for all pairs of positions and residues $\hat{C}^{ij}_{\alpha\beta}$ in their respective evaluation MSA to the corresponding target pair $C^{ij}_{\alpha\beta}$ in the target MSA using the Pearson correlation coefficient $\rho(\{C^{ij}_{\alpha\beta}\}, \{\hat{C}^{ij}_{\alpha\beta}\})$ (Fig. 2, Top Row). In the synthetic tests we evaluate this statistic using 500K sequences for both the target and evaluation MSAs, while for the natural test we compare 500K evaluation sequences to the available 10K target sequences. Mi3 accurately reproduces the target covariance scores in all tests ($\rho = 0.99$ in Fig. 2a, $\rho = 0.95$ in Fig. 2b, and $\rho = 0.88$ in Fig. 2c respectively). The somewhat lower value for Mi3 in the natural analysis of $\rho = 0.88$ is accounted for entirely by increased estimation error in that test, as only 10K target sequences are available for evaluation, and the expected ρ due only to estimation error is $\rho \sim 0.87$, which is computed by comparing the natural 10K target sequences to the natural 10K training sequences using this metric. The high generative capacity of Mi3 is expected, because Mi3 parameters are optimized to exactly reproduce the joint frequencies of pairs of amino acids from the target MSA. vVAE inference does not include this constraint, and we find that even when trained on the larger (1M) dataset of synthetic sequences, the covariances computed from vVAE's evaluation MSA are generally smaller in magnitude than those of the target, showing smaller correlation with the target than Mi3 ($\rho = 0.9$, Fig. 2a). This amount of error in ρ can primarily be attributed to vVAE specification error, since training GPSMs on 1M sequences minimizes out-of-sample error, and the large evaluation MSAs make

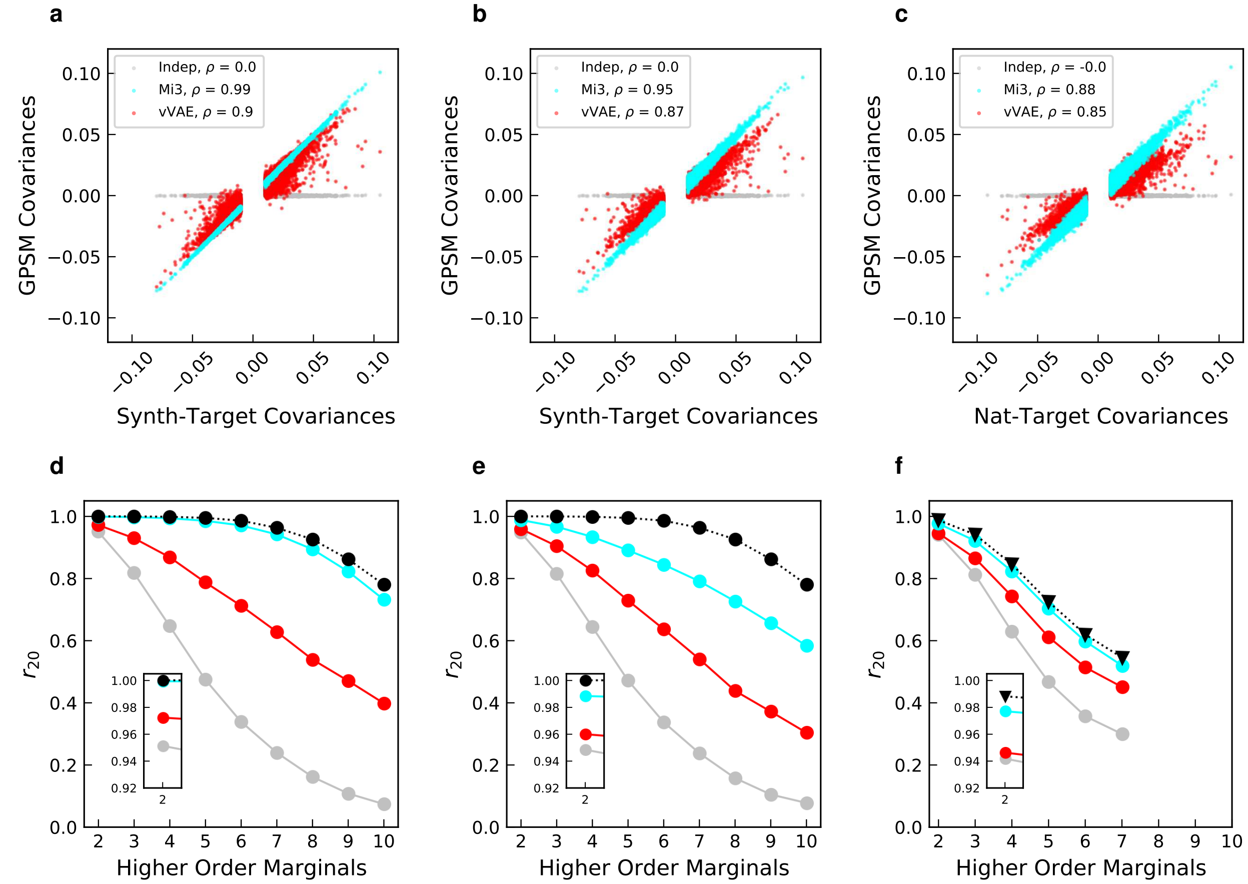

Figure 2. Pairwise covariance correlations (Top Row) and Pearson r_{20} correlations (Bottom Row). Mi3 (cyan), vVAE (red), and Indep (gray) are compared to target distribution (black, where shown). GPSMs were trained on 1M synthetic (**a**, **d**,) 10K synthetic (**b**, **e**), or 10K natural (**c**, **f**) kinase sequences from the corresponding target distribution. **Top Row**, These covariance correlation plots show whether the covariances of the GPSMs match those of their respective targets at the fundamental pairwise level. Synthetic target covariances (x-axis) and GPSM-generated evaluation covariances (y-axis) were computed from 500K-sequence MSAs (**a**, **b**). In contrast, due to limited availability of natural protein sequence data, natural target covariances were computed from a 10K-sequence target MSA, whereas GPSM-generated covariances (y-axis) were computed from 500K-sequence evaluation MSAs (**c**). For each covariance scatterplot, a Pearson correlation coefficient ρ was computed between the GPSM covariance set and the target covariance set. Generative capacity for Mi3 and Indep appear insensitive to training sample size for this metric. Covariances around zero were removed for easier plotting. **Bottom Row**, These plots show whether the GPSMs, on average, have higher order marginals (HOMs) that match those of their respective targets between orders of two and ten. Insets emphasize pairwise r_{20}. Pearson r_{20} correlation (r_{20}, y-axis) is plotted as a function of HOMs (x-axis). Synthetic r_{20} were computed from 6M artificial GPSM-generated evaluation sequences compared to 6M synthetic target sequences (**d**, **e**) as colored lines, and the black dotted line here in the synthetic r_{20} analysis denotes a generative capacity upper limit set by finite sampling of two non-overlapping target MSAs of 6M sequences each. Both Mi3 and vVAE's generative capacity appear sensitive to decreasing synthetic training sample size for this metric. In contrast, for the natural analysis, the black line uses triangles to indicate a difference in how it is computed compared to the synthetic analysis (**f**). Here, the generative capacity upper limit is an estimate computed between non-overlapping evaluation MSAs of 6M and 10K sequences from the artificial Mi3 distribution trained on natural sequences (**f**, cyan), and only orders two through seven were plotted due to finite sampling effects at higher orders. In the natural analysis, Mi3 performs as close to the target as is measurable, within error, for this amount of data.

estimation error negligible. vVAE covariances are further scaled down slightly when fit to the synthetic 10K dataset ($\rho = 0.87$, Fig. 2b). Indep cannot reproduce covariances by definition, so ρ is zero in all tests, as expected. The generative capacity trends for this metric are consistent between the synthetic and natural analyses for all GPSMs, showing the behavior is not due to artificial properties of our synthetic target model.

These results confirm that vVAE can model epistasis in protein sequence datasets, since it generates pairwise mutational covariances which are correlated with the target values, even in the absence of explicit constraints for reproducing these statistics. However, it scales down the strength of pairwise covariances in both the synthetic and natural analyses and the correlation with the target is lower than 1. Mi3, in contrast, is constrained by design to fit the covariance scores and does so nearly perfectly.

Higher order marginal statistics

A more stringent test of GPSM generative capacity is to measure the model's ability to reproduce sequence covariation involving more than two positions, or higher-order covariation. We characterize these higher-order covariation patterns in the target MSA and GPSM-generated evaluation MSAs by computing the frequency of non-contiguous amino acid n-tuples, or higher-order marginals (HOMs) corresponding to subsequences, and compare their frequency in each MSA to corresponding values in the target MSAs. For increasing values of n the number of possible n-tuple combinations increases rapidly, requiring increasingly large evaluation MSAs to accurately estimate the frequency of individual n-tuples. For this reason, we limit n-tuple length to $n \leq 10$ and only compute a limited subset of all possible position sets for each n. For each n we randomly choose 3K position sets, compute the frequencies of the top twenty most frequent n-tuples for each corresponding position set in the target and evaluation MSAs, as these are well sampled, and for each position set compute the Pearson correlation between these top twenty frequencies. We then average the correlation values for each n over all position sets. We call this metric the Pearson correlation r_{20}.[13] In this test, estimation error is non-zero because of the extremely large MSAs required to compute n-tuple frequencies, particularly for high $n > 5$ (see Supplementary Information). We expect the estimation error caused by finite sampling in the evaluation MSAs by computing the r_{20} scores between two non-overlapping MSAs generated by the synthetic target model, which are of the same size as our evaluation MSAs.

In Fig. 2, Bottom Row, we plot the HOM r_{20} for varying n. The expected estimation error (black line) represents a generative capacity upper bound, giving the highest measurable r_{20} given the evaluation MSA size of 6M for the synthetic analysis and 10K for the natural analysis. The r_{20} for Mi3 fit to 1M training sequences is very close to the validation upper-bound for all n, suggesting it has accurately fit the synthetic target distribution and its specification error is close to zero (Fig. 2d). This is expected since the synthetic target model in this test is a Potts model. With 10K training sequences, Mi3 r_{20} scores are lower than the 1M result for all n (Fig. 2e), which illustrates that Mi3 is affected by out-of-sample error for typical dataset sizes, as previously described.[45] Indep has much lower r_{20} scores than Mi3, as expected, since it does not model pairwise epistasis by design. Its r_{20} scores are similar across all experiments, suggesting that it is not strongly affected by out-of-sample error. This is expected because its parameters are optimized for reproducing single-site frequency statistics only, which can be accurately estimated even from small training MSAs.[45] The r_{20} score for vVAE lies between Mi3 and Indep for all training datasets and n. With 1M training sequences, vVAE r_{20} decreases to 0.4 for $n = 10$, reflecting specification error (Fig. 2a). With 10K synthetic training sequences, vVAE r_{20} decreases further for all n due to the addition of out-of-sample error (Fig. 2e).

Whereas the pairwise covariation correlations represent a preliminary indication that vVAE captures epistasis but mispredicts its strength, the r_{20} results reinforce this finding and extend it into higher orders. Unlike Mi3, vVAE shows specification error even when fit to large datasets from a model which only contains pairwise epistatic interaction terms (Fig. 2d). Because higher-order covariation statistics are constrained by the pairwise statistics, and vVAE mispredicts the pairwise statistics, we expect that vVAE will exhibit specification error for higher-order epistasis, even though our synthetic test does not address this issue directly. When considered together with the r_{20}, the pairwise covariance correlations reveal a novel insight, which is that that when trying to gauge GPSM generative capacity at higher orders of covariation, the pairwise statistics alone can be misleading. The relative magnitudes of r_{20} between models at $n = 2$ are different at higher n, and the performance decrease for high n is more severe for vVAE than Mi3 (Fig. 2, Bottom Row). We emphasize that the r_{20} performance decrease for Mi3 when tested against both synthetic and natural MSA targets can be accounted for entirely by out-of-sample error.

Hamming distance distributions

The next metric we use to characterize the generative capacity of the models is the pairwise Hamming distance distribution $d(S, S')$. The Hamming distance between two protein sequences is the number of amino acids that are different between them, and we obtain a distribution for an MSA by comparing all pairs of sequences. Because it characterizes the amount of sequence diversity in an MSA, recapitulation of the Hamming distance distribution has been used in the past as a measure for GPSM performance.[2,13,36,37] In Fig. 3, we compare the pairwise Hamming distance distribution for each GPSM to that of the target distribution, computed with evaluation and target MSAs of 10K sequences each. To quantify the difference between the GPSM and target distributions for this metric, we use the total variation distance (TVD),[46] which equals 1 when the distributions have no overlap and is 0 when they are identical, defined by $\mathrm{TVD}[f,g] = 1/2\int|f(x)-g(x)|dx$.

All models reproduce the mode Hamming distance of ~ 179. For Mi3, we report the same $\mathrm{TVD} = 0.007$ when trained on either 1M (Fig. 3a) or 10K (Fig. 3b) synthetic sequences, showing no specification error, as expected. When trained on 10K natural sequences, Mi3 TVD increases to 0.012 (Fig. 3c), for reasons discussed further below. Indep severely underestimates the probability of both low and high Hamming distances, as observed at the distribution tails, with $\mathrm{TVD} \sim 0.24$ across all experiments. vVAE performs in between Mi3 and Indep, but much closer to Mi3 than Indep with respect to TVD. Performance differences across all GPSMs for this metrics appear to indicate that out-of-sample error has a consistent and detectable, though negligible, effect on the fundamental sequence diversity of artificial GPSM-generated MSAs. That Mi3 and vVAE are highly performant and comparable to each other, but not Indep, corroborates our earlier findings that epistasis is relevant to accurate modelling of protein sequence diversity (Fig. 2). However, because Indep performs much closer to Mi3 and vVAE for this metric than on any other, and also because this metric cannot discriminate well between Mi3 and vVAE, it could mean that reproducing the Hamming distance distribution is a much easier, and perhaps separate, hurdle for GPSMs, than is reproducing higher-order covariation.

To emphasize the decay of the tails, we rescale all the distributions by their maxima and re-center them around their modes to give them the same peak, and then plot them on a log-log scale (Fig. 3, Bottom Row). The relevance of the distribution's tails lies in their power-law behavior as they approach zero, where the function's exponent is related to the intrinsic dimension of the dataset and therefore to the number of informative latent factors needed to explain the data.[36,37,47] A well-specified GPSM ought to reproduce this exponent, and therefore the tail's decay, since it is a topological property intrinsic to the dataset and independent from the particular choice of variables used to describe the probability density.[47] The trend of slightly decreasing generative capacity as training samples decrease, as observed in Fig. 3, Top Row, is emphasized in Fig. 3, Bottom Row. In this rendering of the Hamming distance distribution, differences in GPSM generative capacity can be observed at both low (Left Tail) and high (Right Tail) sequence diversity. The Mi3 distribution closely overlaps the target distribution with both 1M (Fig. 3d) and 10K (Fig. 3e) synthetic training sequences. In the 10K natural experiment, Mi3 deviates noticeably from the target on the left tail (Fig. 3f), which represents less evolutionarily diverged sequences. This could be an artifact of the phylogenetic relationships between sequences present in the natural dataset, which may have been incompletely removed by our phylogenetic filtering step for this data (see Methods, Supplementary Information), or it could be due to estimation error in measuring the target distribution, as only 10K target sequences are available to estimate the black line in the natural analysis. As before, Indep performance is consistently low across all experiments. vVAE performance at low sequence diversity (Fig. 3, Bottom Row, Left Tails) decreases for smaller training dataset size.

Statistical Energy Correlations

A fourth metric we use to evaluate generative capacity is statistical energy $E(S)$ of individual sequences in the dataset, which we express using the negative logarithm of the predicted sequence probability $p(S)$, where $E(S) = -\log p(S)$. $E(S)$ can be computed analytically for Mi3 and Indep, and estimated for vVAE by importance sampling (see Methods).

This statistic directly evaluates accuracy of the GPSM distribution values $p_\theta(S)$ for a limited number of individual sequences, which has been used to validate GPSMs by comparison to corresponding experimental fitness values.[13,19,48,49] In Fig. 4, we compare artificial statistical energies from the GPSM distribution $p_\theta(S)$ to those of the target distribution $p^0(S)$ for a 1K test MSA generated from $p^0(S)$. As before, we use the 1M (Fig. 4, Left Column) and 10K (Fig. 4, Right Column) training MSA sizes, and quantify GPSM generative capacity for this metric by the Pearson correlation coefficient $\rho(\{E(S)\},\{\hat{E}(S)\})$ between target energies $E(S)$ and GPSM energies $\hat{E}(S)$. Mi3 reproduces the synthetic target distribution at both training MSA sizes. Because Mi3 should have very low specification error on the synthetic target, as it is well specified by design, the small amount of error must be due to remaining out-of-sample or numerical errors. As expected, Indep poorly reproduces the target values, with $\rho = 0.6$ for both MSA training sizes. vVAE exhibits greater specification error than Mi3 on the 1M training set with correlation of $\rho = 0.94$, and exhibits further out-of-sample error on the 10K training set with $\rho = 0.89$.

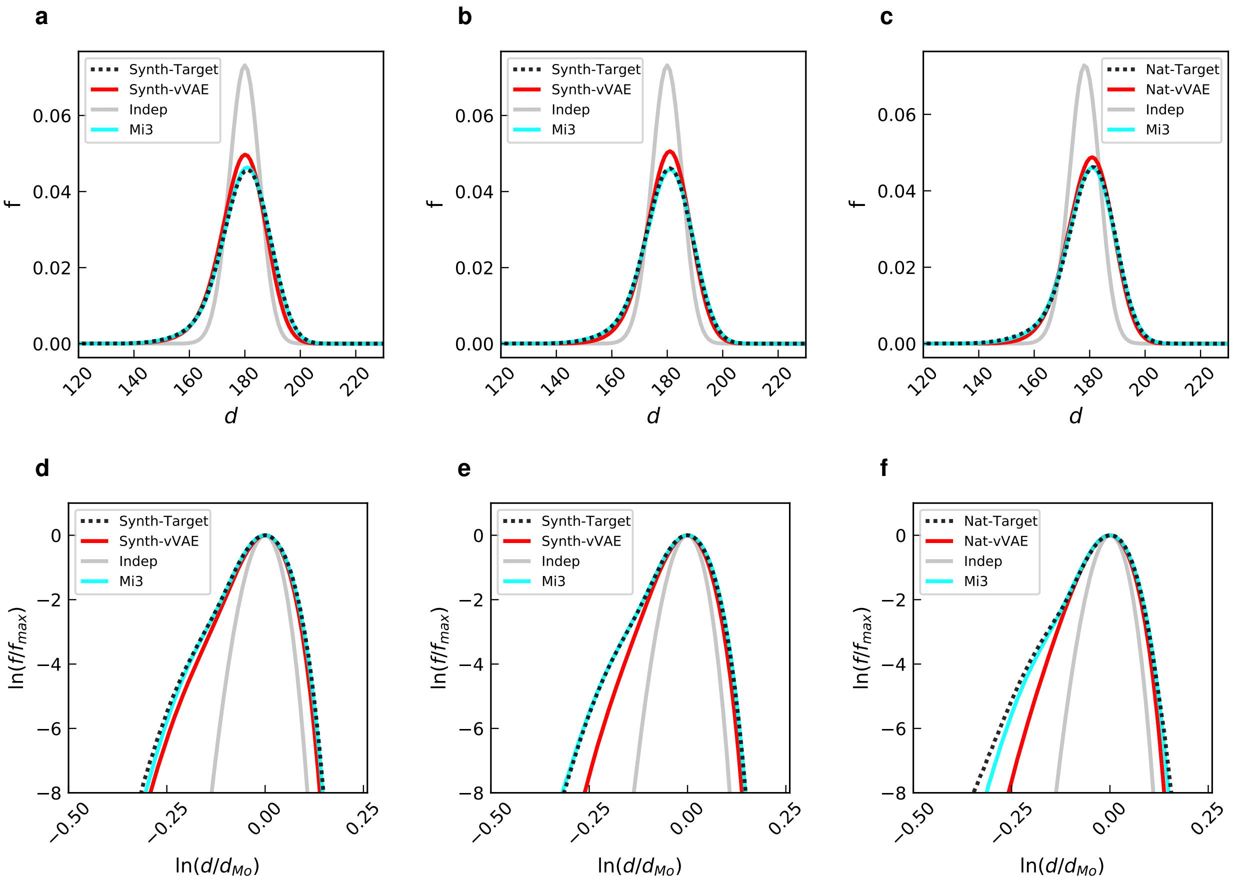

Figure 3. Pairwise Hamming distance distributions. These plots illustrate whether sequences generated from the GPSMs reproduce the overall sequence diversity of their respective targets. Mi3 (cyan), vVAE (red), and Indep (gray) distributions are compared to target distribution (dotted black). GPSMs were trained on 1M synthetic (**a**, **d**), 10K synthetic (**b**, **e**), or 10K natural (**c**, **f**) sequences from the corresponding target distribution. All Hamming distributions were computed from 30K-sequence MSAs, except for the natural target, which was computed from a 10K-sequence target MSA due to data limitations. **Top Row**, Hamming distances d (x-axis) are shown about the mode, and frequency f is normalized as a fraction of total (y-axis). Mi3 perfectly matches the target distribution, and vVAE performs closely to Mi3. **Bottom Row**, Re-scaled logarithmic Hamming distance distributions better discriminate between GPSMs with respect to generative capacity than the normal Hamming distance distribution. Before being log-scaled, the Hamming distances d are normalized by the mode d_{Mo} (x-axis), and frequencies f are normalized by the maximum Hamming distance f_{max} (y-axis). This transformation highlights minute differences between distributions at low frequencies in the tails of the distributions on the left- and right-hand sides. Only vVAE appears sensitive to training sample size for this metric at the log-log scale.

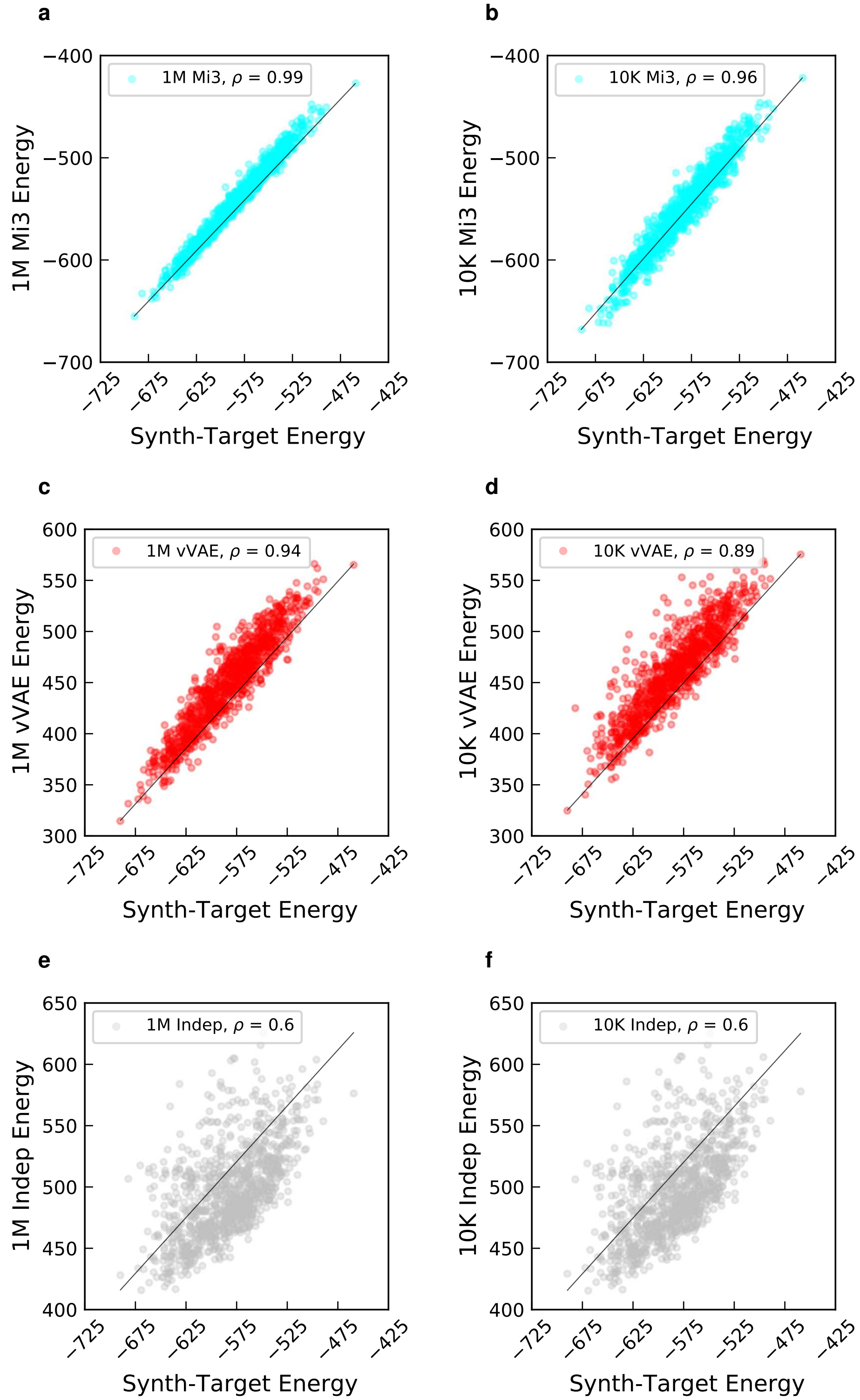

Figure 4. Statistical energy correlations. Statistical energies $E(S)$ of 1K synthetic test sequences from the target distribution as evaluated by Mi3 (cyan), vVAE (red), Indep (gray). Each GPSM was trained on 1M (**a**, **c**, **e**) or 10K (**b**, **d**, **f**) sequences from the synthetic target distribution. For each scatterplot, Pearson correlation coefficient ρ was computed between each GPSM's statistical energy and that of the synthetic target distribution for each sequence. Only Indep is insensitive to decreased training sample size for this metric.

Conclusions

In this study we reveal the steep challenges, limits, and errors faced by contemporaneous GPSMs trained on either synthetic or currently available natural sequence datasets. Recent state-of-the-art GPSM studies have benchmarked their models by comparing $p_\theta(S)$ to experimental fitness values from deep mutational scans,[19,21] or by generating artificial sequences that appear to fold into realistic structures based on *in silico* folding energy.[24] However, these strategies for model comparison present their own challenges. Point mutation fitness experiments in practice may measure particular contributors to fitness, including replicative capacity, drug resistance, protein stability, and enzymatic activity,[19] and are subject to significant experimental error and other limitations, e.g. those imposed by conservation.[50] On the other hand, *in silico* computational chemistry approaches rely on energy functionals that may lack the precision needed to meaningfully discriminate between highly similar sequences.[51] Additionally, neither protein function nor fitness rely exclusively on the thermodynamic stability of the static native structure, but more so on the protein's conformational dynamics,[52–56] which are not fully described by folding energy values alone.[57] This could mean that despite generating sequences with realistic *in silico* folding energy, a GPSM may still not be capturing crucial higher-order epistatic effects. Neither point mutation fitness effects, nor *in silico* folding energy estimations, are directly related to mutational covariation statistics observed in an MSA, and thus they are indirect metrics of GPSM statistical accuracy. Our comparison of higher-order covariation patterns between target MSAs and GPSM-generated evaluation MSAs is a direct measurement of GPSM statistical accuracy, and our study emphasizes higher-order covariation whereas previous studies rarely go beyond the pairwise level.[2,13,34,35,58]

Benchmarking coevolution-based protein sequence models in data rich and data poor regimes, as done here, is an effective method for ascertaining where data-driven effects stop, and algorithmic failure begins.[39] In our synthetic analysis track, we have demonstrated the extent to which a vanilla VAE (vVAE) can capture higher-order covariation at orders between three and ten when the target distribution is known, its statistical properties are measurable with a high degree of certainty, and major forms of error are removed, minimized, or accounted for. When given a large number of training and target sequences, we found vVAE's generative capacity to be between that of a site-independent model (Indep) and a pairwise Hamiltonian (Mi3) for all measurements. In the synthetic r_{20} tests, our results show that vVAE generative capacity is well below Mi3, raising questions about whether vanilla VAEs can capture higher-order epistasis significantly better than a pairwise Potts model. In our synthetic analysis the target distribution is a Potts model, and therefore we expect Mi3 to fit the target distribution well by design. However, we find Mi3 also outperforms vVAE where we do not have this expectation, such as on the natural target distribution and on a target distribution specified by a vVAE, as shown in Supplementary Information, suggesting Mi3 generally outperforms vVAE on protein sequence data.

The Hamming distance distributions, pairwise covariation correlations, and statistical energy correlations have been used in the past to measure GPSM accuracy, but we find that they can be inadequate or misleading indicators of a GPSM's ability to capture covariation at higher orders. Taken together, our results suggest that, of the metrics we tested, only r_{20} provides the granularity needed to discriminate between different GPSM's ability to model higher-order epistasis, and this is because r_{20} directly tests the model's ability to capture higher-order covariation.

Comparative benchmarking of our r_{20} results between the natural and synthetic analyses has shown for the first time how inherent limitations of natural sequence data obfuscate GPSM performance with respect to higher-order covariation. Our results also show that there are too few sequences in typical natural datasets to accurately gauge the generative capacity of a GPSM, without also performing a synthetic test. In our natural analysis, both the GPSM training and evaluation processes are data-starved to the degree typical of publicly available protein sequence datasets, allowing out-of-sample and estimation error to frustrate efforts to discriminate between model performance, giving a glimpse into the scale of error that could be present in the results of recently published GPSM studies.

Although our results suggest VAE-GPSMs are less effective for capturing higher-order epistasis than pairwise Potts models, they have demonstrated utility in unsupervised learning and clustering. One VAE-GPSM has generated artificial sequences that share a "hallucinated" homology to natural proteins in the training set, which could mean that their folded structures would perform similar functions to their hallucinated natural homologs.[22] Another has shown that a VAE-GPSM's latent space may capture phylogenetic relationships[20] better than PCA[59] and t-SNE.[60] These VAE-GPSMs furnished a latent space that immediately allowed for function-based protein classification, a benefit unavailable to pairwise Potts models without some effort.

Our innovative epistasis-oriented methodology focuses on measuring higher-order covariation, with the potential for broad applicability to various sequentially ordered data. r_{20}-like measurements become possible when the data are sufficient in number, and the correlation structures between elements, both within and across samples, are statistically detectable and meaningful in some context, be it visual, biophysical, or linguistic. The convergence between data categories such as images, proteins, and language with respect to generative modelling evaluation offers the exciting opportunity of a wider, interdisciplinary audience for the work proposed here. Conversely, further development of sophisticated, data-intensive, and direct generative capacity metrics of GPSMs could reveal nuances of the correlation structure of protein sequence datasets that distinguish them from other datasets, helping to explain why r_{20}-like metrics can detect higher-order covariation in artificial MSAs, whereas other metrics cannot. Our work represents not only a revision of currently prevailing paradigms of GPSM benchmarking, but also a challenge to generative modelling more broadly, to consider how higher-order covariation metrics and epistasis can inform their models and results.

Methods

Sequence dataset preparation

For the natural analysis, we use an MSA of the kinase protein superfamily which we have previously curated using sequences from the Uniprot/TREMBL database.[40] This MSA is composed of $\sim$20K sequences of length 232, obtained by filtering a larger set of $\sim$291K sequences to remove any sequences with more than 50% sequence identity to another (see Supplementary Information).

For the synthetic analysis, synthetic sequences are generated from an "original" Potts Hamiltonian model trained on the natural protein-kinase MSA, as detailed below and shown in Fig.1, Stage 1. This synthetic MSA is considered to be the target for the synthetic analysis. To clarify, our focus in this work is to quantify model generative capacity against the target distribution, and not the training distribution itself. Therefore, we have ensured that our synthetic training and target sequences are non-overlapping sets, even though they come from the same target model. The statistical differences between the two sets are nontrivial, and have been detected experimentally by our measurements (see Results).

In Stage 2, we process the sequences into a train-validation-test split. The processed MSAs are then one-hot encoded and fed into the GPSMs in Stage 3, where training occurs. In Stage 4, we generate evaluation MSAs from each model, and in Stage 5 we perform our generative capacity measurements by comparing the artificial MSAs to the appropriate target MSA.

Mi3

The Mi3 model is a pairwise Potts Hamiltonian model fit to sequence data using the "Mi3-GPU" software we have developed previously,[32] which performs "inverse Ising inference" to infer parameters of Potts models using a Markov-Chain Monte-Carlo (MCMC) algorithm which entails very few approximations. This software allows us to fit statistically accurate Potts models to MSA data. We have examined Mi3's generative capacity and out-of-sample error in earlier work,[32,45] which we summarize here.

A Potts model is the maximum entropy model for $p(S)$ constrained to reproduce the bivariate marginals $f^{ij}_{\alpha\beta}$ of an MSA, i.e. the frequency of amino acid combination α, β ,at positions i, j. The probability distribution $p_\theta(S)$ for the Potts model takes the form

$$p_\theta(S) = \frac{e^{-E(S)}}{Z} \quad \text{with} \quad E(S) = \sum_i^L h^i_{s_i} + \sum_{i<j} J^{ij}_{s_i s_j} \tag{1}$$

where Z is a normalization constant, $Z = \sum_S e^{-E(S)}$, and "coupling" $J^{ij}_{\alpha\beta}$ and "field" h^i_α parameters are compactly denoted by the vector $\theta = \{h^i_\alpha, J^{ij}_{\alpha\beta}\}$. The number of free parameters of the model (non-independent couplings and fields) is equal to the number of non-independent bivariate marginals, which can be shown to be $\frac{L(L-1)}{2}(q-1)^2 + L(q-1)$ for q amino acids,[15] which is 10.7M parameters for our model. This implies that the Potts model is well specified to reproduce the bivariate marginals when generating sequences from $p_\theta(S)$.

The Mi3 model inference procedure maximizes the log-likelihood with regularization. Maximizing the Potts log-likelihood can be shown to be equivalent to minimizing the difference between the dataset MSA bivariate marginals $f^{ij}_{\alpha\beta}$ and the model bivariate marginals of sequences generated from $p(S)$. To account for finite sampling error in the estimate of $f^{ij}_{\alpha\beta}$ for an MSA of N sequences we add a small pseudocount of size $1/N$, as described previously.[45] We also add a regularization penalty to the likelihood affecting the coupling parameters $J^{ij}_{\alpha\beta}$ to bias them towards 0, of form $\lambda \sum \mathrm{SCAD}(J^{ij}_{\alpha\beta}, \lambda, \alpha)$ using the SCAD function which behaves like $\lambda |J^{ij}_{\alpha\beta}|$ for small $J^{ij}_{\alpha\beta}$ but gives no bias for large $J^{ij}_{\alpha\beta}$[61] , using a small regularization strength of $\lambda = 0.001$ for all inferences which causes little model bias.

To generate synthetic MSAs from the Potts model we use MCMC over the trial distribution $p_\theta(S)$ until the Markov-Chains reach equilibrium, as described previously.[32] We can directly evaluate $E(S)$ the negative log-probability of any sequences for the Mi3 model using equation 1 up to a constant Z, and this constant can be dropped without affecting our results.

Indep

The Indep model is the maximum entropy model for $p(S)$ constrained to reproduce the univariate marginals of an MSA, and is commonly called a "site-independent" model because the sequence variations at each site are independent of the variation at other sites. Because it does not fit the bivariate marginals, it cannot model covariation between positions. It takes the form

$$p_\theta(S) = \frac{e^{-E(S)}}{Z} \quad \text{with} \quad E(S) = \sum_i^L h^i_{s_i} \tag{2}$$

where Z is a normalization constant, $Z = \sum_S e^{-E(S)}$, and "field" parameters h^i_α for all positions i and amino acids α are compactly referred to by the vector $\theta = \{h^i_\alpha\}$. The fields of the Indep model generally have different values from the fields of the Potts model. Unlike for the Potts model, maximum likelihood parameters can be determined analytically to be $h^i_\alpha = -\log f^i_\alpha$ where f^i_α are the univariate marginals of the dataset MSA. When fitting the Indep model to a dataset MSA of N sequences, we add a pseudocount of $1/N$ to the univariate marginals to give model marginals $\hat{f}^i_\alpha$, to account for finite sample error in the univariate marginal estimates. The model distribution simplifies to a product over positions, as $p_\theta(S) = \sum \hat{f}^i_{s_i}$. The number of independent field parameters is $L(q-1)$ which equals 4.6K parameters for our model.

To generate sequences from the independent model we independently generate the residues at each position i by a weighted random sample from the marginals f^i_α, and we directly evaluate the log probability of each sequence $E(S)$ from equation 2.

vVAE

The vanilla variational autoencoder (vVAE) is a deep, symmetrical, and undercomplete autoencoder neural network composed of a separate encoder $q_\phi(Z|S)$ and decoder $p_\theta(S|Z)$,[62] which map input sequences S to regions within a low-dimensional latent space Z and back. The probability distribution for the vVAE is defined as

$$p_\theta(S) = \int p_\theta(S|Z)p(Z)dZ \tag{3}$$

where the latent space distribution is a unit Normal distribution, $p(Z) = \mathcal{N}[0,1](Z)$. Training of a VAE can be understood as maximization of the dataset log-likelihood with the addition of a Kullback-Leibler regularization term $\mathrm{D}_{\mathrm{KL}}[q_\phi(Z|S), p_\theta(Z|S)]$, where $p_\theta(Z|S)$ is the posterior of the decoder.[28,29]

Our VAE architecture is intentionally "vanilla",[33] and our encoder and decoder use a simple architecture unlike more sophisticated VAE implementations which use convolutional layers,[19] multi-stage training,[63] disentanglement learning,[33] Riemannian Brownian motion priors,[64] and more. It reflects a simple VAE architecture that has been implemented as a VAE-GPSM in prior work.[21] This allows us to directly interrogate the assumptions and performance of vanilla variational autoencoding with respect to the training and evaluation of GPSMs in this work.

Our encoder and decoder have 3 layers each and employ minimal standard normalization and regularization strategies.[65,66] vVAE's latent bottleneck layer has 7 nodes, and the model in total has 2.7M inferred parameters. The input layer of the encoder accepts a one-hot encoded sequence and the decoder's output layer values can be interpreted as a Bernoulli distribution of the same dimensions as a one-hot encoded sequence. We have tested various vanilla VAE architectures and hyperparameters with our datasets, as well as DeepSequence,[19] and found qualitatively similar generative capacity results.

To generate a sequence from vVAE, we generate a random sample in latent space from the latent distribution $p(Z)$ pass this value to the decoder to obtain a Bernoulli distribution, from which we sample once. To evaluate the negative log-probability of a sequence $E(S)$ we use importance sampling, averaging over 1K samples from the latent distribution $q_\phi(Z|S)$.[20] Other publications use the Evidence Lower Bound (ELBO) estimate as an approximation of the negative log-probability,[19] and we have verified that the ELBO and the negative log-probability are nearly identical in our tests and have equal computational complexity.

References

1. Levy, R. M., Haldane, A. & Flynn, W. F. Potts hamiltonian models of protein co-variation, free energy landscapes, and evolutionary fitness. *Current Opinion in Structural Biology* **43**, 55–62 (2017).
2. Cocco, S., Feinauer, C., Figliuzzi, M., Monasson, R. & Weigt, M. Inverse statistical physics of protein sequences: a key issues review. *Reports on Progress in Physics* **81**, 032601 (2018). Publisher: IOP Publishing.
3. Tubiana, J., Cocco, S. & Monasson, R. Learning compositional representations of interacting systems with restricted boltzmann machines: Comparative study of lattice proteins. *Neural Computation* **31**, 1671–1717 (2019). URL https://doi.org/10.1162/neco_a_01210.
4. Lapedes, A. S., Giraud, B., Liu, L. & Stormo, G. D. *Correlated mutations in models of protein sequences: phylogenetic and structural effects* (Institute of Mathematical Statistics, 1999). Pages: 236-256 Publication Title: Statistics in molecular biology and genetics.
5. Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein–protein interaction by message passing. *Proceedings of the National Academy of Sciences* **106**, 67–72 (2009). Publisher: National Academy of Sciences Section: Physical Sciences.
6. Haldane, A., Flynn, W. F., He, P., Vijayan, R. S. K. & Levy, R. M. Structural propensities of kinase family proteins from a Potts model of residue co-variation. *Protein Science: A Publication of the Protein Society* **25**, 1378–1384 (2016).
7. Domingo, J., Baeza-Centurion, P. & Lehner, B. The Causes and Consequences of Genetic Interactions (Epistasis). *Annual Review of Genomics and Human Genetics* **20**, 433–460 (2019). Publisher: Annual Reviews.
8. Noel, J. K., Morcos, F. & Onuchic, J. N. Sequence co-evolutionary information is a natural partner to minimally-frustrated models of biomolecular dynamics. *F1000Research* **5** (2016). URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4755392/.
9. Morcos, F., Jana, B., Hwa, T. & Onuchic, J. N. Coevolutionary signals across protein lineages help capture multiple protein conformations. *Proceedings of the National Academy of Sciences* **110**, 20533–20538 (2013). URL https://www.pnas.org/content/110/51/20533. Publisher: National Academy of Sciences Section: Biological Sciences.
10. Sułkowska, J. I., Morcos, F., Weigt, M., Hwa, T. & Onuchic, J. N. Genomics-aided structure prediction. *Proceedings of the National Academy of Sciences* **109**, 10340–10345 (2012). URL https://www.pnas.org/content/109/26/10340. Publisher: National Academy of Sciences Section: Biological Sciences.
11. Hopf, T. A. *et al.* Mutation effects predicted from sequence co-variation. *Nature biotechnology* **35**, 128–135 (2017).
12. Biswas, A., Haldane, A., Arnold, E. & Levy, R. M. Epistasis and entrenchment of drug resistance in HIV-1 subtype B. *eLife* **8**, e50524 (2019). Publisher: eLife Sciences Publications, Ltd.
13. Haldane, A., Flynn, W. F., He, P. & Levy, R. M. Coevolutionary Landscape of Kinase Family Proteins: Sequence Probabilities and Functional Motifs. *Biophysical Journal* **114**, 21–31 (2018).
14. Figliuzzi, M., Barrat-Charlaix, P. & Weigt, M. How Pairwise Coevolutionary Models Capture the Collective Residue Variability in Proteins? *Molecular Biology and Evolution* **35**, 1018–1027 (2018). URL https://doi.org/10.1093/molbev/msy007. https://academic.oup.com/mbe/article-pdf/35/4/1018/24597926/msy007.pdf.
15. Bialek, W. & Ranganathan, R. Rediscovering the power of pairwise interactions. *arXiv:0712.4397 [q-bio]* (2007). ArXiv: 0712.4397.
16. Weinreich, D. M., Lan, Y., Wylie, C. S. & Heckendorn, R. B. Should evolutionary geneticists worry about higher-order epistasis? *Current opinion in genetics & development* **23**, 700–707 (2013).
17. Haq, O., Andrec, M., Morozov, A. V. & Levy, R. M. Correlated electrostatic mutations provide a reservoir of stability in hiv protease. *PLoS Comput Biol* **8**, e1002675EP (2012).
18. Haq, O., Levy, R., Morozov, A. & Andrec, M. Pairwise and higher-order correlations among drug-resistance mutations in hiv-1 subtype b protease. *BMC Bioinformatics* **10**, S10 (2009).
19. Riesselman, A. J., Ingraham, J. B. & Marks, D. S. Deep generative models of genetic variation capture the effects of mutations. *Nature Methods* **15**, 816–822 (2018).

20. Ding, X., Zou, Z. & Brooks Iii, C. L. Deciphering protein evolution and fitness landscapes with latent space models. *Nature Communications* **10**, 5644 (2019). Number: 1 Publisher: Nature Publishing Group.

21. Sinai, S., Kelsic, E., Church, G. M. & Nowak, M. A. Variational auto-encoding of protein sequences. *arXiv:1712.03346 [cs, q-bio]* (2018). ArXiv: 1712.03346.

22. Costello, Z. & Martin, H. G. How to Hallucinate Functional Proteins. *arXiv:1903.00458 [q-bio]* (2019). ArXiv: 1903.00458.

23. Gupta, A. & Zou, J. Feedback GAN for DNA optimizes protein functions. *Nature Machine Intelligence* **1**, 105–111 (2019). Number: 2 Publisher: Nature Publishing Group.

24. Madani, A. *et al.* ProGen: Language Modeling for Protein Generation. *arXiv:2004.03497 [cs, q-bio, stat]* (2020). ArXiv: 2004.03497.

25. Vig, J. *et al.* BERTology Meets Biology: Interpreting Attention in Protein Language Models. *bioRxiv* 2020.06.26.174417 (2020). Publisher: Cold Spring Harbor Laboratory Section: New Results.

26. Elnaggar, A. *et al.* ProtTrans: Towards Cracking the Language of Life's Code Through Self-Supervised Deep Learning and High Performance Computing. *bioRxiv* 2020.07.12.199554 (2020). Publisher: Cold Spring Harbor Laboratory Section: New Results.

27. Choromanski, K. *et al.* Rethinking Attention with Performers. *arXiv:2009.14794 [cs, stat]* (2020). URL http://arxiv.org/abs/2009.14794. ArXiv: 2009.14794.

28. Kingma, D. P. & Welling, M. Auto-Encoding Variational Bayes. *arXiv:1312.6114 [cs, stat]* (2014). URL http://arxiv.org/abs/1312.6114. ArXiv: 1312.6114.

29. Rezende, D. J., Mohamed, S. & Wierstra, D. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *arXiv:1401.4082 [cs, stat]* (2014). ArXiv: 1401.4082.

30. Luce, R. D., Bush, R. R. & Galanter, E. (eds.) *Handbook of mathematical psychology: I.* Handbook of mathematical psychology: I. (John Wiley, Oxford, England, 1963). Pages: xiii, 491.

31. Frawley, W. J. *International Encyclopedia of Linguistics* (Oxford University Press, 2003). Publication Title: International Encyclopedia of Linguistics.

32. Haldane, A. & Levy, R. M. Mi3-GPU: MCMC-based inverse Ising inference on GPUs for protein covariation analysis. *Computer Physics Communications* 107312 (2020).

33. Ding, Z. *et al.* Guided Variational Autoencoder for Disentanglement Learning. *arXiv:2004.01255 [cs]* (2020). ArXiv: 2004.01255.

34. Shimagaki, K. & Weigt, M. Selection of sequence motifs and generative Hopfield-Potts models for protein families. *Physical Review E* **100**, 032128 (2019). URL https://link.aps.org/doi/10.1103/PhysRevE.100.032128. Publisher: American Physical Society.

35. Hawkins-Hooker, A. *et al.* Generating functional protein variants with variational autoencoders. *bioRxiv* 2020.04.07.029264 (2020). Publisher: Cold Spring Harbor Laboratory Section: New Results.

36. Facco, E., Pagnani, A., Russo, E. T. & Laio, A. The intrinsic dimension of protein sequence evolution. *PLOS Computational Biology* **15**, e1006767 (2019). Publisher: Public Library of Science.

37. Granata, D. & Carnevale, V. Accurate Estimation of the Intrinsic Dimension Using Graph Distances: Unraveling the Geometric Complexity of Datasets. *Scientific Reports* **6**, 31377 (2016).

38. El-Gebali, S. *et al.* The Pfam protein families database in 2019. *Nucleic Acids Research* **47**, D427–D432 (2019). Publisher: Oxford Academic.

39. AlQuraishi, M. ProteinNet: a standardized data set for machine learning of protein structure. *BMC Bioinformatics* **20**, 311 (2019). URL https://doi.org/10.1186/s12859-019-2932-0.

40. Consortium, T. U. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Research* **47**, D506–D515 (2019). URL https://academic.oup.com/nar/article/47/D1/D506/5160987. Publisher: Oxford Academic.

41. Burnham, K. P. & Anderson, D. R. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach* (Springer-Verlag, New York, 2002), 2 edn.

42. Yaser S. Abu-Mostafa, Malik Magdon-Ismail, Hsuan-Tien Lin-Learning From Data_ A short course-AMLBook.com (2012).pdf | Statistical Classification | Machine Learning.

43. Everitt, B. S. & Skrondal, A. The Cambridge Dictionary of Statistics. *Cambridge University Press* 480 (2010).

44. Mohri, M., Rostamizadeh, A. & Talwalkar, A. *Foundations of Machine Learning, second edition* (MIT Press, 2018). Google-Books-ID: dWB9DwAAQBAJ.

45. Haldane, A. & Levy, R. M. Influence of multiple-sequence-alignment depth on Potts statistical models of protein covariation. *Physical Review. E* **99**, 032405 (2019).

46. Levin, D. A. & Peres, Y. *Markov Chains and Mixing Times* (American Mathematical Soc., 2017). Google-Books-ID: f208DwAAQBAJ.

47. Ansuini, A., Laio, A., Macke, J. H. & Zoccolan, D. Intrinsic dimension of data representations in deep neural networks. In *Advances in Neural Information Processing Systems*, 6111–6122 (2019).

48. Riesselman, A. *et al.* Accelerating Protein Design Using Autoregressive Generative Models. *bioRxiv* 757252 (2019). Publisher: Cold Spring Harbor Laboratory Section: New Results.

49. Figliuzzi, M., Jacquier, H., Schug, A., Tenaillon, O. & Weigt, M. Coevolutionary Landscape Inference and the Context-Dependence of Mutations in Beta-Lactamase TEM-1. *Molecular Biology and Evolution* **33**, 268–280 (2016).

50. Haddox, H. K., Dingens, A. S., Hilton, S. K., Overbaugh, J. & Bloom, J. D. Mapping mutational effects along the evolutionary landscape of HIV envelope. *eLife* **7**, e34420 (2018). Publisher: eLife Sciences Publications, Ltd.

51. Alley, E. C., Khimulya, G., Biswas, S., AlQuraishi, M. & Church, G. M. Unified rational protein engineering with sequence-based deep representation learning. *Nature Methods* **16**, 1315–1322 (2019). URL https://www.nature.com/articles/s41592-019-0598-1. Number: 12 Publisher: Nature Publishing Group.

52. Wei, G., Xi, W., Nussinov, R. & Ma, B. Protein Ensembles: How Does Nature Harness Thermodynamic Fluctuations for Life? The Diverse Functional Roles of Conformational Ensembles in the Cell. *Chemical Reviews* **116**, 6516–6551 (2016). URL https://doi.org/10.1021/acs.chemrev.5b00562. Publisher: American Chemical Society.

53. Anishchenko, I., Ovchinnikov, S., Kamisetty, H. & Baker, D. Origins of coevolution between residues distant in protein 3D structures. *Proceedings of the National Academy of Sciences* **114**, 9122–9127 (2017). URL https://www.pnas.org/content/114/34/9122. Publisher: National Academy of Sciences Section: Biological Sciences.

54. Sailer, Z. R. & Harms, M. J. Molecular ensembles make evolution unpredictable. *Proceedings of the National Academy of Sciences* **114**, 11938–11943 (2017). URL https://www.pnas.org/content/114/45/11938. Publisher: National Academy of Sciences Section: Biological Sciences.

55. Petrović, D., Risso, V. A., Kamerlin, S. C. L. & Sanchez-Ruiz, J. M. Conformational dynamics and enzyme evolution. *Journal of The Royal Society Interface* **15**, 20180330 (2018). URL https://royalsocietypublishing.org/doi/full/10.1098/rsif.2018.0330. Publisher: Royal Society.

56. Nussinov, R., Tsai, C.-J. & Jang, H. Protein ensembles link genotype to phenotype. *PLOS Computational Biology* **15**, e1006648 (2019). URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1006648. Publisher: Public Library of Science.

57. Campitelli, P., Modi, T., Kumar, S. & Ozkan, S. B. The Role of Conformational Dynamics and Allostery in Modulating Protein Evolution. *Annual Review of Biophysics* **49**, 267–288 (2020). Publisher: Annual Reviews.

58. Haq, O., Levy, R. M., Morozov, A. V. & Andrec, M. Pairwise and higher-order correlations among drug-resistance mutations in HIV-1 subtype B protease. *BMC Bioinformatics* **10**, S10 (2009).

59. Bishop, C. M. Pattern recognition and machine learning (2006). ISBN: 9781493938438 9780387310732 Library Catalog: cds.cern.ch Publisher: Springer.

60. Maaten, L. v. d. & Hinton, G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).

61. Fan, J. & Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360 (2001). URL https://doi.org/10.1198/016214501753382273. https://doi.org/10.1198/016214501753382273.

62. Charte, D., Charte, F., García, S., del Jesus, M. J. & Herrera, F. A practical tutorial on autoencoders for nonlinear feature fusion: Taxonomy, models, software and guidelines. *Information Fusion* **44**, 78–96 (2018).

63. Dai, B. & Wipf, D. Diagnosing and Enhancing VAE Models. *arXiv:1903.05789 [cs, stat]* (2019). URL http://arxiv.org/abs/1903.05789. ArXiv: 1903.05789.

64. Kalatzis, D., Eklund, D., Arvanitidis, G. & Hauberg, S. Variational Autoencoders with Riemannian Brownian Motion Priors. *arXiv:2002.05227 [cs, stat]* (2020). ArXiv: 2002.05227.

65. Ioffe, S. & Szegedy, C. Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. *arXiv:1502.03167 [cs]* (2015). ArXiv: 1502.03167.

66. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **15**, 1929–1958 (2014).

Acknowledgements

This research was supported by the National Science Foundation, grant number NSF-GCR 1934848, and the National Institute of Health, grant number R35-GM132090. This research includes calculations carried out on HPC resources supported in part by the National Science Foundation through major research instrumentation grant number 1625061 and by the US Army Research Laboratory under contract number W911NF-16-2-0189.

Author contributions statement

F.M., R.M.L, V.C., A.H. conceived the experiments. F.M., V.C., A.H. performed the experiments. F.M., R.M.L, V.C., A.H. analyzed the results. F.M., V.C., A.H. wrote the bulk of the codebase, Q.N. made a contribution to the codebase. F.M., R.M.L, V.C., A.H. wrote the paper. All authors reviewed the manuscript.

Figures

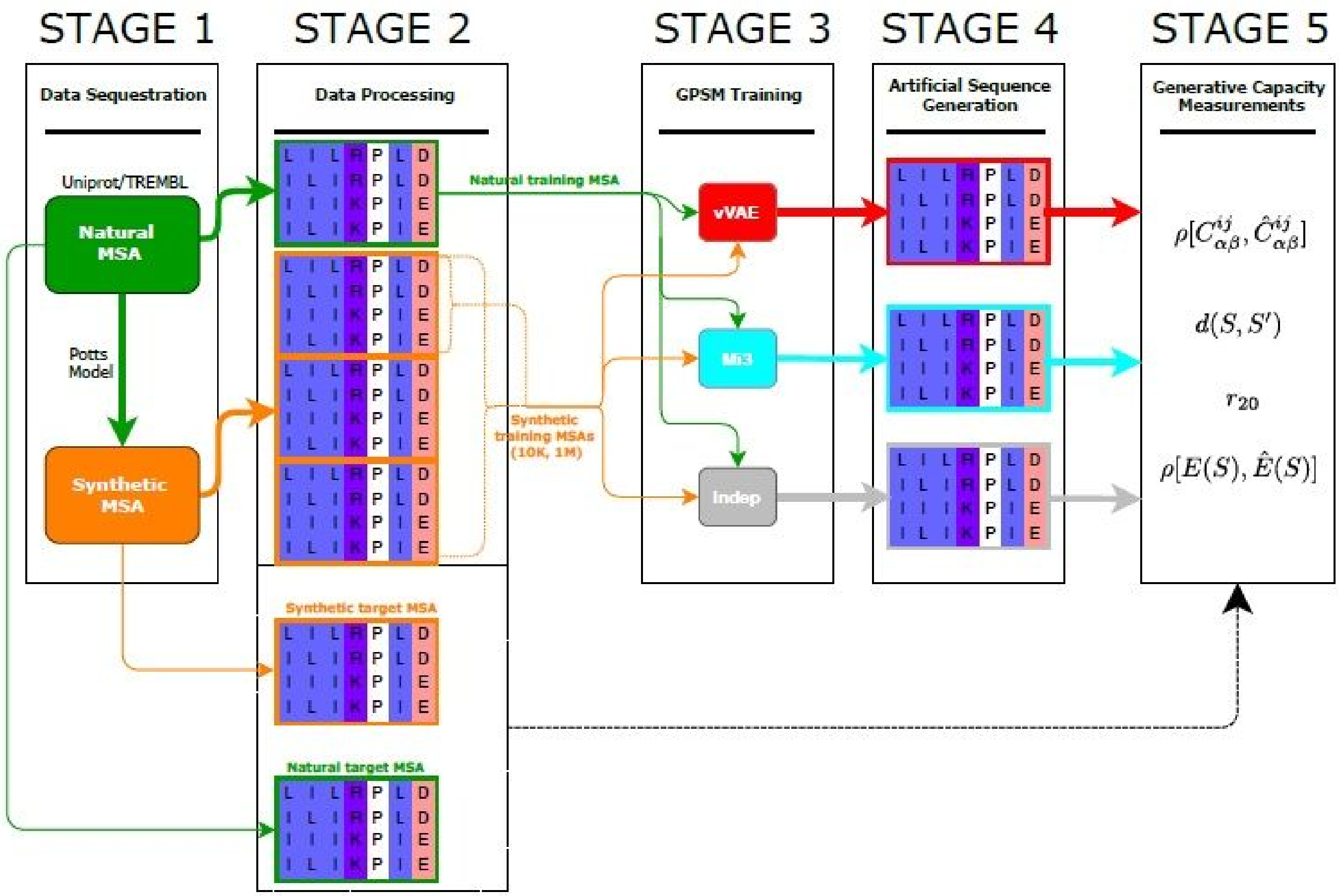

Figure 1

5-stage analysis pipeline diagram. Stage 1: Data Sequestration. Two different MSAs are sequestered, one natural and the other synthetic, generated by a Potts model fit to the natural MSA. Stage 2: Data Processing. Sequences are indexed and split. Non-overlapping training, target, and evaluation MSAs are shown. Stage 3: GPSM Training. GPSMs are trained. Stage 4: Artificial Sequence Generation. Evaluation MSAs are generated from each GPSM. Stage 5: Generative Capacity Measurements. Computation and visualization of generative capacity metrics is performed.

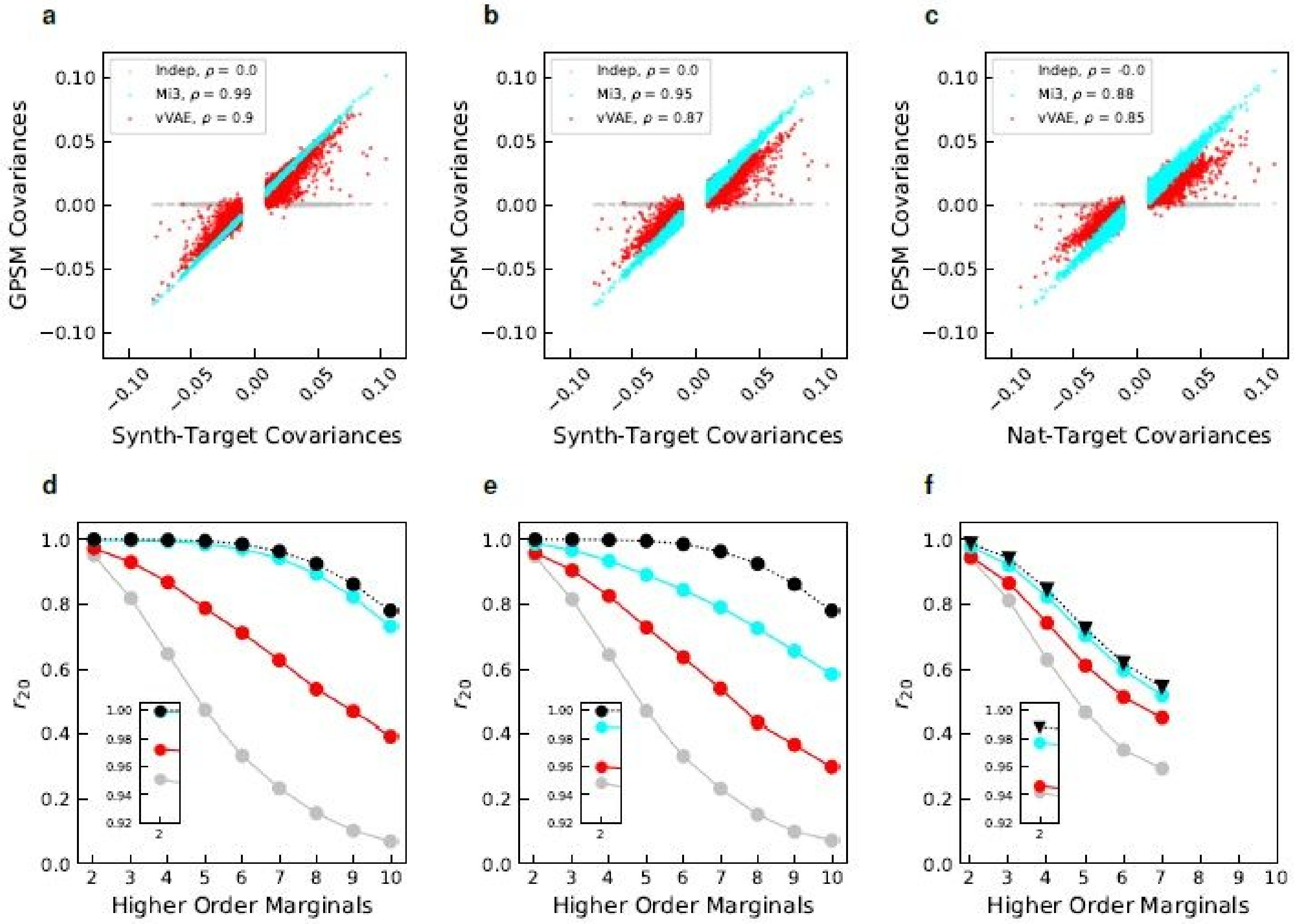

Figure 2

Pairwise covariance correlations (Top Row) and Pearson r20 correlations (Bottom Row). Mi3 (cyan), vVAE (red), and Indep (gray) are compared to target distribution (black, where shown). GPSMs were trained on 1M synthetic (a, d,) 10K synthetic (b, e), or 10K natural (c, f) kinase sequences from the corresponding target distribution. Top Row, These covariance correlation plots show whether the covariances of the GPSMs match those of their respective targets at the fundamental pairwise level. Synthetic target covariances (x-axis) and GPSM-generated evaluation covariances (y-axis) were computed from 500K-sequence MSAs (a, b). In contrast, due to limited availability of natural protein sequence data, natural target covariances were computed from a 10K-sequence target MSA, whereas GPSM-generated covariances (y-axis) were computed from 500K-sequence evaluation MSAs (c). For each covariance scatterplot, a Pearson correlation coefficient r was computed between the GPSM covariance set and the target covariance set. Generative capacity for Mi3 and Indep appear insensitive to training sample size for this metric. Covariances around zero were removed for easier plotting. Bottom Row, These plots show whether the GPSMs, on average, have higher order marginals (HOMs) that match those of their respective targets between orders of two and ten. Insets emphasize pairwise r20. Pearson r20 correlation (r20, y-axis) is plotted as a function of HOMs (x-axis). Synthetic r20 were computed from 6M artificial GPSM-generated evaluation sequences compared to 6M synthetic target sequences (d, e) as colored lines, and the black dotted line here in the synthetic r20 analysis denotes a generative capacity upper limit set by finite sampling of two non-overlapping target MSAs of 6M sequences each. Both Mi3 and vVAE's generative capacity appear sensitive to decreasing synthetic training sample size for this metric. In contrast, for the natural analysis, the black line uses triangles to indicate a difference in how it is computed compared to the synthetic analysis (f). Here, the generative capacity upper limit is an estimate computed between non-overlapping evaluation MSAs of 6M and 10K sequences from the artificial Mi3 distribution trained on natural sequences (f, cyan), and only orders two through seven were plotted due to finite sampling effects at higher orders. In the natural analysis, Mi3 performs as close to the target as is measurable, within error, for this amount of data.

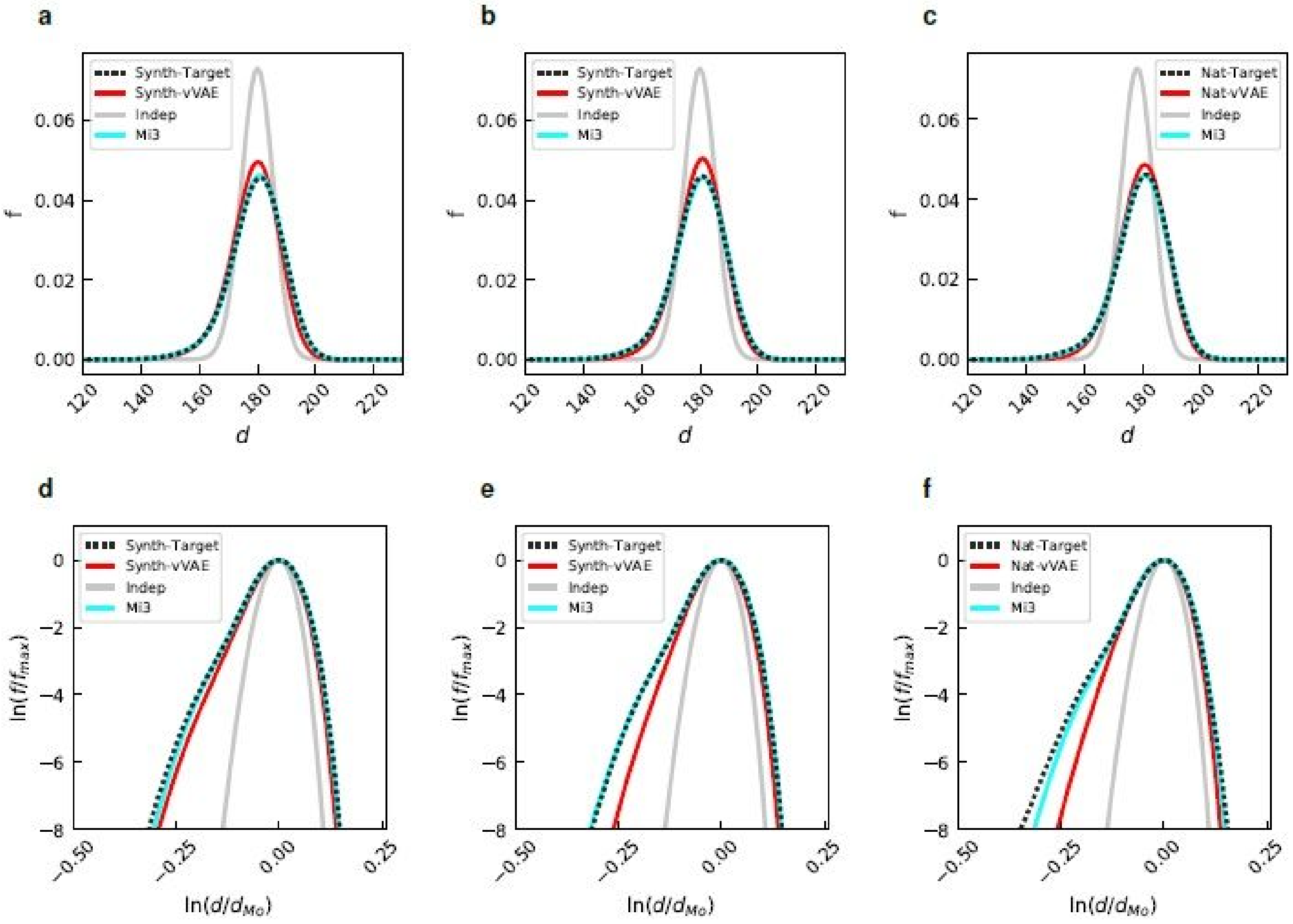

Figure 3

Pairwise Hamming distance distributions. These plots illustrate whether sequences generated from the GPSMs reproduce the overall sequence diversity of their respective targets. Mi3 (cyan), vVAE (red), and Indep (gray) distributions are compared to target distribution (dotted black). GPSMs were trained on 1M synthetic (a, d), 10K synthetic (b, e), or 10K natural (c, f) sequences from the corresponding target distribution. All Hamming distributions were computed from 30K-sequence MSAs, except for the natural target, which was computed from a 10K-sequence target MSA due to data limitations. Top Row, Hamming distances d (x-axis) are shown about the mode, and frequency f is normalized as a fraction of total (y-axis). Mi3 perfectly matches the target distribution, and vVAE performs closely to Mi3. Bottom Row, Re-scaled logarithmic Hamming distance distributions better discriminate between GPSMs with respect to generative capacity than the normal Hamming distance distribution. Before being log-scaled, the Hamming distances d are normalized by the mode dMo (x-axis), and frequencies f are normalized by the maximum Hamming distance fmax (y-axis). This transformation highlights minute differences between distributions at low frequencies in the tails of the distributions on the left- and right-hand sides. Only vVAE appears sensitive to training sample size for this metric at the log-log scale.

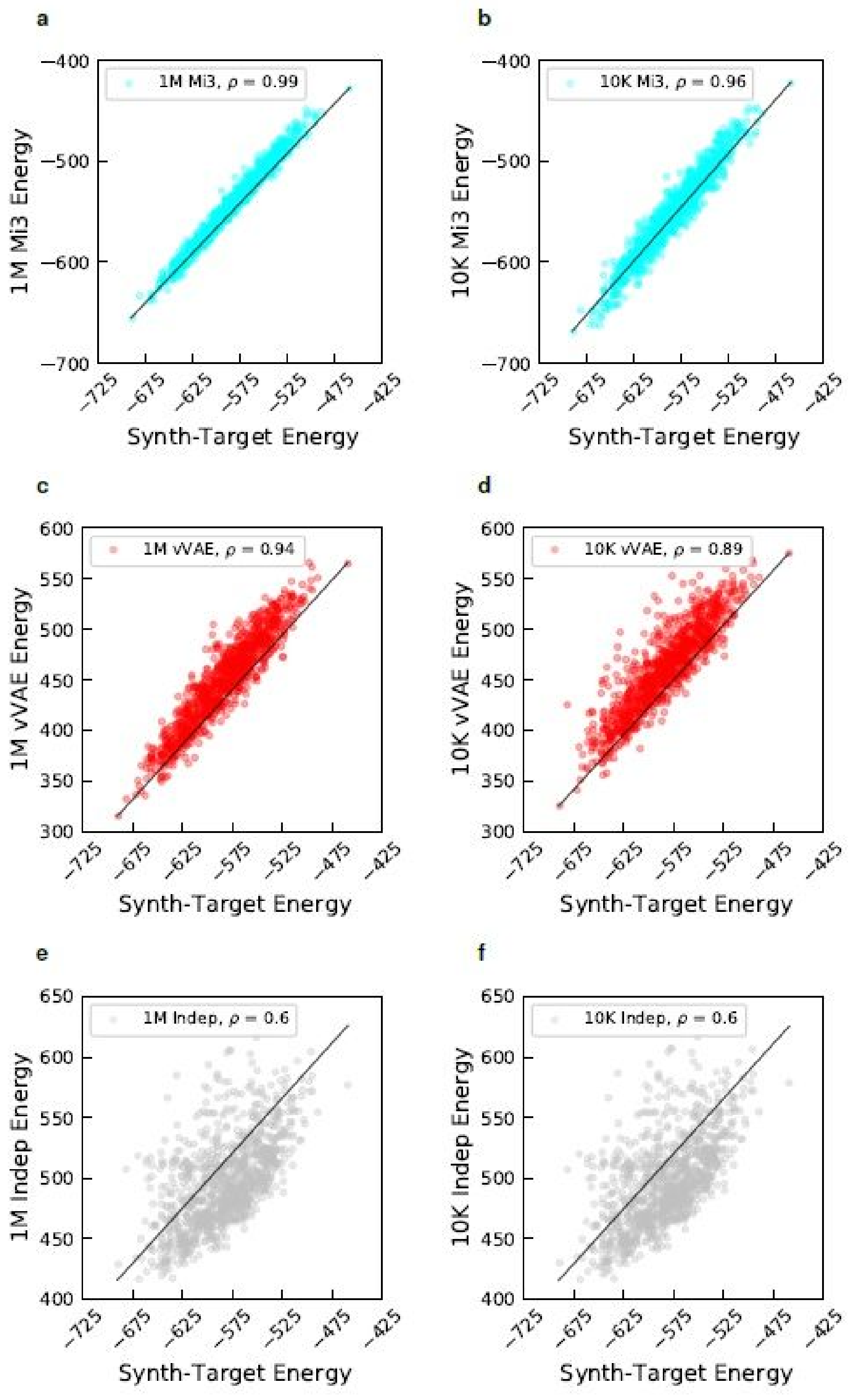

Figure 4

Statistical energy correlations. Statistical energies E(S) of 1K synthetic test sequences from the target distribution as evaluated by Mi3 (cyan), vVAE (red), Indep (gray). Each GPSM was trained on 1M (a, c, e) or 10K (b, d, f) sequences from the synthetic target distribution. For each scatterplot, Pearson correlation coefficient r was computed between each GPSM's statistical energy and that of the synthetic target distribution for each sequence. Only Indep is insensitive to decreased training sample size for this metric.

Supplementary Files

- gencapacitysupplement.pdf