Supplementary Material

for

Boltzmann Machine Learning and Regularization Methods for Inferring Evolutionary Fields and Couplings from a Multiple Sequence Alignment

(Article DOI: 10.1109/TCBB.2020.2993232)

in

IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2020

Sanzo Miyazawa

sanzo.miyazawa@gmail.com

2021-06-07

S.1 Methods

S.1.1 The Inverse Potts model for protein homologous sequences

Let us consider probability distributions $P(\boldsymbol{\sigma})$ of amino acid sequences $\boldsymbol{\sigma}(\equiv(\sigma_1,\ldots,\sigma_L))$, which satisfy the following constraints that single-site and two-site marginal probabilities must be equal to a given frequency $P_i(a_k)$ of amino acid a_k at each site i and a given frequency $P_{ij}(a_k,a_l)$ of amino acid pair (a_k,a_l) for site pair (i,j), respectively.

$$P(\sigma_i=a_k) \equiv \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\delta_{\sigma_i a_k} = P_i(a_k) \quad \text{(S1)}$$

$$P(\sigma_i=a_k,\sigma_j=a_l) \equiv \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\delta_{\sigma_i a_k}\delta_{\sigma_j a_l} = P_{ij}(a_k,a_l) \quad \text{(S2)}$$

where $\sigma_i, a_k \in \{\text{amino acids, deletion}\}$ $k=1,\ldots,q$, $q \equiv |\{\text{amino acids, deletion}\}| = 21$, $i,j=1,\ldots,L$, and $\delta_{\sigma_i a_k}$ is the Kronecker delta. The sequence distribution $P(\boldsymbol{\sigma}|h,J)$ with the maximum entropy can be represented as

$$P(\boldsymbol{\sigma}|h,J) = \frac{1}{Z_{\boldsymbol{\sigma}}} e^{-\psi_N(\boldsymbol{\sigma}|h,J)} \quad , \quad Z_{\boldsymbol{\sigma}} = \sum_{\boldsymbol{\sigma}} e^{-\psi_N(\boldsymbol{\sigma}|h,J)} \quad \text{(S3)}$$

$$\psi_N(\boldsymbol{\sigma}|h,J) = -\left[\sum_i \{h_i(\sigma_i) + \sum_{j(>i)} J_{ij}(\sigma_i,\sigma_j)\}\right] \quad \text{(S4)}$$

where Lagrange multipliers $h_i(a_k)$ and $J_{ij}(a_k,a_l)$ are interaction potentials called fields and couplings, and $\psi_N(\boldsymbol{\sigma}|h,J)$ is referred to here as evolutionary energy.

Fields $h_i(a_k)$ and couplings $J_{ij}(a_k,a_l)$ provide useful information to understand protein evolution [16] and also to predict residue-residue contacts in protein structures on the basis of coevolutional residue substitutions [1], [2], [8], [71].

For given single-site $P_i(a_k)$ and two-site frequencies $P_{ij}(a_k,a_l)$, which are evaluated from a multiple sequence alignment, inferring $h_i(a_k)$ and $J_{ij}(a_k,a_l)$ have been attempted as the Inverse Potts problem by the Boltzmann machine learning [14], [22], by the mean field approximation [1], [2], [8], by the Gaussian approximation [72], by maximizing a pseudo-likelihood [10], [11], [68], [73], and by minimizing a cross entropy in the adaptive cluster expansion [12].

S.1.2 The sample average of evolutionary energy

According to the Potts model, the sample average of $\psi_N(\boldsymbol{\sigma}_N)$ over natural sequences, $\boldsymbol{\sigma}_N$, fixed in protein evolution is equal to the ensemble average of $\psi_N(\boldsymbol{\sigma})$ over sequences, $\boldsymbol{\sigma}$. Sample averages are calculated with a sample weight $w_{\boldsymbol{\sigma}_N}$ for each homologous sequence, which is used to reduce phylogenetic biases in the set of homologous sequences; for example, the sample average of evolutionary energy is calculated as follows.

$$\overline{\psi_N(\boldsymbol{\sigma}_N)} \equiv \frac{\sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N}\psi_N(\boldsymbol{\sigma}_N)}{\sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N}} \quad \text{(S5)}$$

$$= \langle\psi_N(\boldsymbol{\sigma})\rangle_{\boldsymbol{\sigma}} \quad \text{(S6)}$$

where $\overline{\psi_N(\boldsymbol{\sigma}_N)}$ denotes a sample average of $\psi_N(\boldsymbol{\sigma}_N)$ with a sample weight $w_{\boldsymbol{\sigma}_N}$ for each homologous sequence $\boldsymbol{\sigma}_N$, and $\langle\psi_N(\boldsymbol{\sigma})\rangle_{\boldsymbol{\sigma}}$ is the ensemble average of $\psi_N(\boldsymbol{\sigma})$ that obeys a Boltzmann distribution.

S.1.3 Ensemble average by a Gaussian Approximation for the distribution of the evolutionary energies of random sequences

The ensemble average over sequences, for example, of $\psi_N(\boldsymbol{\sigma})$ is estimated by the Gaussian approximation [16], [53], in which the distribution of the evolutionary energies of random sequences is approximated as a Gaussian distribution, $\mathcal{N}(\bar{\psi},\delta\psi^2)$. The mean $\bar{\psi}$ and variance $\delta\psi^2$ are evaluated as those of evolutionary energies of random sequences whose amino acid composition is equal to the average amino acid composition of sequences in a protein family.

$$\langle\psi_N(\boldsymbol{\sigma})\rangle_{\boldsymbol{\sigma}} \equiv \left[\sum_{\boldsymbol{\sigma}} \psi_N(\boldsymbol{\sigma})\exp(-\psi_N(\boldsymbol{\sigma}))\right] / Z_{\boldsymbol{\sigma}} \quad \text{(S7)}$$

$$\approx \frac{\int \psi_N \exp(-\psi_N))\,\mathcal{N}(\bar{\psi},\delta\psi^2)\,d\psi_N}{\int \exp(-\psi_N)\,\mathcal{N}(\bar{\psi},\delta\psi^2)\,d\psi_N} \quad \text{(S8)}$$

$$= \bar{\psi}(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}) - \delta\psi^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}) \quad \text{(S9)}$$

where $\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}$ is the sample-average amino acid composition of natural sequences in a protein family.

S.1.4 Relationships between evolutionary energy $\psi_N(\boldsymbol{\sigma})$, fitness $m(\boldsymbol{\sigma})$, and folding free energy $\Delta G_{ND}(\boldsymbol{\sigma})$ of protein $\boldsymbol{\sigma}$ [16]

In [16], it was proved by assuming the detailed balance principle that the equilibrium distribution of protein sequences must be the Boltzmann distribution of their Malthusian fitness m as well as that of $\Delta\psi_{ND}$. On the other hand, a protein folding theory [50], [51], [52], [53] based on a random energy model (REM) indicates that it can be approximated to the Boltzmann distribution of the folding free energy divided by selective temperature, $\Delta G_{ND}/(k_B T_s)$.

$$P^{\text{eq}}(\boldsymbol{\mu}) = \frac{P^{\text{mut}}(\boldsymbol{\mu})\exp(4N_e m(\boldsymbol{\mu})(1-q_m))}{\sum_{\boldsymbol{\nu}} P^{\text{mut}}(\boldsymbol{\nu})\exp(4N_e m(\boldsymbol{\nu})(1-q_m)))} \quad \text{(S10)}$$

$$= \frac{P^{\text{mut}}(\overline{\boldsymbol{\mu}})\exp(-(\psi_N(\boldsymbol{\mu}) - \psi_D(\overline{\boldsymbol{f}(\boldsymbol{\mu})},T)))}{\sum_{\boldsymbol{\nu}} P^{\text{mut}}(\overline{\boldsymbol{\nu}})\exp(-(\psi_N(\boldsymbol{\nu}) - \psi_D(\overline{\boldsymbol{f}(\boldsymbol{\nu})},T)))} \quad \text{(S11)}$$

$$\simeq \frac{P^{\text{mut}}(\boldsymbol{\mu})\exp(-\Delta G_{ND}(\boldsymbol{\mu},T)/(k_B T_s))}{\sum_{\boldsymbol{\nu}} P^{\text{mut}}(\boldsymbol{\nu})\exp(-\Delta G_{ND}(\boldsymbol{\nu},T)/(k_B T_s))} \quad \text{(S12)}$$

where $p^{\text{mut}}(\boldsymbol{\sigma})$ is the probability of a sequence $(\boldsymbol{\sigma})$ randomly occurring in a mutational process and depends only on the amino acid composition of the sequence $\boldsymbol{f}(\boldsymbol{\sigma})$, q_m is the frequency of a single mutant gene in a population, k_B is the Boltzmann constant, T is growth temperature, T_s is selective temperature that quantifies how strong the folding constraints are in protein evolution, $\overline{\boldsymbol{f}(\boldsymbol{\sigma})} \equiv \sum_{\boldsymbol{\sigma}} \boldsymbol{f}(\boldsymbol{\sigma})P(\boldsymbol{\sigma})$ and $\log P^{\text{mut}}(\overline{\boldsymbol{\sigma}}) \equiv \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\log(\prod_i P^{\text{mut}}(\sigma_i))$. Then, the following relationships are derived for sequences for which $f(\boldsymbol{\mu}) = \overline{f(\boldsymbol{\mu})}$.

$$4N_e m(\boldsymbol{\mu})(1-q_m) = -\Delta\psi_{ND}(\boldsymbol{\mu},T) + \text{constant} \quad \text{(S13)}$$

$$\simeq \frac{-\Delta G_{ND}(\boldsymbol{\mu},T)}{k_B T_s} + \text{constant} \quad \text{(S14)}$$

The selective advantage of $\boldsymbol{\nu}$ to $\boldsymbol{\mu}$ is represented as follows for $f(\boldsymbol{\mu}) = f(\boldsymbol{\nu}) = \overline{f(\boldsymbol{\sigma})}$.

$$4N_e s(\boldsymbol{\mu}\to\boldsymbol{\nu})(1-q_m)$$

$$= (4N_e m(\boldsymbol{\nu}) - 4N_e m(\boldsymbol{\mu}))(1-q_m) \quad \text{(S15)}$$

$$= -(\Delta\psi_{ND}(\boldsymbol{\nu},T) - \Delta\psi_{ND}(\boldsymbol{\mu},T)) = -(\psi_N(\boldsymbol{\nu}) - \psi_N(\boldsymbol{\mu})) \quad \text{(S16)}$$

$$\simeq -(\Delta G_{ND}(\boldsymbol{\nu},T) - \Delta G_{ND}(\boldsymbol{\mu},T))/(k_B T_s)$$

$$= -(G_N(\boldsymbol{\nu}) - G_N(\boldsymbol{\mu}))/(k_B T_s) \quad \text{(S17)}$$

$$\psi_N(\boldsymbol{\mu}) \simeq G_N(\boldsymbol{\mu})/(k_B T_s) \qquad \psi_D(\boldsymbol{\mu}) \simeq G_D(\boldsymbol{\mu})/(k_B T_s) \quad \text{(S18)}$$

where $G_N(\boldsymbol{\sigma})$ and $G_D(\boldsymbol{\sigma})$ are the free energies of the native and the denatured states of sequence $\boldsymbol{\sigma}$. It should be noted here that only sequences for which $f(\boldsymbol{\sigma}) = \overline{f(\boldsymbol{\sigma})}$ contribute significantly to the partition functions in Eq. S11, and other sequences may be ignored.

S.1.5 Relationships among selective temperature (T_s), glass transition temperature (T_g), and melting temperature (T_m) of protein

The distribution of conformational energies in the denatured state (molten globule state), which consists of conformations as compact as the native conformation, is approximated in the random energy model (REM), particularly the independent interaction model (IIM) [53], to be equal to the energy distribution of the randomized sequences, which is approximated by the energy distribution of the random sequences with the same amino acid composition and then by a Gaussian distribution, in the native conformation. That is, the partition function Z for the denatured state is written as follows with the number density per energy $n(E)$ of conformations that is approximated by a product of a Gaussian probability density and the total number of conformations whose logarithm is proportional to the chain length.

$$Z = \int \exp(\frac{-E}{k_B T})\, n(E) dE \tag{S19}$$

$$n(E) \approx \exp(\omega L) \mathcal{N}(\bar{E}(\boldsymbol{f}(\boldsymbol{\sigma}_N)), \delta E^2(\boldsymbol{f}(\boldsymbol{\sigma}_N))) \tag{S20}$$

where ω is the conformational entropy per residue in the compact denatured state, and $\mathcal{N}(\bar{E}(\boldsymbol{f}(\boldsymbol{\sigma}_N)), \delta E^2(\boldsymbol{f}(\boldsymbol{\sigma}_N)))$ is the Gaussian probability density with mean $\bar{E}$ and variance δE^2, which depend only on the amino acid composition, $\boldsymbol{f}(\boldsymbol{\sigma}_N)$, of the protein sequence, $\boldsymbol{\sigma}_N$. The free energy of the denatured state is approximated as follows.

$$G_D(\boldsymbol{\sigma}_N, T) \approx \bar{E}(\boldsymbol{f}(\boldsymbol{\sigma}_N)) - \frac{\delta E^2(\boldsymbol{f}(\boldsymbol{\sigma}_N))}{2k_B T} - k_B T \omega L \tag{S21}$$

$$= \bar{E}(\boldsymbol{f}(\boldsymbol{\sigma}_N)) - \delta E^2(\boldsymbol{f}(\boldsymbol{\sigma}_N)) \frac{\vartheta(T/T_g)}{k_B T} \tag{S22}$$

$$\psi_D(\boldsymbol{\sigma}_N, T) \approx \bar{\psi}(\boldsymbol{f}(\boldsymbol{\sigma})) - \delta\psi^2(\boldsymbol{f}(\boldsymbol{\sigma}))\vartheta(T/T_g)\frac{T_s}{T} \tag{S23}$$

$$\vartheta(\frac{T}{T_g}) \equiv \begin{cases} (1 + T^2/T_g^2)/2 & \text{for } T > T_g \\ T/T_g & \text{for } T \le T_g \end{cases} \tag{S24}$$

where $\bar{E}$ ($\bar{\psi}$) and δE^2 ($\delta\psi^2$) are estimated as the mean and variance of interaction energies E (ψ_N) of the randomized sequences, which are approximated by random sequences, in the native conformation; $\bar{E} \simeq k_B T_s \bar{\psi}$ and $\delta E^2 \simeq (k_B T_s)^2 \delta\psi^2$. T_g is the glass transition temperature of the protein at which entropy becomes zero [50], [51], [52], [53].

$$-\frac{\partial G_D}{\partial T}|_{T=T_g} = 0 \tag{S25}$$

The conformational entropy per residue ω in the compact denatured state can be represented with T_g.

$$\omega L = \frac{\delta E^2}{2(k_B T_g)^2} \tag{S26}$$

Thus, unless $T_g < T_m$, a protein will be trapped at local minima on a rugged free energy landscape before it can fold into a unique native structure.

The ensemble average of $\Delta G_{ND}(\boldsymbol{\sigma}, T)$ over sequences, which is observable as the sample averages of $\Delta G_{ND}(\boldsymbol{\sigma}_N, T)$ over homologous sequences fixed in protein evolution, is estimated as follows [16].

$$\langle \Delta G_{ND}(\boldsymbol{\sigma}, T) \rangle_{\boldsymbol{\sigma}} \equiv \sum_{\boldsymbol{\sigma}} \Delta G_{ND}(\boldsymbol{\sigma}, T) P^{\mathrm{eq}}(\boldsymbol{\sigma})$$

$$\approx \sum_{\{\boldsymbol{\sigma} | f(\boldsymbol{\sigma}) = \overline{f(\boldsymbol{\sigma}_N)}\}} \Delta G_{ND}(\boldsymbol{\sigma}, T) P^{\mathrm{eq}}(\boldsymbol{\sigma}) \tag{S27}$$

$$= \langle G_N(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\sigma}} - G_D(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}, T) \tag{S28}$$

where the ensemble averages of $G_N(\boldsymbol{\sigma})$ over sequences is also estimated in the Gaussian approximation [53].

$$\langle G_N(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\sigma}}$$

$$\approx \int E \exp(-\frac{E}{k_B T_s})\, \mathcal{N}(\bar{E}(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}), \delta E^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}))\, dE \tag{S29}$$

$$= \bar{E}(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}) - \frac{\delta E^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)})}{k_B T_s} \tag{S30}$$

The sample averages of $\Delta G_{ND}(\boldsymbol{\sigma}_N, T)$ and $\psi_N(\boldsymbol{\sigma}_N)$ over homologous sequences fixed in protein evolution are equal to their ensemble averages over sequences [16].

$$\overline{\Delta G_{ND}(\boldsymbol{\sigma}_N, T)}/(k_B T_s)$$

$$= \langle \Delta G_{ND}(\boldsymbol{\sigma}, T) \rangle_{\boldsymbol{\sigma}}/(k_B T_s) \tag{S31}$$

$$\approx [\, \delta E^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)})\, [\, \vartheta(T/T_g) T_s/T - 1\,]/(k_B T_s)^2 \tag{S32}$$

$$= \delta\psi^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)})\, [\, \vartheta(T/T_g) T_s/T - 1\,] \tag{S33}$$

$$= \overline{\Delta G_{ND}(\boldsymbol{\sigma}_N, T_g)}\, /\, (k_B T_s') \tag{S34}$$

$$T_s' = T_s(T_s/T_g - 1)/(\vartheta(T/T_g) T_s/T - 1) \tag{S35}$$

$$\overline{\psi_N(\boldsymbol{\sigma}_N)} \equiv \frac{\sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N} \psi_N(\boldsymbol{\sigma}_N)}{\sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N}} \tag{S36}$$

$$= \langle \psi_N(\boldsymbol{\sigma}) \rangle_{\boldsymbol{\sigma}} \approx \bar{\psi}(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}) - \delta\psi^2(\overline{\boldsymbol{f}(\boldsymbol{\sigma}_N)}) \tag{S37}$$

where the sample averages are calculated with a sample weight $w_{\boldsymbol{\sigma}_N}$ for each homologous sequence, which is used to reduce phylogenetic biases in the set of homologous sequences. $\Delta G_{ND}(\boldsymbol{\sigma}_N, T_g)$ corresponds to the energy gap [50] between the native and the glass states, and T_s' will be the selective temperature if $\Delta G_{ND}(\boldsymbol{\sigma}_N, T_g)$ is used for selection instead of $\Delta G_{ND}(\boldsymbol{\sigma}_N, T)$.

The folding free energy becomes equal to zero at the melting temperature T_m; $\langle \Delta G_{ND}(\boldsymbol{\sigma}_N, T_m) \rangle_{\boldsymbol{\sigma}} = 0$. Thus, the following relationship must be satisfied [50], [51], [52], [53].

$$\vartheta(\frac{T_m}{T_g})\frac{T_s}{T_m} = \frac{T_s}{2T_m}(1 + \frac{T_m^2}{T_g^2}) = 1 \quad \text{with } T_s \le T_g \le T_m \tag{S38}$$

S.1.6 Boltzmann machine learning

The cross entropy with a regularization term, S, which corresponds to a negative log-posterior-probability per instance, is minimized.

$$S \equiv \frac{-1}{\sum_\tau 1} \sum_\tau \log P(\boldsymbol{\sigma}^\tau) + R \tag{S39}$$

where R is a regularization term, and τ denotes an instance. According to [14], instead of h_i and J_{ij}, we use the new parameters ϕ_i and ϕ_{ij} for minimization, which are Lagrange multipliers in the maximum entropy model corresponding to $[\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\delta_{\sigma_i a_k} - P_i(a_k)]$ and $[\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\delta_{\sigma_i a_k}\delta_{\sigma_j a_l} - P_{ij}(a_k, a_l) - \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\delta_{\sigma_i a_k} P_j(a_l) - P_i(a_k)\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\delta_{\sigma_j a_l} + 2P_i(a_k)P_j(a_l)]$ in the

maximum entropy model, respectively. The partial derivatives of the cross entropy can be easily calculated:

$$\frac{\partial S}{\partial \phi_i(a_k)} = \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\delta_{\sigma_i a_k} - P_i(a_k) + \frac{\partial R}{\partial \phi_i(a_k)} \quad \text{(S40)}$$

$$\begin{aligned}\frac{\partial S}{\partial \phi_{ij}(a_k,a_l)} &= [\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\delta_{\sigma_i a_k}\delta_{\sigma_j a_l} - P_{ij}(a_k,a_l) \\ &\quad - \sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\delta_{\sigma_i a_k}P_j(a_l) - P_i(a_k)\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\delta_{\sigma_j a_l} \\ &\quad + 2P_i(a_k)P_j(a_l)]] + \frac{\partial R}{\partial \phi_{ij}(a_k,a_l)} \quad \text{(S41)}\end{aligned}$$

The relationships between (h_i, J_{ij}) and (ϕ_i, ϕ_{ij}) are as follows.

$$h_i(a_k) = \phi_i(a_k) - \sum_{j(\neq i)}\sum_l \phi_{ij}(a_k,a_l))P_j(a_l) \quad \text{(S42)}$$

$$J_{ij}(a_k,a_l) = \phi_{ij}(a_k,a_l) \quad \text{(S43)}$$

The single-site and two-site frequencies, $P_i(a_k)$ and $P_{ij}(a_k, a_l)$, are evaluated from homologous sequences, each of which has a sample weight $w_{\boldsymbol{\sigma}_N}$, in a multiple sequence alignment.

$$P_i(a_k) = \sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N}\delta_{\sigma_{N,i}a_k} / \sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N} \quad \text{(S44)}$$

$$P_{ij}(a_k,a_l) = \sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N}\delta_{\sigma_{N,i}a_k}\delta_{\sigma_{N,j}a_l} / \sum_{\boldsymbol{\sigma}_N} w_{\boldsymbol{\sigma}_N} \quad \text{(S45)}$$

where $\boldsymbol{\sigma}_N$ denotes natural sequences.

$\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\delta_{\sigma_i a_k}$ and $\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma})\delta_{\sigma_i a_k}\delta_{\sigma_j a_l}$ are estimated by a Markov chain Monte Carlo method with the Metropolis-Hastings algorithm [23], [24], and then a gradient-descent algorithm is used to minimize the cross entropy S; the Metropolis-Hastings algorithm was employed rather than the Gibbs sampler [25], because calculating full conditionals require more computation time.

S.1.7 Regularization

Couplings $\phi_{ij}(a_k, a_l)$ are expected to be significant between residues that are closely located in a 3D protein structure and complex. Thus, they are expected to be sparse, because the number of residue-residue contacts in a protein 3D structure is between 2 and 4 per residue depending on a criterion, in comparison with the number of residue pairs, $L(L-1)/2$, where L is a protein length [74]. Here, to take account of the sparsity of the couplings, the elastic net [54], [75], [76] and group L_1 regularizations are employed to see the effects of different regularizations. The elastic net regularization [54], [75], [76] is used instead of pure L_1 regularization, which is not strictly convex and can occasionally produce non-unique solutions [54]. Group L_1 is employed to deal with pairwise couplings, $\phi_{ij}(a_k, a_l)$, between residues i and j as a group.

S.1.7.1 An elastic net regularization

An elastic net regularization [54], [75], [76] is a mixture of L_1 and L_2.

$$\begin{aligned}R \equiv \lambda_1 \sum_i \sum_k \{\theta_1|\phi_i(a_k)| + \frac{(1-\theta_1)}{2}\phi_i(a_k)^2\} + \\ \lambda_2 \sum_i \sum_k \sum_{j(>i)} \sum_l \{\theta_2|\phi_{ij}(a_k,a_l)| + \frac{(1-\theta_2)}{2}\phi_{ij}(a_k,a_l)^2\} \quad \text{(S46)}\end{aligned}$$

where $0 \leq \theta_1 \leq 1$ and $0 \leq \theta_2 \leq 1$. If $\theta = 0(1)$, the regularization will be $L_2(L_1)$. In the present work, $L1$ regularization means the elastic net with $\theta = 0.9$ rather than 1.0.

S.1.7.1.1 The soft-thresholding function for L_1 regularization: Let us assume that the learning of fields and couplings (ϕ_i, ϕ_{ij}) is iteratively updated as follows.

$$\phi_\mu(t+1) = \phi_\mu(t) - [\alpha_\mu(t+1) + \beta_\mu(t+1)(\frac{\partial S}{\partial \phi_\mu})_{\boldsymbol{\phi}(t)}] \quad \text{(S47)}$$

$$= \phi_\mu(t+1 \text{ without } L_1) - \gamma_\mu(t+1)(\frac{\partial|\phi_\mu|}{\partial \phi_\mu})_{\boldsymbol{\phi}(t)} \quad \text{(S48)}$$

$$= \text{prox}(\gamma_\mu(t+1)|\phi_\mu|\,,\, \phi_\mu(t+1 \text{ without } L_1)) \quad \text{(S49)}$$

where the suffix μ denotes ${}_i(a_k)$ or ${}_{ij}(a_k, a_l)$, $\phi_\mu(t+1 \text{ without } L_1)$ is $\phi_\mu(t+1)$ which does not include the L_1 regularization term, and the second term is one corresponding to the L_1 terms of the regularization in Eq. S46, and the derivative in the second term may be evaluated as a subderivative at a singular point. Here the proximal operator [77] defined as follows is used for faster convergence.

$$\text{prox}(h(u), x) \equiv \text{argmin}_u(h(u) + \frac{1}{2}\|u-x\|_2^2) \quad \text{(S50)}$$

The proximal operator for L_1 regression is equal to:

$$\begin{aligned}&\text{prox}(\gamma_\mu(t+1)|\phi_\mu|\,,\, \phi_\mu(t+1 \text{ without } L_1)) \\ &= \begin{cases} \text{if } \phi_\mu(t+1 \text{ without } L_1) > \gamma_\mu(t+1) \\ \qquad \phi_\mu(t+1 \text{ without } L_1) - \gamma_\mu(t+1) \\ \text{if } |\phi_\mu(t+1 \text{ without } L_1)| \leq \gamma_\mu(t+1) \\ \qquad 0 \\ \text{if } \phi_\mu(t+1 \text{ without } L_1) < -\gamma_\mu(t+1) \\ \qquad \phi_\mu(t+1 \text{ without } L_1) + \gamma_\mu(t+1) \end{cases} \quad \text{(S51)}\end{aligned}$$

$$\gamma_\mu(t+1) \equiv \begin{cases} \beta_\mu\lambda_1\theta_1 & \text{for } \mu =_i (a_k) \\ \beta_\mu\lambda_2\theta_2 & \text{for } \mu =_{ij} (a_k,a_l) \end{cases} \quad \text{(S52)}$$

S.1.7.2 L_2 regularization for $\phi_i(a_k)$ and group L_1 for $\phi_{ij}(a_k, a_l)$

The regularization terms of the L_2 for $\phi_i(a_k)$ and the group L_1 for $\phi_{ij}(a_k, a_l)$ are as follows.

$$R \equiv \lambda_1 \sum_i \sum_k \frac{1}{2}\{\phi_i(a_k)^2\} + \lambda_2 \sum_i \sum_{j(>i)} \sqrt{\sum_k \sum_l \{\phi_{ij}(a_k,a_l)^2\}} \quad \text{(S53)}$$

S.1.7.2.1 The soft-thresholding function for group L_1 regularization:

$$\boldsymbol{\phi_{ij}}(t+1) = \text{prox}(\gamma_{ij}(t+1)\left\|\boldsymbol{\phi_{ij}}\right\|_2\,,\, \boldsymbol{\phi_{ij}}(t+1 \text{ without group } L_1))$$

$$= \begin{cases} \text{if } \phi_{ij}(t) > 0 \text{ and } \phi_{ij}(t+1 \text{ without group } L_1) > \gamma_{ij}(t+1)\frac{\phi_{ij}(t)}{\|\phi_{ij}(t)\|_2} \\ \qquad \phi_{ij}(t+1 \text{ without group } L_1) - \gamma_{ij}(t+1)\frac{\phi_{ij}(t)}{\|\phi_{ij}(t)\|_2} \\ \text{if } \phi_{ij}(t) < 0 \text{ and } \phi_{ij}(t+1 \text{ without group } L_1) < \gamma_{ij}(t+1)\frac{\phi_{ij}(t)}{\|\phi_{ij}(t)\|_2} \\ \qquad \phi_{ij}(t+1 \text{ without group } L_1) - \gamma_{ij}(t+1)\frac{\phi_{ij}(t)}{\|\phi_{ij}(t)\|_2} \\ \text{if } \left\|\boldsymbol{\phi_{ij}}(t)\right\|_2 = 0 \text{ and } \phi_{ij}(t+1 \text{ without group } L_1) > \gamma_{ij}(t+1) \\ \qquad \phi_{ij}(t+1 \text{ without group } L_1) - \gamma_{ij}(t+1) \\ \text{if } \left\|\boldsymbol{\phi_{ij}}(t)\right\|_2 = 0 \text{ and } \phi_{ij}(t+1 \text{ without group } L_1) < -\gamma_{ij}(t+1) \\ \qquad \phi_{ij}(t+1 \text{ without group } L_1) + \gamma_{ij}(t+1) \\ \text{otherwise} \\ \qquad 0 \end{cases} \quad \text{(S54)}$$

$$\gamma_{ij}(t+1) \equiv \beta_{ij}\lambda_2 \quad \text{(S55)}$$

Here it should be noted that $\gamma_{ij}(t+1)$ must not depend on (a_k, a_l) but may depend on (i, j).

S.1.8 Parameter updates

Given the convexity of the cross entropy function, its minimum can be found by the gradient descent.

S.1.8.1 Modified Adam method (ModAdam)

The modified version of the adaptive learning rate method [62], which is named ModAdam here, has been used.

$$m_\mu(t+1) = \rho_m m_\mu(t) + (1-\rho_m)[(\frac{\partial S}{\partial \phi_\mu})_{\boldsymbol{\phi}(t)}] \quad \text{(S56)}$$

$$v_\mu(t+1) = \rho_v v_\mu(t) + (1-\rho_v)[(\frac{\partial S}{\partial \phi_\mu})_{\boldsymbol{\phi}(t)}]^2 \quad \text{(S57)}$$

$$\kappa(t+1) = \kappa_0 \frac{(1-\rho_v^{t+1})^{1/2}}{1-\rho_m^{t+1}} \frac{1}{\max_\mu(v_\mu(t+1)^{1/2})+\epsilon} \quad \text{(S58)}$$

$$\phi_\mu(t+1) = \phi_\mu(t) - \kappa(t+1) m_\mu(t+1) \quad \text{(S59)}$$

where κ_0 is an initial learning rate, and ρ_m, ρ_v, and $\epsilon/(1-\rho_v^{t+1})^{1/2}$ have been set to 0.9, 0.999, and 10^{-8} according to the Adam method [62]. It should be noted here that unlike the Adam method $\kappa(t+1)$ takes the same value for all parameters, because $v_\mu(t+1)^{1/2}$ is replaced by its maximum in Eq. S58; the condition of Eq. S55 required for the soft-thresholding function is satisfied.

An important property of Adam's update rule is its careful choice of stepsizes. The effective stepsize is upper bounded by $|\Delta\phi(t+1)| \leq \kappa_0 \max((1-\rho_m)/\sqrt{(1-\rho_v)}, 1)$ [62] but essentially all elements of the increment vector $\Delta\phi(t+1)$ are the same order. However, unlike the original Adam, in which $\Delta\phi_\mu(t+1) = -\kappa_0(\sqrt{(1-\rho_v^{t+1})}/(1-\rho_m^{t+1}))(m_\mu(t+1)/(v_\mu(t+1)^{1/2}+\epsilon))$, in this modified version the increment $\Delta\boldsymbol{\phi}(t+1)$ is proportional to $-\boldsymbol{m}(t+1)$.

Thus, α_μ and β_μ in Eq. S47 are defined as follows.

$$\alpha_\mu(t+1) = \kappa(t+1)\rho_m m_\mu(t) \quad \text{(S60)}$$

$$\beta_\mu(t+1) = \kappa(t+1)(1-\rho_m) \quad \text{(S61)}$$

S.1.8.2 Nesterov's Accelerated Momentum/Gradient method (NAG)

The algorithm of Nesterov's Accelerated Momentum/Gradient method (NAG) [56] employed here is a simple version with the constant friction for velocity as follows. This version includes a correction, which is employed in the Adam method, for the bias that the estimate of the first moment of the gradients will be biased towards zero if it is initialized as zero.

$$m_\mu(t+1) = \rho_m m_\mu(t) + (1-\rho_m)[(\frac{\partial S}{\partial \phi_\mu})_{\boldsymbol{\phi}(t)}] \quad \text{(S62)}$$

$$\phi_\mu(t+1) = \phi_\mu(t) - \kappa_0[(1+\rho_m)m_\mu(t+1) - \rho_m m_\mu(t)] \quad \text{(S63)}$$

$$= \phi_\mu(t) - \kappa_0[\rho_m^2 m_\mu(t) + (1-\rho_m^2)(\frac{\partial S}{\partial \phi_\mu})_{\boldsymbol{\phi}(t)}] \quad \text{(S64)}$$

$$= \phi_\mu(t) - \kappa_0[\rho_m m_\mu(t+1) + (1-\rho_m)(\frac{\partial S}{\partial \phi_\mu})_{\boldsymbol{\phi}(t)}] \quad \text{(S65)}$$

where κ_0 is an initial learning rate, and ρ_m has been set to 0.95. The α_μ and β_μ in Eq. S47 for the NAG method are defined as follows by replacing $m_\mu(t+1)$ in Eq. S65 by its estimate, $m_\mu(t+1)/(1-\rho_m^{t+1})$, for the initial condition, $m(0) \equiv 0$.

$$\alpha_\mu(t+1) = \kappa_0 \rho_m^2 [\, m_\mu(t)/(1-\rho_m^{t+1}) \,] \quad \text{(S66)}$$

$$\beta_\mu(t+1) = \kappa_0 [\, \rho_m/(1-\rho_m^{t+1}) + 1 \,](1-\rho_m) \quad \text{(S67)}$$

S.1.8.3 The number of iterations for learning

The objective function is expected to significantly fluctuate in the minimization process, when the first-order methods based on gradients are employed. In addition, the partial derivatives of Eqs. S40 and S41, which are calculated from the pairwise marginal distributions estimated by Markov Chain Monte Carlo samplings, include statistical errors. Thus, even though the learning rate κ is sufficiently small, the cross entropy/log-likelihood are not monotonically improved. However, the cross entropy/log-likelihood can hardly be evaluated for the Boltzmann machine, although its partial-derivatives can be easily calculated and then it can be minimized/maximized. Thus, it is not obvious to judge which set of interactions is the best in the learning process.

Here we monitor the average, D_2^{KL}, of Kullback-Leibler divergences for pairwise marginal distributions over all residue pairs as a rough measure of fitting to the reference distribution.

$$D_2^{KL} \equiv \frac{2}{L(L-1)} \sum_i \sum_{j>i} \sum_k \sum_l P_{ij}(a_k, a_l) \log \frac{(P_{ij}(a_k, a_l)+\epsilon)}{(\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \delta_{\sigma_i a_k} \delta_{\sigma_j a_l} + \epsilon)} \quad \text{(S68)}$$

$$D_1^{KL} \equiv \frac{1}{L} \sum_i \sum_k P_i(a_k) \log \frac{(P_i(a_k)+\epsilon)}{(\sum_{\boldsymbol{\sigma}} P(\boldsymbol{\sigma}) \delta_{\sigma_i a_k} + \epsilon)} \quad \text{(S69)}$$

where $\epsilon = 10^{-5}$ is employed to prevent the logarithm of zero. The iteration of parameter updates has been stopped when $\min D_2^{KL}$ over the iteration numbers larger than 1000 does not improve during a certain number, 100, of iterations, and the number of iterations passes over a certain threshold, 1200 iterations. Then the fields and couplings and Monte Carlo samples corresponding to the $\min D_2^{KL}$ over the iteration numbers larger than 1000 are selected.

S.1.9 A gauge employed to compare $h_i(a_k)$ and $J_{ij}(a_k, a_l)$ between various models

The ψ_N of Eq. S4 is invariant under a certain transformation of fields and couplings, $J_{ij}(a_k, a_l) \to J_{ij}(a_k, a_l) - J_{ij}^1(a_k) - J_{ji}^1(a_l) + J_{ij}^0$, $h_i(a_k) \to h_i(a_k) - h_i^0 + \sum_{j\neq i} J_{ij}^1(a_k)$ for any $J_{ij}^1(a_k)$, J_{ij}^0 and h_i^0. Therefore, in order to compare h and J between various models, a certain gauge must be used. Here we use the following gauge that we call the Ising gauge.

$$h_i(\cdot) = \sum_q J_{ij}(a_k, \cdot) = \sum_q J_{ij}(a_q, \cdot) = 0 \quad \text{(S70)}$$

where "$\cdot$" denotes the reference state, which is the average over all states for the Ising gauge. Any gauge can be transformed to this gauge by the following transformation.

$$J_{ij}^{\mathrm{I}}(a_k, a_l) \equiv J_{ij}(a_k, a_l) - J_{ij}(\cdot, a_l) - J_{ij}(a_k, \cdot) + J_{ij}(\cdot, \cdot) \quad \text{(S71)}$$

$$h_i^{\mathrm{I}}(a_k) \equiv h_i(a_k) - h_i(\cdot) + \sum_{j\neq i} (J_{ij}(a_k, \cdot) - J_{ij}(\cdot, \cdot)) \quad \text{(S72)}$$

S.2 Figures

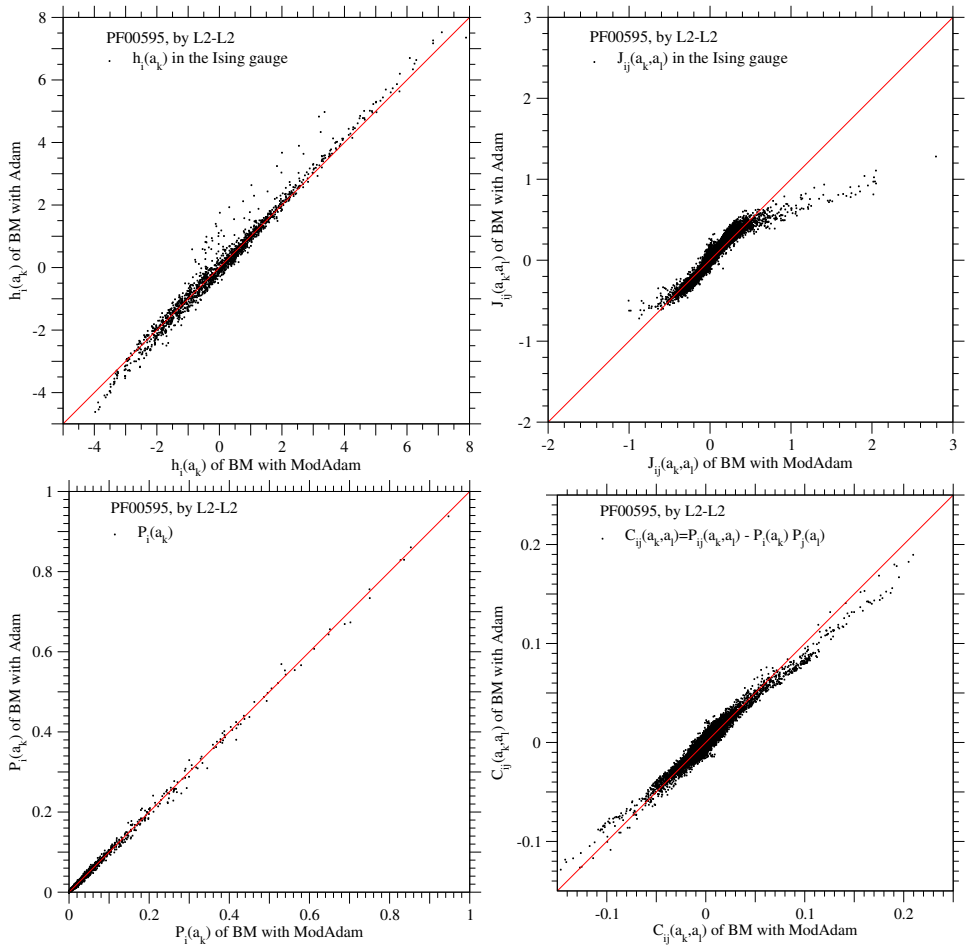

Fig. S1. Comparison of the Adam with the ModAdam gradient-descent method in each of the inferred fields and couplings and the recovered single-site marginals and pairwise correlations for PF00595. The upper left and upper right figures are the comparisons of the inferred fields and couplings in the Ising gauge, respectively, and the lower left and lower right figures are the comparisons of the recovered single-site frequencies and pairwise correlations, respectively. The abscissas and ordinates correspond to the quantities estimated by the modified Adam and Adam methods for gradient descent, respectively. The regularization model L2-L2 is employed for both methods. The solid lines show the equal values between the ordinate and abscissa. The values of hyper-parameters are listed in Table 2. The overlapped points of $J_{ij}(a_k, a_l)$ in the units 0.001 and of $C_{ij}(a_k, a_l)$ in the units 0.0001 are removed.

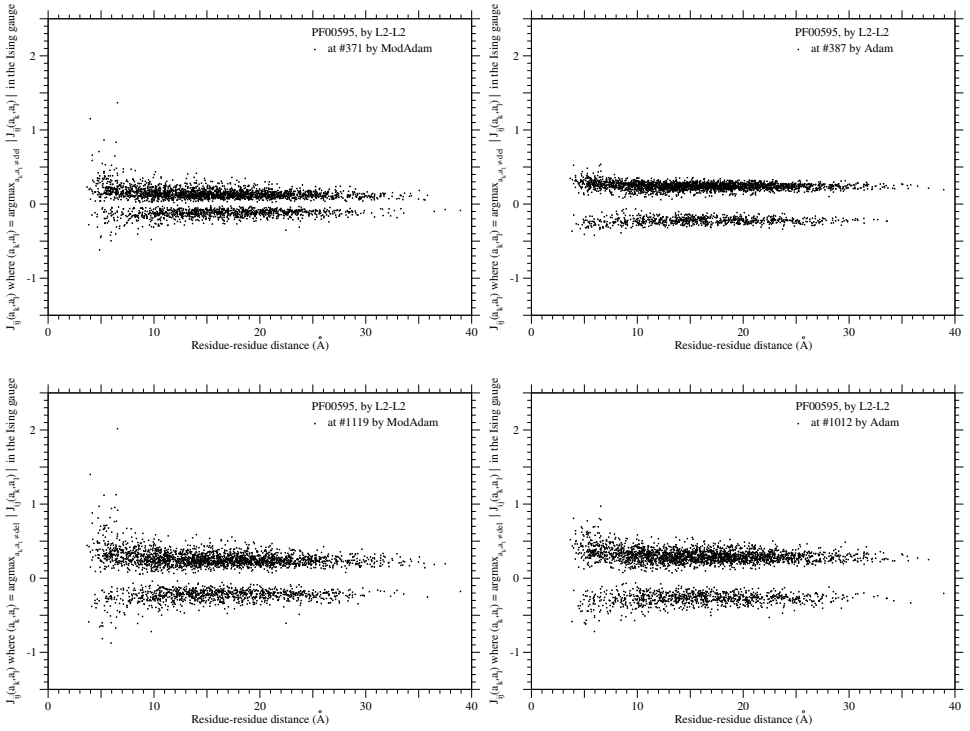

Fig. S2. Differences in the learning of coupling parameters, $J_{ij}(a_k, a_l)$, between the ModAdam and Adam gradient-descent methods for PF00595. All $J_{ij}(a_k, a_l)$ where $(a_k, a_l) = \mathrm{argmax}_{a_k, a_l \neq \mathrm{deletion}} |J_{ij}(a_k, a_l)|$ in the Ising gauge are plotted against the distance between *i*th and *j*th residues. The upper left and lower left figures are for the iteration numbers 371 and 1119 in a learning process by the modified Adam method, respectively. The upper right and lower right figures are for the iteration numbers 387 and 1012 in a learning process by the Adam method, respectively. These iteration numbers correspond to $\min D_2^{\mathrm{KL}}$ over the iteration numbers smaller than 400 and those over the iteration numbers larger than 1000. The regularization model L2-L2 is employed for both methods. The learning processes by both methods are shown in Figs. 2 and 5. Please notice that more strong couplings tend to be inferred for closely located residues pairs by the modified Adam method than by the Adam method. The values of hyper-parameters are listed in Table 2.

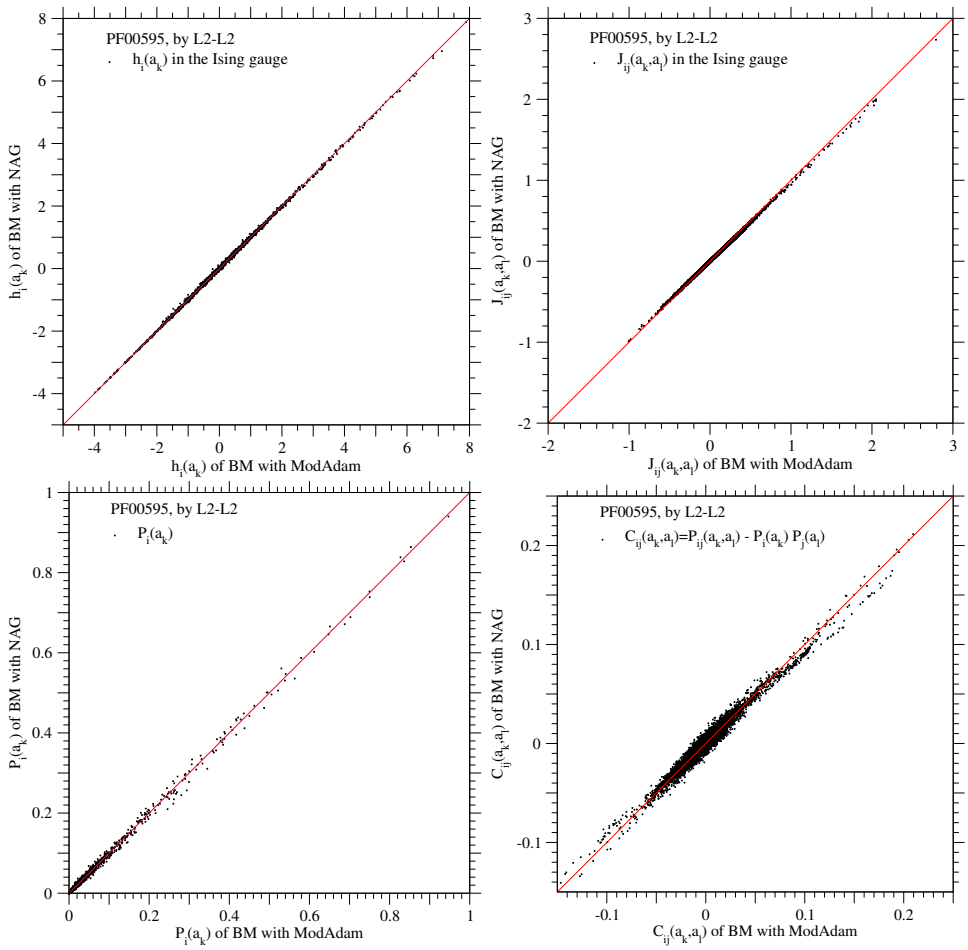

Fig. S3. Comparison of the NAG with the ModAdam gradient-descent method in each of the inferred fields and couplings and the recovered single-site marginals and pairwise correlations for PF00595. The upper left and upper right figures are the comparisons of the inferred fields and couplings in the Ising gauge, respectively, and the lower left and lower right figures are the comparisons of the recovered single-site frequencies and pairwise correlations, respectively. The abscissas and ordinates correspond to the quantities estimated by the modified Adam and NAG methods for gradient descent, respectively. The regularization model L2-L2 is employed for both methods. The solid lines show the equal values between the ordinate and abscissa. The values of hyper-parameters are listed in Table 2. The overlapped points of $J_{ij}(a_k, a_l)$ in the units 0.001 and of $C_{ij}(a_k, a_l)$ in the units 0.0001 are removed.

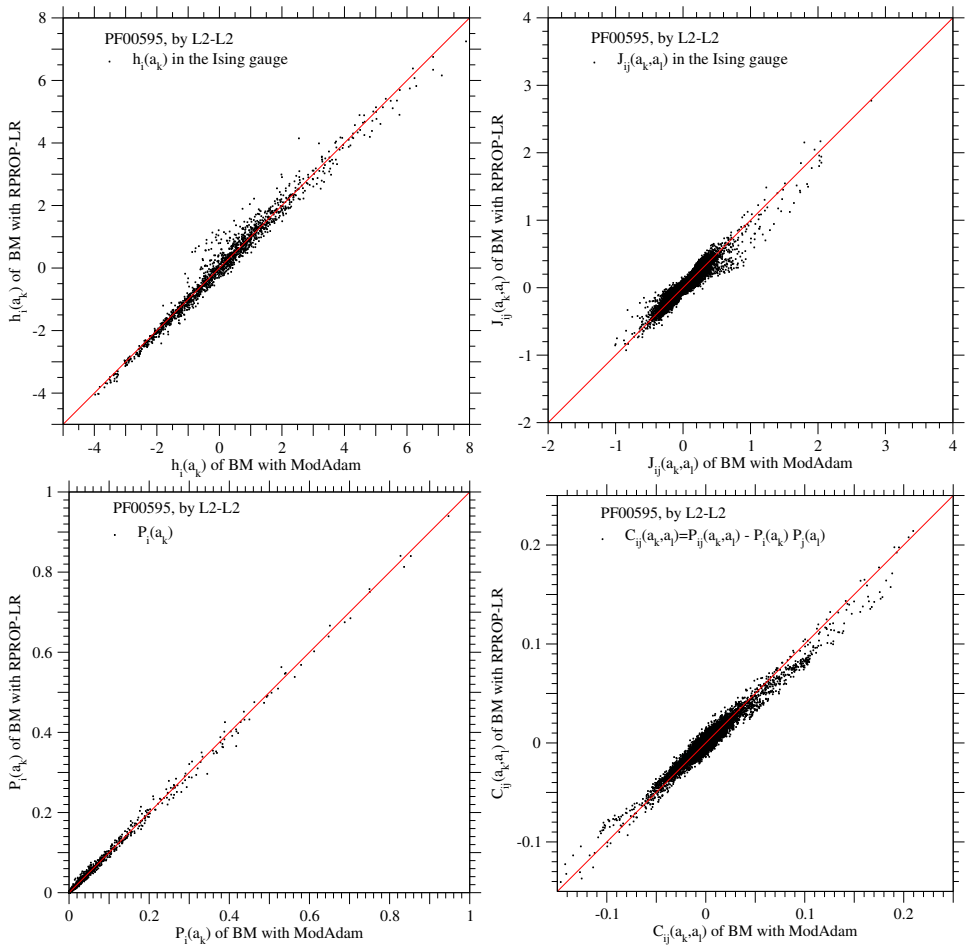

Fig. S4. Comparison of the RPROP-LR with the ModAdam gradient-descent method in each of the inferred fields and couplings and the recovered single-site marginals and pairwise correlations for PF00595. The upper left and upper right figures are the comparisons of the inferred fields and couplings in the Ising gauge, respectively, and the lower left and lower right figures are the comparisons of the recovered single-site frequencies and pairwise correlations, respectively. The abscissas and ordinates correspond to the quantities estimated by the modified Adam and RPROP-LR method for gradient descent, respectively. The regularization model L2-L2 is employed for both methods. The solid lines show the equal values between the ordinate and abscissa. The values of hyper-parameters are listed in Table 2. The overlapped points of $J_{ij}(a_k, a_l)$ in the units 0.001 and of $C_{ij}(a_k, a_l)$ in the units 0.0001 are removed.

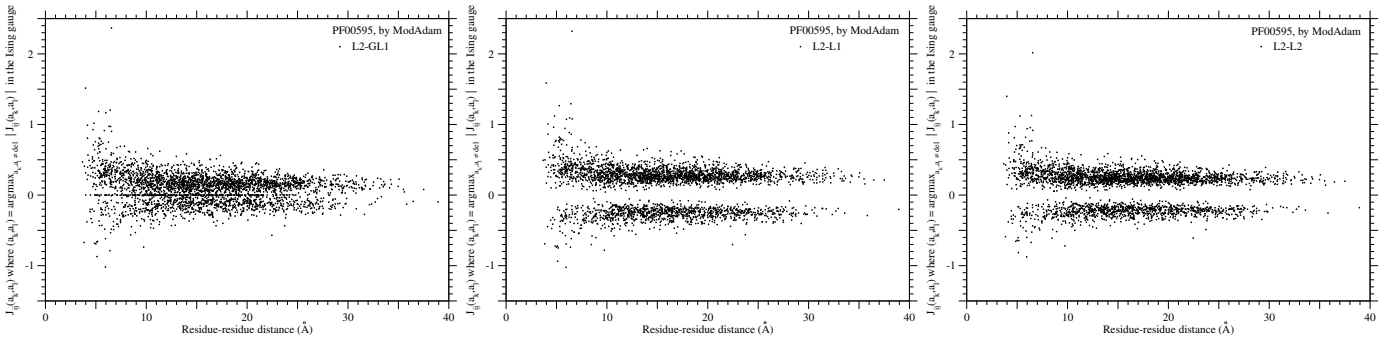

Fig. S5. Differences of inferred couplings J_{ij} among the regularization models for PF00595. All $J_{ij}(a_k, b_l)$ where $(a_k, a_l) = \mathrm{argmax}_{a_k,a_l \neq \mathrm{deletion}} |J_{ij}(a_k, a_l)|$ in the Ising gauge are plotted against the distance between ith and jth residues. The protein family PF00595 is employed. The regularization models L2-GL1, L2-L1, and L2-L2 are employed for the left, middle, and right figures, respectively. The values of regularization parameters are listed in Table 2.

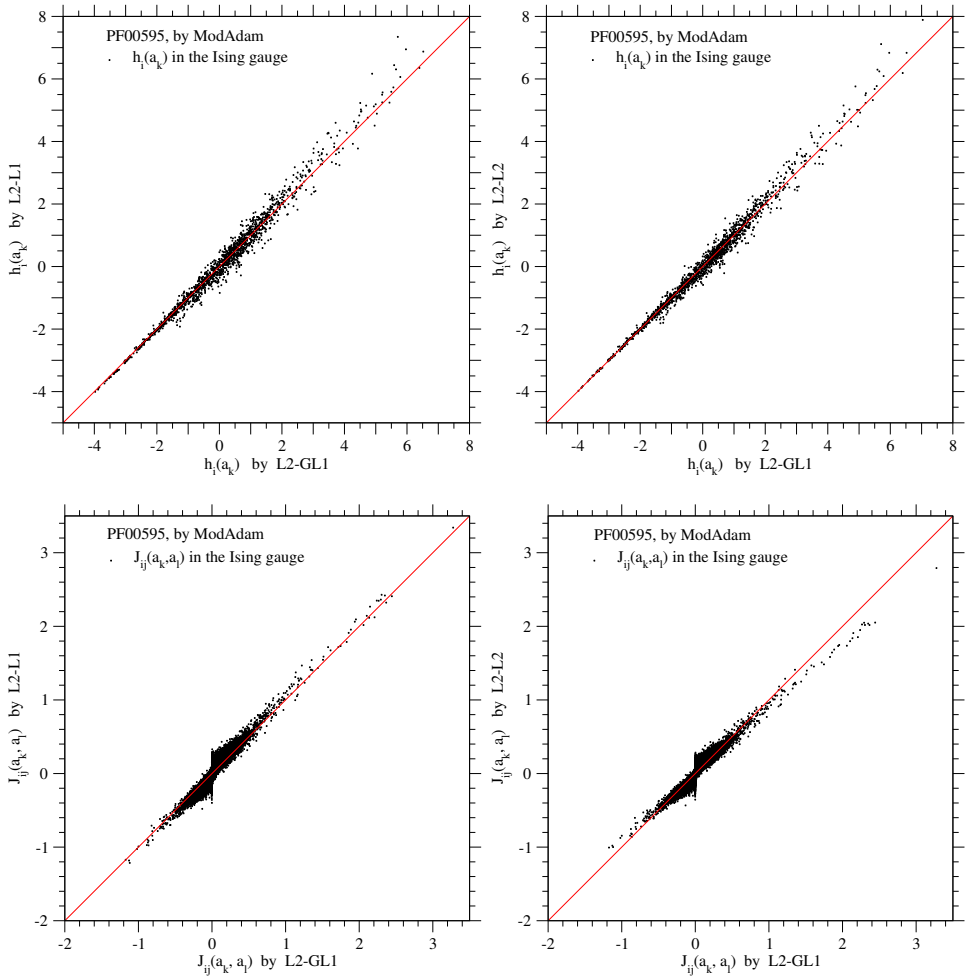

Fig. S6. Comparisons of inferred fields $h_i(a_k)$ and couplings $J_{ij}(a_k, a_l)$ in the Ising gauge between the regularization models for PF00595. The upper and lower figures show the comparisons of fields and couplings in the Ising gauge, respectively. All abscissa correspond to the fields or couplings inferred by the L2-GL1. The ordinates in the left and right figures correspond to the fields or couplings inferred by the L2-L1 and L2-L2 models, respectively. The values of regularization parameters are listed in Table 2. The solid lines show the equal values between the ordinate and abscissa. The overlapped points of $J_{ij}(a_k, a_l)$ in the units 0.001 are removed.

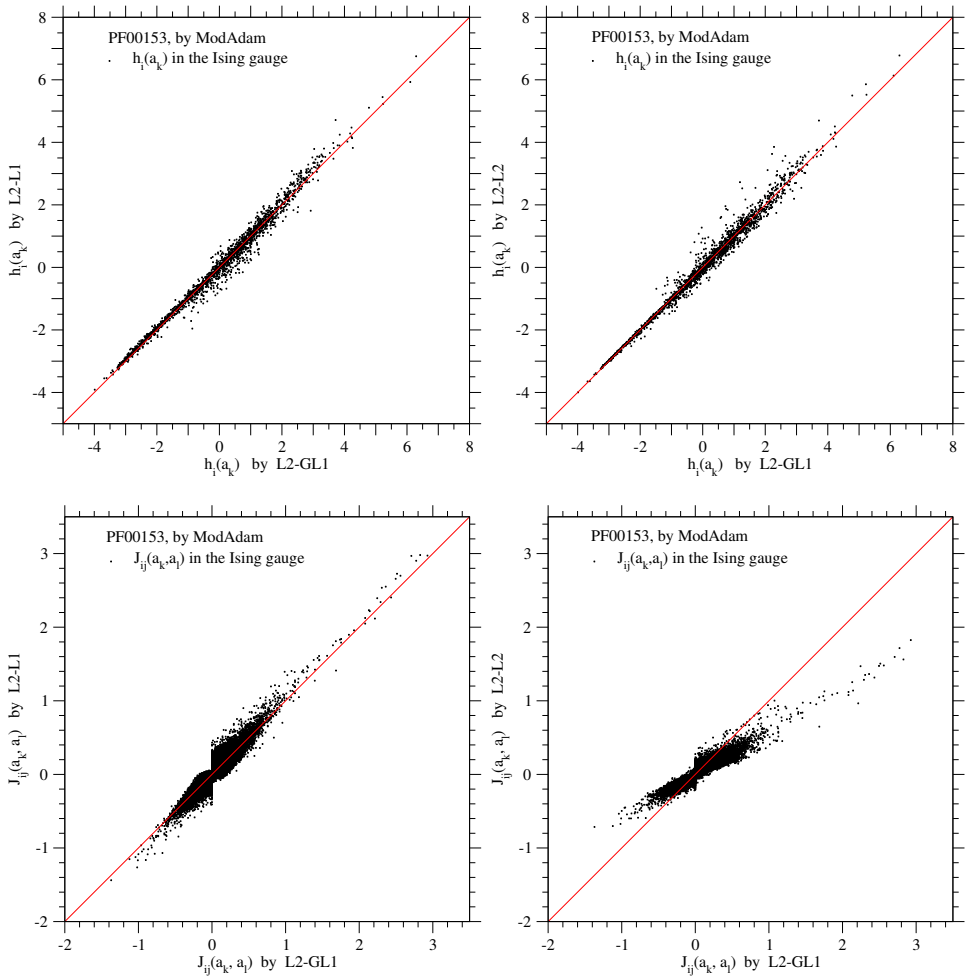

Fig. S7. Comparisons of inferred fields $h_i(a)$ and couplings $J_{ij}(a, b)$ in the Ising gauge between the regularization models for PF00153. The upper and lower figures show the comparisons of fields and couplings in the Ising gauge, respectively. All abscissa correspond to the fields or couplings inferred by the L2-GL1. The ordinates in the left and right figures correspond to the fields or couplings inferred by the L2-L1 and L2-L2 models, respectively. The values of regularization parameters are listed in Table 3. The solid lines show the equal values between the ordinate and abscissa. The overlapped points of $J_{ij}(a_k, a_l)$ in the units 0.001 are removed.

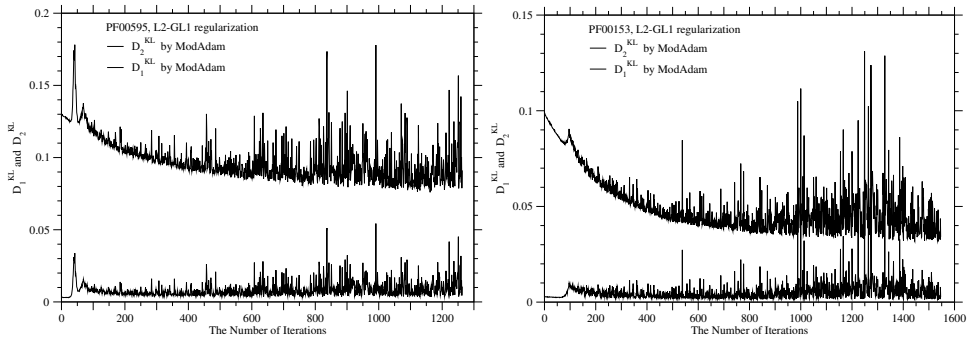

Fig. S8. Learning processes by the L2-GL1 model and the ModAdam method for PF00595 and PF00153. The averages of Kullback-Leibler divergences, D^2_{KL} for pairwise marginal distributions and D^1_{KL} for single-site marginal distributions, are drawn against iteration number in the learning processes with the L2-GL1 model and the ModAdam method for PF00595 and PF00153 in the left and right figures, respectively. The values of hyper-parameters are listed in Tables 2 and 3 as well as others.

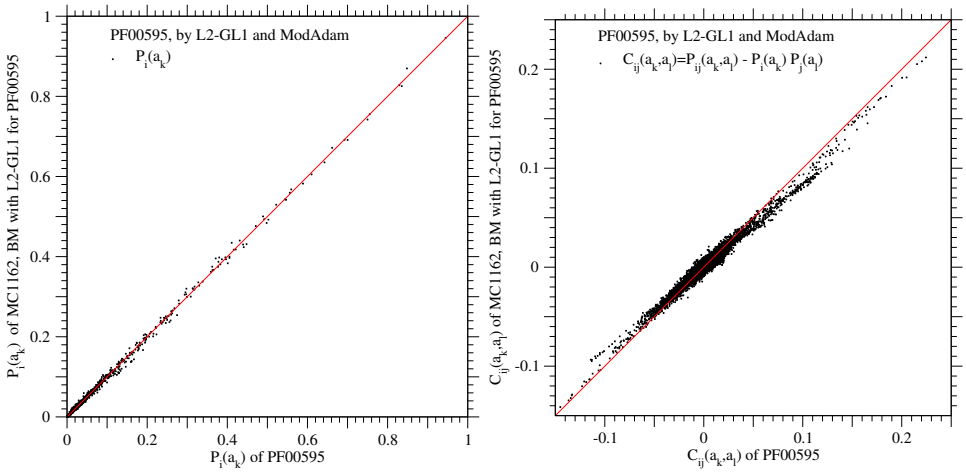

Fig. S9. Recoverabilities of the single-site frequencies and pairwise correlations of PF00595 by the Boltzmann machine learning with the L2-GL1 model and the ModAdam method. The left and right figures are for single-site frequencies and pairwise correlations, respectively; $D_1^{KL} = 0.003695$ and $D_2^{KL} = 0.07594$. The solid lines show the equal values between the ordinate and abscissa. The overlapped points of $C_{ij}(a_k, a_l)$ in the units 0.0001 are removed. See Table 2 for the regularization parameters employed.

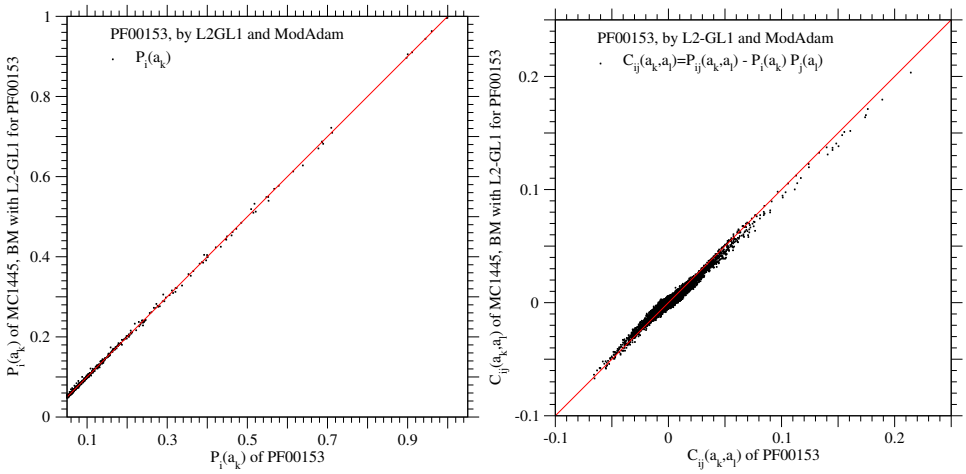

Fig. S10. Recoverabilities of the single-site frequencies and pairwise correlations of PF00153 by the Boltzmann machine learning with the L2-GL1 model and the ModAdam method. The left and right figures are for single-site frequencies and pairwise correlations, respectively; $D_1^{KL} = 0.001120$ and $D_2^{KL} = 0.03176$. The solid lines show the equal values between the ordinate and abscissa. The overlapped points of $C_{ij}(a_k, a_l)$ in the units 0.0001 are removed. See Table 3 for the regularization parameters employed.

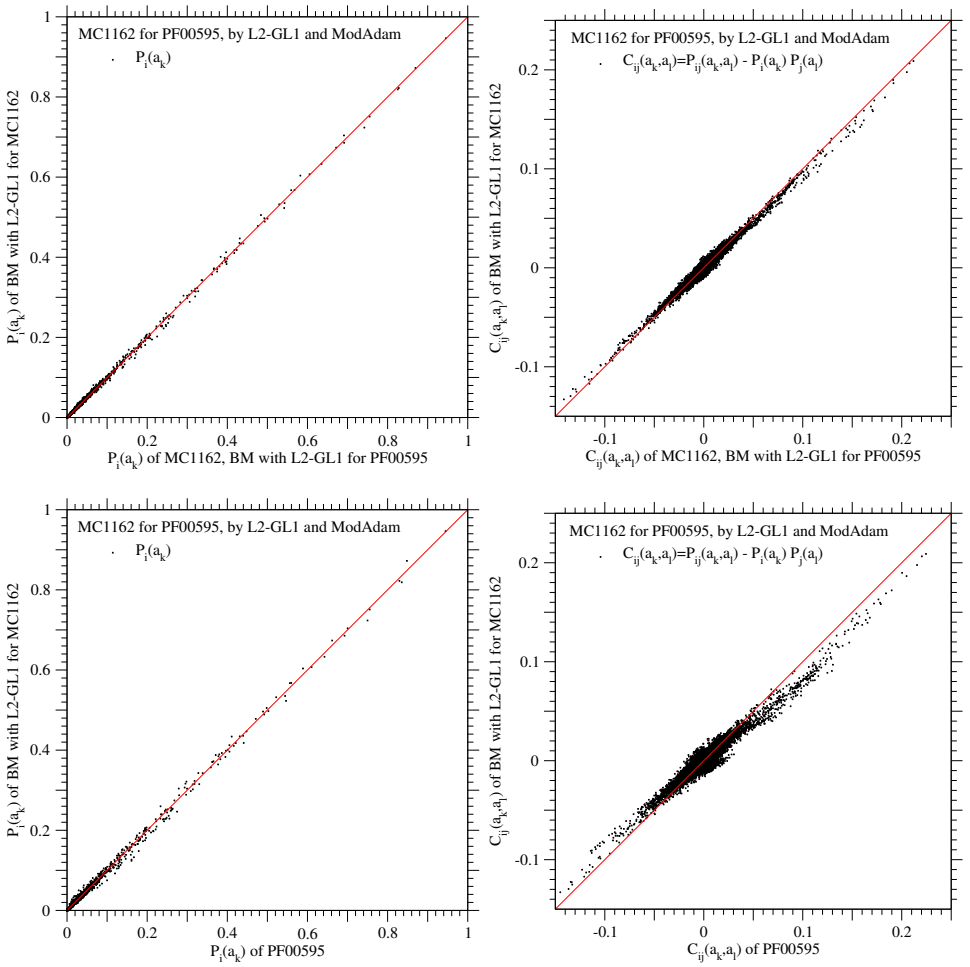

Fig. S11. Recoverabilities of the single-site frequencies and pairwise correlations by the Boltzmann machine learning with the L2-GL1 model and the ModAdam method for the protein-like sequences, the MCMC samples that are obtained by the same Boltzmann machine for PF00595. The MCMC samples obtained by the Boltzmann machine learning with the L2-GL1 model and the ModAdam method for PF00595 are employed as protein-like sequences for which the Boltzmann machine learning with the same model and method is executed again in order to examine how precisely the marginals of the protein-like sequences can be recovered. The marginals recovered by the Boltzmann machine learning for the MCMC samples are compared to those of the MCMC samples in the upper figures, and to those of PF00595 in the lower figures. The left and right figures are for the single-site probabilities and pairwise correlations, respectively. The solid lines show the equal values between the ordinate and abscissa. The overlapped points of $C_{ij}(a_k, a_l)$ in the units 0.0001 are removed. See Table 2 for the regularization parameters employed.

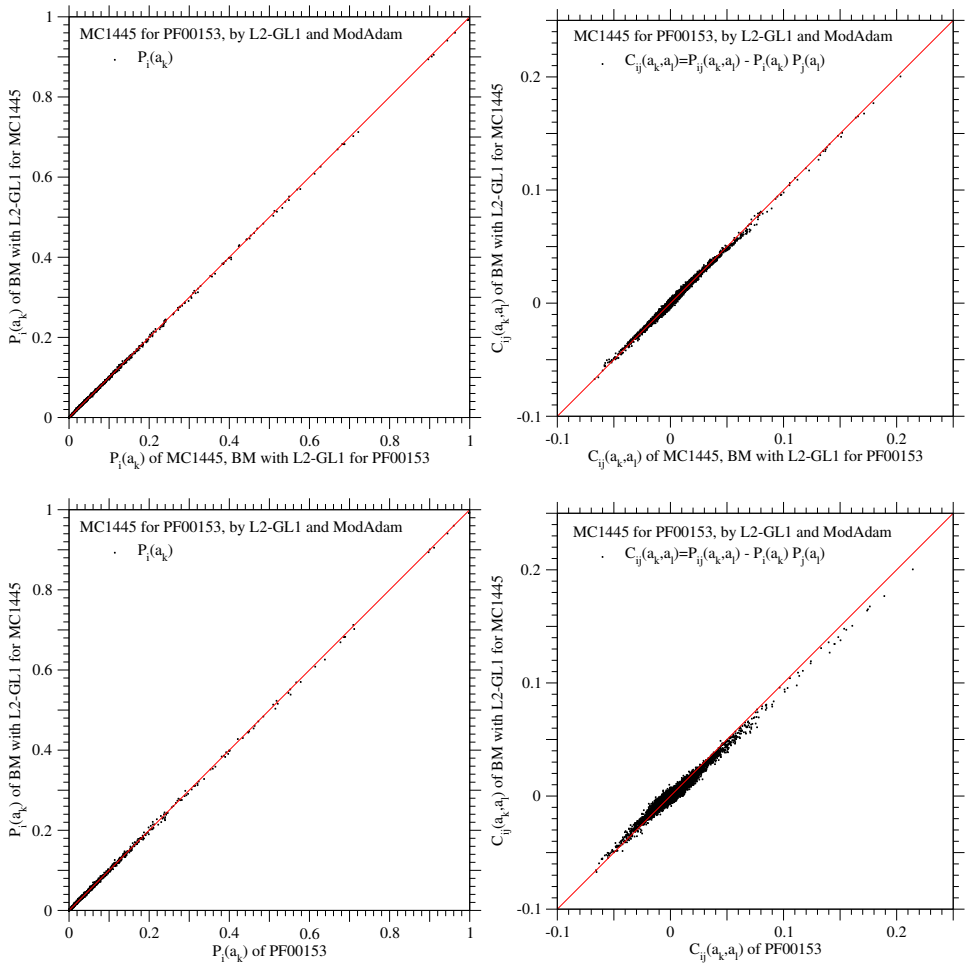

Fig. S12. Recoverabilities of the single-site frequencies and pairwise correlations by the Boltzmann machine learning with the L2-GL1 model and the ModAdam method for the protein-like sequences, the MCMC samples that are obtained by the same Boltzmann machine for PF00153. The MCMC samples obtained by the Boltzmann machine learning with the L2-GL1 model and the ModAdam method for PF00153 are employed as protein-like sequences for which the Boltzmann machine learning with the same model and method is executed again in order to examine how precisely the marginals of the protein-like sequences can be recovered. The marginals recovered by the Boltzmann machine learning for the MCMC samples are compared to those of the MCMC samples in the upper figures, and to those of PF00153 in the lower figures. The left and right figures are for the single-site probabilities and pairwise correlations, respectively. The solid lines show the equal values between the ordinate and abscissa. The overlapped points of $C_{ij}(a_k, a_l)$ in the units 0.0001 are removed. See Table 3 for the regularization parameters employed.