Optimization Problems in Structural Biology with Applications to Protein Design

Mikhail Karasikov

Supervisor:
Prof. Yury Maximov
Skoltech

Co-advisor:
Dr. Sergei Grudinin
Inria

Skolkovo Institute of Science and Technology

Moscow
9 June 2017

The project

Proteins

Amino acid sequences that fold into **spacial shapes** under certain environmental conditions (atomic arrangement in 3D)

The goal

Investigation of the protein design problem
(predict amino acid sequences **given** a protein **spacial shape**)

Applications and importance

Design proteins that posses given properties:

- Disease therapeutics
- Novel enzymes
- Self-assembling proteins/peptides

Challenges of optimization approach to protein design

Issues

- Objective is undefined
- NP-hard optimization problem
- Huge dimensionality
- Requires biological experiments to assess the performance

Problems

1. Define an objective
2. Control the amino acid occurrence in predicted sequences
3. Solve the arisen optimization problem

Problems of computational structural biology for proteins of length m

$$\mathcal{A} = \{\mathrm{Ala}, \mathrm{Arg}, \mathrm{Asn}, \mathrm{Asp}, \mathrm{Cys}, \mathrm{Glu}, \mathrm{Gln}, \mathrm{Gly}, \mathrm{His}, \ldots, \mathrm{Trp}, \mathrm{Tyr}, \mathrm{Val}\}$$

$$\mathrm{Cys} = [\underbrace{\mathrm{N}, \mathrm{C}_\alpha, \mathrm{C}, \mathrm{H}, \mathrm{O}}_{\substack{\text{backbone part} \\ \mathbb{S}_\mathrm{b}=\mathbb{R}^{3\times 3}}}, \underbrace{\mathrm{H}_\alpha, \mathrm{C}_\beta, \mathrm{H}_{\beta_1}, \mathrm{H}_{\beta_2}, \mathrm{S}_\gamma, \mathrm{H}_\gamma}_{\substack{\text{side-chain part (rotamer)} \\ \mathcal{R}}}]$$

$S^m_\mathrm{t} \subset \mathcal{A}^m \times \mathbb{S}^m_\mathrm{b} \times \mathcal{R}^m$ — full proteins

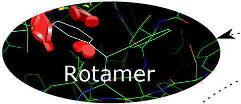
$S^m_\mathrm{r} \subset \mathcal{A}^m \times \mathbb{S}^m_\mathrm{b}$ — reduced proteins

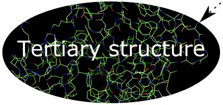
φ_r — rotamer prediction

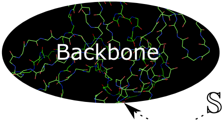
φ_f — protein folding

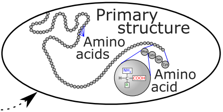
φ_d — protein design

$$\begin{tikzcd}
 & S^m_\mathrm{t} \arrow[dl, "\pi_\mathrm{tb}"'] \arrow[d, shift right, "\pi_\mathrm{tr}"'] \arrow[dr, shift left, "\pi_\mathrm{ta}"'] & \\
 & S^m_\mathrm{r} \arrow[u, shift right, "\varphi_\mathrm{r}"'] \arrow[dl, "\pi_\mathrm{rb}"] \arrow[dr, shift right, "\pi_\mathrm{ra}"'] & \\
\mathbb{S}^m_\mathrm{b} \arrow[r, "\varphi_\mathrm{d}"'] & 2^{\mathcal{A}^m} & \mathcal{A}^m \arrow[l, "\boldsymbol{a}\mapsto\{\boldsymbol{a}\}"] \arrow[ul, shift right, "\varphi_\mathrm{cg}"'] \arrow[uul, shift left, "\varphi_\mathrm{f}"']
\end{tikzcd}$$









Related work

Khoury, G. A., Smadbeck, J., Kieslich, C. A., and Floudas, C. A. (2014).
Protein folding and de novo protein design for biotechnological applications.
Trends in Biotechnology, **32**(2), 99–109.

Samish, I., Macdermaid, C., Perez-Aguilar, J., and Saven, J. (2011).
Theoretical and computational protein design.
Annual Review of Physical Chemistry, **62**(1), 129–149.

Liu, Y., Zeng, J., and Gong, H. (2014).
Improving the orientation-dependent statistical potential using a reference state.
Proteins, **82**(10), 2383–2393.

Protein backbone similarity function $\rho(b', b)$

$\boldsymbol{b} \in \mathbb{S}_{\mathrm{b}}^{m} = \mathbb{R}^{m \times 3 \times 3}$ — a native protein backbone

$\boldsymbol{b}' \in \mathbb{S}_{\mathrm{b}}^{m}$ — an arbitrary protein backbone (model structure)

- Root-mean-square deviation of atomic positions

$$\underbrace{\mathsf{RMSD}(\boldsymbol{b}', \boldsymbol{b})}_{\in[0,\infty)} = \Bigg(\frac{1}{3m} \min_{\substack{\boldsymbol{t}\in\mathbb{R}^3 \\ \mathbf{S}\in\mathrm{SO}(3)}} \sum_{i=1}^{m} \sum_{k=1}^{3} \|\boldsymbol{b}_{ik} - \mathbf{S}\boldsymbol{b}'_{ik} + \boldsymbol{t}\|_2^2\Bigg)^{1/2}$$

- Template modeling score ($\rho_{\mathsf{TM}} = 1 - \mathsf{TM\text{-}score}$)

$$\underbrace{\mathsf{TM\text{-}score}(\boldsymbol{b}', \boldsymbol{b})}_{\in(0,1]} = \frac{1}{m} \max_{\substack{\boldsymbol{t}\in\mathbb{R}^3 \\ \mathbf{S}\in\mathrm{SO}(3)}} \sum_{i=1}^{m} \left(1 + \frac{\|\boldsymbol{b}_{i2} - \mathbf{S}\boldsymbol{b}'_{i2} + \boldsymbol{t}\|_2^2}{d_0^2}\right)^{-1}$$

- **Global distance test scores** ($\rho_{\mathsf{GDT\text{-}TS}} = 1 - \mathsf{GDT\text{-}TS}$)

$$\underbrace{\mathsf{GDT\text{-}TS}(\boldsymbol{b}', \boldsymbol{b})}_{\in[0,1]} = \frac{1}{4m} \max_{\substack{\boldsymbol{t}\in\mathbb{R}^3 \\ \mathbf{S}\in\mathrm{SO}(3)}} \sum_{i=1}^{m} \sum_{j=1}^{4} \mathbb{1}\left[\|\boldsymbol{b}_{i2} - \mathbf{S}\boldsymbol{b}'_{i2} + \boldsymbol{t}\|_2 < c_j\right],$$

$c_{1,2,3,4} = 1, 2, 4, 8\mathrm{A}$, $\mathbb{1}[\cdot]$ — the truth $\{0, 1\}$ predicate.

The protein design problem statement

Given a protein backbone $\boldsymbol{b}^0 \in \mathbb{S}_{\mathrm{b}}^m = \mathbb{R}^{m\times 3\times 3}$ — coordinated of atom triplets [N, C$_\alpha$, C] for m undefined amino acids in 3D space.

Find sequences $\boldsymbol{a} \in \mathcal{A}^m$ that fold to shapes close to backbone $\boldsymbol{b}^0$:

$$\varphi_{\mathrm{d}}(\boldsymbol{b}^0) = \operatorname*{Arg\,min}_{\boldsymbol{a}\in\mathcal{A}^m} \rho(\boldsymbol{b}^0, \underbrace{(\pi_{\mathrm{tb}} \circ \varphi_{\mathrm{f}})(\boldsymbol{a})}_{\text{native backbone of } \boldsymbol{a}}).$$

We propose a two-stage solution

1. Build an approximate scoring function

$$S(\boldsymbol{a}, \boldsymbol{b}^0) \approx S^*(\boldsymbol{a}, \boldsymbol{b}^0) := \rho(\boldsymbol{b}^0, (\pi_{\mathrm{tb}} \circ \varphi_{\mathrm{f}})(\boldsymbol{a}))$$

2. Solve optimization problem

$$S(\boldsymbol{a}, \boldsymbol{b}^0) \to \min_{\boldsymbol{a}\in\mathcal{A}^m}$$

Building a protein backbone scoring function

Given a similarity function $\rho : \bigcup_{m=1}^{\infty} \mathbb{S}_{\mathrm{b}}^m \times \mathbb{S}_{\mathrm{b}}^m \to \mathbb{R}$
and a set of protein backbone domains $\mathcal{D}_1, \ldots, \mathcal{D}_n$:

$$\mathcal{D}_i = \left\{ P_j^i = (\boldsymbol{a}^i, \boldsymbol{b}^{ij}) \mid j = 0, \ldots, t_i \right\} \subset \mathcal{A}^{m_i} \times \mathbb{S}_{\mathrm{b}}^{m_i},$$

where $P_0^i = (\boldsymbol{a}^i, \boldsymbol{b}^{i0}) \in S_{\mathrm{r}}^{m_i}$ is the native protein with sequence $\boldsymbol{a}^i$.

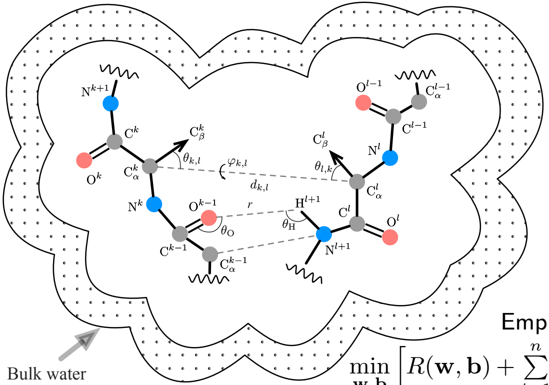
Build a scoring function $S : \bigcup_{m=1}^{\infty} \mathcal{A}^m \times \mathbb{S}_{\mathrm{b}}^m \to \mathbb{R}$ that
approximates actual scoring function S^* on domains $\mathcal{D}_1, \ldots, \mathcal{D}_n$:

$$S(P_j^i) \approx S^*(\boldsymbol{a}^i, \boldsymbol{b}^{ij}) = \rho(\boldsymbol{b}^{ij}, \underbrace{(\pi_{\mathrm{tb}} \circ \varphi_{\mathrm{f}})(\boldsymbol{a}^i)}_{\boldsymbol{b}^{i0}}).$$

Quality criteria:

- Loss$(S; P_0, \mathcal{D}) = \left| \max_{P' \in \mathcal{D} \setminus \{P_0\}} S^*(P') - S^*(\operatorname*{arg\,max}_{P' \in \mathcal{D} \setminus \{P_0\}} S(P')) \right|$,
- Z-score$(S; P_0, \mathcal{D}) = \dfrac{S^*\left(\operatorname*{arg\,max}_{P' \in \mathcal{D} \setminus \{P_0\}} S(P') \right) - \mathbb{E}_{P \sim \mathcal{D} \setminus \{P_0\}} S^*(P)}{\sqrt{\mathbb{D}_{P \sim \mathcal{D} \setminus \{P_0\}} S^*(P)}}$,
- Pearson's, Spearman's rank, and Kendall's tau corr. coeff.

Model and features



Feature extraction:

$$\mathbf{f} : \bigcup_{m=1}^{\infty} \mathcal{A}^m \times \mathbb{S}_b^m \to \mathbb{R}^k$$

Linear model:

$$S(P) = \langle \mathbf{w}, \mathbf{f}(P) \rangle$$

$$\tilde{\mathbf{f}}(P_j^i) := \begin{bmatrix} \mathbf{f}(P_j^i) \\ \beta \boldsymbol{e}_i \end{bmatrix} \in \mathbb{R}^{k+n}$$

$$\tilde{\mathbf{w}} := \begin{bmatrix} \mathbf{w} \\ \boldsymbol{b} \end{bmatrix} \in \mathbb{R}^{k+n}$$

Empirical risk minimization:

$$\min_{\mathbf{w},\mathbf{b}} \left[R(\mathbf{w}, \mathbf{b}) + \sum_{i=1}^{n} \sum_{j=0}^{t_i} L\left(S(P_j^i) + b_i,\ S^*(P_j^i, P_0^i) \right) \right]$$

$$\alpha \left(\|\mathbf{w}\|_2^2 + \frac{1}{\beta^2} \|\mathbf{b}\|_2^2 \right) + \sum_{i=1}^{n} \sum_{j=0}^{t_i} \left(S(P_j^i) + b_i - S^*(P_j^i, P_0^i) \right)^2 \to \min_{\mathbf{w},\mathbf{b}}$$

$$\alpha \|\tilde{\mathbf{w}}\|_2^2 + \sum_{i=1}^{n} \sum_{j=0}^{t_i} \left(\left\langle \tilde{\mathbf{w}}, \tilde{\mathbf{f}}(P_j^i) \right\rangle - S^*(P_j^i, P_0^i) \right)^2 \to \min_{\tilde{\mathbf{w}}}$$ — ridge regression

Protein design as optimization of pairwise decomposable function

Proposed scoring function is pairwise decomposable:

$$S(\boldsymbol{a}, \boldsymbol{b}) = \sum_{k=1}^{m} \sum_{l=1}^{m} E_{kl}^{\boldsymbol{b}}(a_k, a_l) \to \min_{\boldsymbol{a} \in \mathcal{A}^m} .$$

Suppose $\mathcal{A} = \{a^1, \dots, a^t\}$. Reduction to BQP:

$$\sum_{k,l=1}^{m} E_{kl}^{\boldsymbol{b}}(a_k, a_l) = \sum_{k,l=1}^{m} \sum_{i,j=1}^{t} E_{kl}^{\boldsymbol{b}}(a^i, a^j) \underbrace{\mathbb{1}[a_k = a^i]}_{x_i^k} \underbrace{\mathbb{1}[a_l = a^j]}_{x_j^l} .$$

Assume $\mathbf{Q} = \left[[E_{kl}^{\boldsymbol{b}}(a^i, a^j)]_{i,j=1}^{t} \right]_{k,l=1}^{m}$. Equivalent BQP problem:

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & \boldsymbol{x}^\intercal \mathbf{Q} \boldsymbol{x} \\ \text{subject to} \quad & \boldsymbol{x} = [\boldsymbol{x}^{1\intercal}, \dots, \boldsymbol{x}^{m\intercal}]^\intercal \\ & \boldsymbol{x}^k \in \{0, 1\}^t, \quad k = 1, \dots, m, \\ & \|\boldsymbol{x}^k\|_0 = 1, \qquad k = 1, \dots, m. \end{aligned}$$

Probabilistic problem statement for protein design

Posterior probability maximization

$$p(\boldsymbol{a}|\boldsymbol{b}^0) \to \max_{\boldsymbol{a}\in\mathcal{A}^m},$$

where likelihood is defined as follows:

$$p(\boldsymbol{b}^0|\boldsymbol{a}) \propto \exp\left(-\frac{\sum_{k=1}^m \sum_{l=1}^m E_{kl}^{\boldsymbol{b}}(a_k, a_l)}{T}\right).$$

Note that

$$p(\boldsymbol{b}^0|\boldsymbol{a}) \to \max_{\boldsymbol{a}\in\mathcal{A}^m} \iff \sum_{k=1}^m \sum_{l=1}^m E_{kl}^{\boldsymbol{b}}(a_k, a_l) \to \min_{\boldsymbol{a}\in\mathcal{A}^m}$$

Let us introduce the prior distribution

$$p(a_1, \ldots, a_m) = C \prod_{a\in\mathcal{A}} \mathcal{N}\left(m_a \middle| \, mp_a, m\sigma_a^2\right),$$

where

$\mathcal{N}\left(x \middle| \, \mu, \sigma^2\right) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$, $m_a = \sum_{i=1}^m \mathbb{1}[a_i = a]$, $a \in \mathcal{A}$,
p_a and σ_a for each $a \in \mathcal{A}$ — the distribution parameters,
C is the normalizing constant.

Energy corrections for prior distribution

Lemma (Energy corrections, Karasikov 2016)

Let prior distribution $p(a_1, \ldots, a_m)$ *be defined by formula* $p(a_1, \ldots, a_m) = C \prod\limits_{a \in \mathcal{A}} \mathcal{N}\left(m_a \middle| \, mp_a, m\sigma_a^2\right).$

Then the problem of maximization the posterior probability $p(\boldsymbol{a}|\boldsymbol{b}^0) \to \max\limits_{\boldsymbol{a} \in \mathcal{A}^m}$ *is equivalent to minimization of the total energy*

$$\sum_{k=1}^{m} \sum_{l=1}^{m} \left[E_{kl}(a_k, a_l) + E'_{kl}(a_k, a_l)\right] \to \min_{a_1, \ldots, a_m},$$

where energy correction terms are introduced as follows:

$$E'_{kl}(a_k, a_l) := \begin{cases} \frac{T}{2m} \cdot \frac{1 - 2p_{a_k}}{\sigma_{a_k}^2}, & a_k = a_l; \\ -\frac{T}{2m} \cdot \left(\frac{p_{a_k}}{\sigma_{a_k}^2} + \frac{p_{a_l}}{\sigma_{a_l}^2}\right), & a_k \neq a_l. \end{cases}$$

Computational experiment

Goals:

1. Investigation of the scoring quality depending on the size of the training set and width of the kernel for feature smoothing
2. To compare the quality of proposed scoring function with state-of-the-art methods
3. Investigation of the contribution of energy corrections introduced to regulate the occurrence ratio of amino acids of different types in predicted sequences

Data:

- Protein models from the CASP[5-11] competition
- Per 300 NMA protein models for each native from CASP within RMSD range $[0.5, 6]\mathrm{A}$ using 100 first normal modes
- Native protein structure from the test set of SCWRL4

Dependency on smoothing

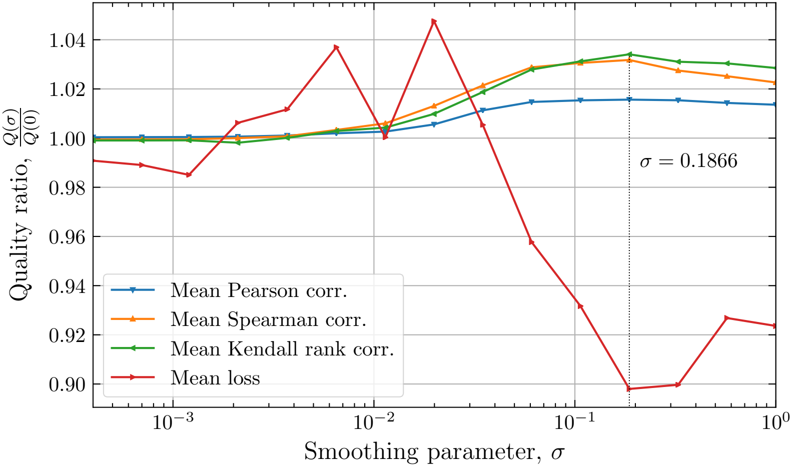


Figure: Performance on the CASP10 (stage1 and stage2 together) dataset depending on width of the smoothing kernel $\sigma^a = \sigma^r = \sigma^h = \sigma^s = \sigma$ when training on the CASP[5-9] datasets without smoothing ($\sigma = 0$).

Learning curves

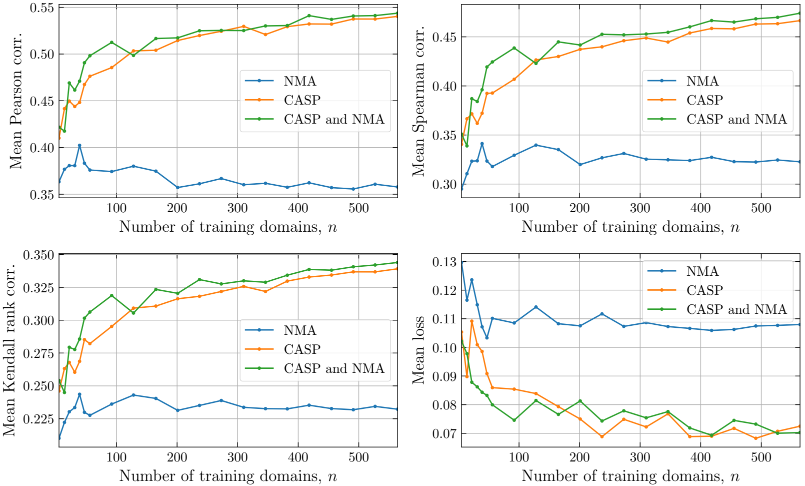


Figure: Learning curves for Performance on validation depending on the number of backbone domains used when training. Training: random subsets of CASP[5-10]. Validation: CASP11 (stage1 and stage2 together).

Performance comparison

QA Method	CASP11 Stage1			CASP11 Stage2		
	Loss	PCC	SCC	Loss	PCC	SCC
This study	**0.083**	0.645	0.522	**0.057**	**0.441**	**0.426**
ProQ2	0.090	0.643	0.506	0.058	0.372	0.366
VoroMQA	0.108	0.561	0.426	0.069	0.401	0.386
Wang-SVM	0.109	**0.655**	**0.535**	0.085	0.362	0.351
Dope	0.111	0.542	0.416	0.077	0.304	0.324
RWplus	0.135	0.536	0.433	0.084	0.295	0.314

Table: Performance on the CASP11 dataset. Quality criteria: Mean metric loss (Loss), Pearson (PCC), and Spearman (SCC) correlation coefficients between predicted scores and actual scores S^*. Trained on CASP[5-10].

Contribution of energy correction terms

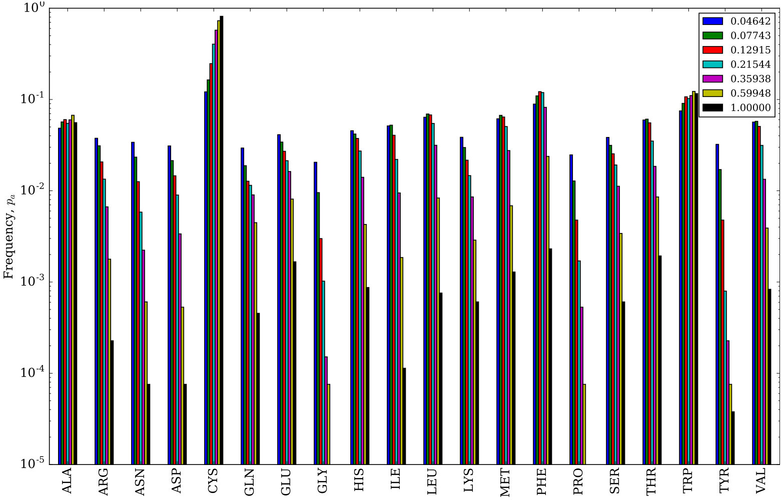


Figure: Average occurrence frequency of different amino acids in predicted sequences depending on temperature factor $\beta = 1/T$. Averaged on dataset SCWRL4.

Conclusions

Proposed protein scoring function:

- uses interpretable **physical model**;
- uses only conformation of the **backbone**;
- is **robust** to errors in side-chain positions;
- is **smooth** function of atomic positions;
- achieves the **state-of-the-art** performance;
- is **pairwise decomposable**.

Proposed energy correction terms:

- control the frequency of occurrence of different amino acids in predicted sequences

Further research directions:

- Outliers detection in the training set
- Incorporation of the physics-based features
- Experiments of protein design performance based on the proposed scoring function

Personal contribution

- Proposed a novel method for protein scoring function
- Developed a program package for the protein quality assessment and executables for Windows, Linux, and MacOS
- Proposed energy correction terms for regulating the frequency of occurrence of different amino acids in sequences predicted
- Conducted experimental comparison of different techniques for convex relaxations and general methods of discrete optimization when solving the arising BQP problem that proved the applicability of the semidefinite relaxation with random sampling as the best method among the tested ones

Backup: convex relaxations for BQP

$$\underset{\boldsymbol{x}\in\{0,1\}^n}{\text{minimize}} \quad \boldsymbol{x}^\mathsf{T}\mathbf{Q}\boldsymbol{x}$$
$$\text{subject to} \quad \mathbf{A}\boldsymbol{x} = \mathbf{1}_m,$$

1. Continuous

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} \left\{\boldsymbol{x}^\mathsf{T}(\mathbf{Q}-\lambda_{\min}\mathbf{I}_n)\boldsymbol{x} + \lambda_{\min}\mathbf{1}_n^\mathsf{T}\boldsymbol{x} \;\middle|\; \mathbf{A}\boldsymbol{x} = \mathbf{1}_m,\ \mathbf{0}_n \leqslant \boldsymbol{x} \leqslant \mathbf{1}_n\right\}$$

2. Lagrange (dual problem)

$$\min_{\boldsymbol{\lambda}\in\mathbb{R}^n,\,\boldsymbol{u}\in\mathbb{R}^m} \left\{\gamma - r(\boldsymbol{u}) \;\middle|\; \gamma \leqslant 0,\ \begin{bmatrix} \mathbf{P}(\boldsymbol{\lambda}) & \frac{1}{2}\boldsymbol{q}(\boldsymbol{\lambda},\boldsymbol{u}) \\ \frac{1}{2}\boldsymbol{q}^\mathsf{T}(\boldsymbol{\lambda},\boldsymbol{u}) & -\gamma \end{bmatrix} \in \mathcal{S}_+^{n+1}\right\}$$

3. Positive semidefinite (SDP)

$$\min_{\boldsymbol{x},\,\mathbf{X}} \left\{\mathrm{Tr}\,(\mathbf{Q}\mathbf{X}) \;\middle|\; \begin{matrix} \mathbf{A}\boldsymbol{x} = \mathbf{1}_m, \\ \mathbf{A}\mathbf{X} = \mathbf{1}_m\boldsymbol{x},^\mathsf{T} \end{matrix} \quad \begin{matrix} X_{ij} \in [0,1], \\ X_{ii} = x_i, \\ i,j = 1,\ldots,n, \end{matrix} \quad \begin{bmatrix} \mathbf{X} & \boldsymbol{x} \\ \boldsymbol{x}^\mathsf{T} & 1 \end{bmatrix} \in \mathcal{S}_+^{n+1}\right\}$$

Backup: relaxation → approximate solution

Random sampling for SDP relaxation

$$\boldsymbol{x}' \sim \mathcal{N}(\boldsymbol{x}, \underbrace{\mathbf{X} - \boldsymbol{x}\boldsymbol{x}^\intercal}_{\in \mathcal{S}_+^n})$$

Rounding

Projection $\hat{\boldsymbol{x}} \in \mathsf{Proj}_V \boldsymbol{x}$, where $V = \{\boldsymbol{x} \in \{0,1\}^n \mid \mathbf{A}\boldsymbol{x} = \mathbf{1}_m\}$ is computed as follows:

$$\hat{x}_i^k := \begin{cases} 1, & i = \underset{j=1,\dots,t}{\widehat{\arg\max}}\, x_j^k, \\ 0, & \text{otherwise}, \end{cases}$$

where $k = 1, \dots, m$, $\underset{j}{\widehat{\arg\max}}\, x_j = \min(\underset{j}{\operatorname{Arg\,max}}\, x_j)$.

Backup: optimization results for rotamer prediction

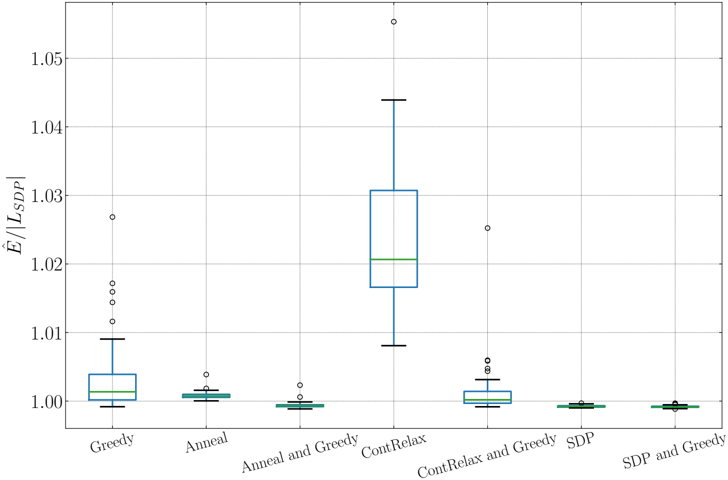


Figure: Box plots for normalized approximate optimal values obtained by different optimization methods. Averaged over first 40 structures from the SCWRL4 dataset.

Backup: optimization results for protein design

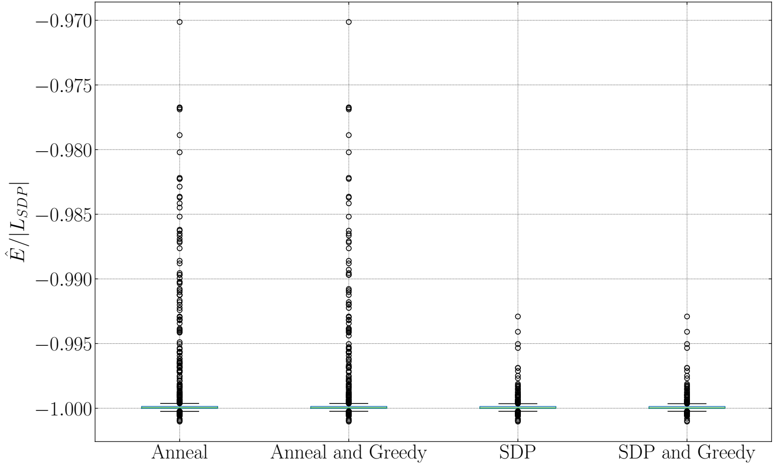


Figure: Box plots for normalized approximate optimal values obtained by different optimization methods. Averaged over structures from the SCWRL4 dataset and sequence lengths $m = 5, 10, 15, 20, 25, 30$.

Backup: optimization results for protein design

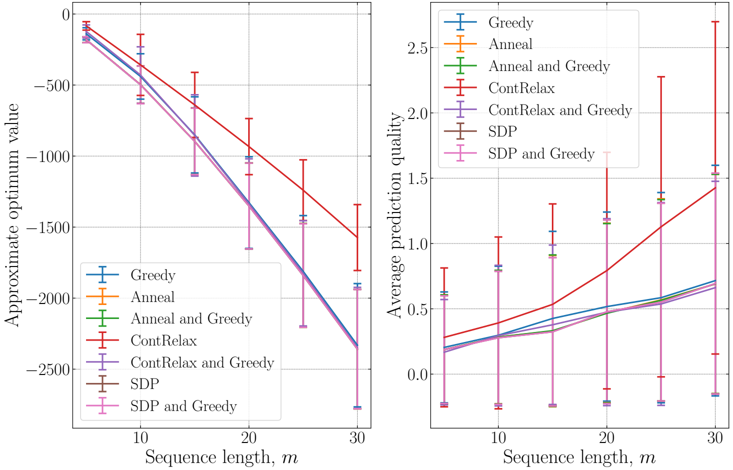
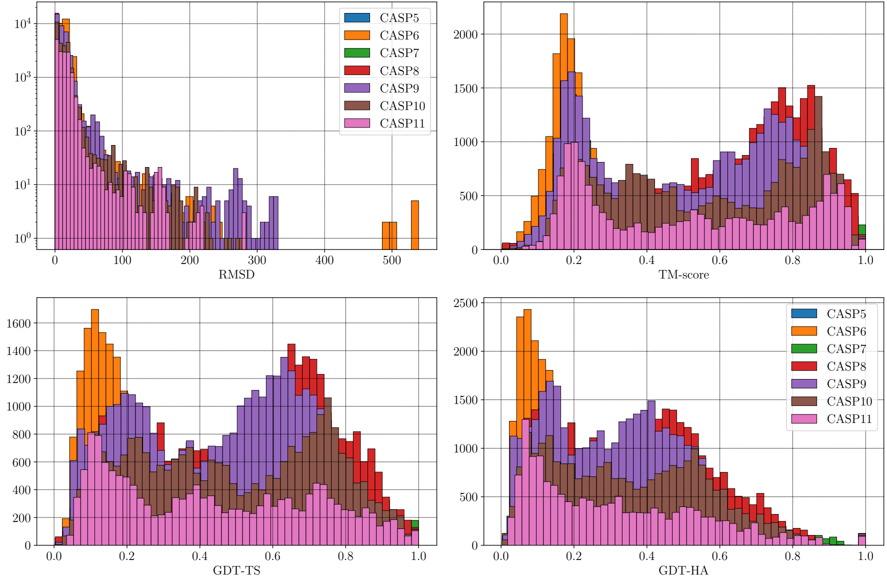


Figure: Upper bounds on the optimal value and Average ratio of correctly predicted amino acids. Averaged over 352 protein structures in the SCWRL4 dataset. DFIRE-C$_\alpha$ (Zhang et al., 2004) energy function.

Backup: distributions for different protein backbone scores


CASP5
CASP6
CASP7
CASP8
CASP9
CASP10
CASP11
10^0
10^1
10^2
10^3
10^4
RMSD
TM-score
GDT-TS
GDT-HA
CASP5
CASP6
CASP7
CASP8
CASP9
CASP10
CASP11