Efficient Generative Modelling of Protein Structure Fragments using a Deep Markov Model

*Christian B. Thygesen[1,2], Ahmad Salim Al-Sibahi[1], Christian S. Steenmans[2], Lys S. Moreta[1], Anders B. Sørensen[*2], Thomas Hamelryck[*1]*

[1] Department of Computer Science, University of Copenhagen; [2] Evaxion Biotech A/S; [*] Equal contribution

ABSTRACT

Fragment libraries are often used in protein structure prediction, simulation and design as a means to significantly reduce the vast conformational search space.

Current state-of-the-art methods for fragment library generation do not properly account for aleatory and epistemic uncertainty, respectively due to the dynamic nature of proteins and experimental errors in protein structures.

Additionally, they typically rely on information that is not generally or readily available, such as homologous sequences, related protein structures and other complementary information.

To address these issues, we developed BIFROST, a novel take on the fragment library problem based on a Deep Markov Model architecture combined with directional statistics for angular degrees of freedom, implemented in the deep probabilistic programming language Pyro.

BIFROST is a probabilistic, generative model of the protein backbone dihedral angles conditioned solely on the amino acid sequence.

BIFROST generates fragment libraries with a quality on par with current state-of-the-art methods at a fraction of the runtime, while requiring considerably less information and allowing efficient evaluation of probabilities.

INTRODUCTION

Fragment libraries [1] find wide application in protein structure prediction, simulation, design and experimental determination [2-4]. Fragment libraries are used in a divide-and-conquer approach, whereby a full-length protein is divided into a manageable sub-set of shorter stretches of amino acids for which backbone conformations (figure 1) are sampled. Typically, sampling is done using a finite set of fragments derived from experimentally determined protein structures. Fragment libraries are used in state-of-the-art protein structure prediction frameworks such as Rosetta [5], I-TASSER [6], and AlphaFold [7].

Here, we present BIFROST - Bayesian Inference for FRagments Of protein STructures – the first deep, generative, probabilistic model of protein backbone angles that solely uses the amino acid sequence as input. BIFROST is based on a Deep Markov Model (DMM) architecture [8] and represents the angular variables (ϕ and ψ) in a principled way using directional statistics [9].

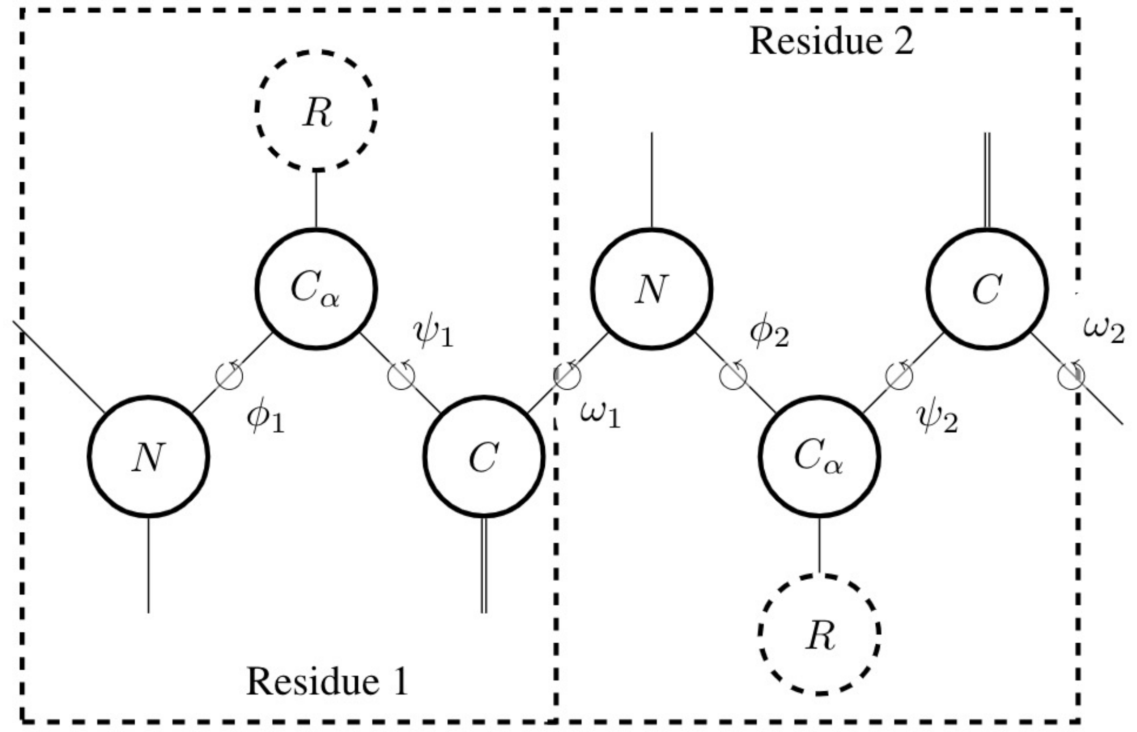
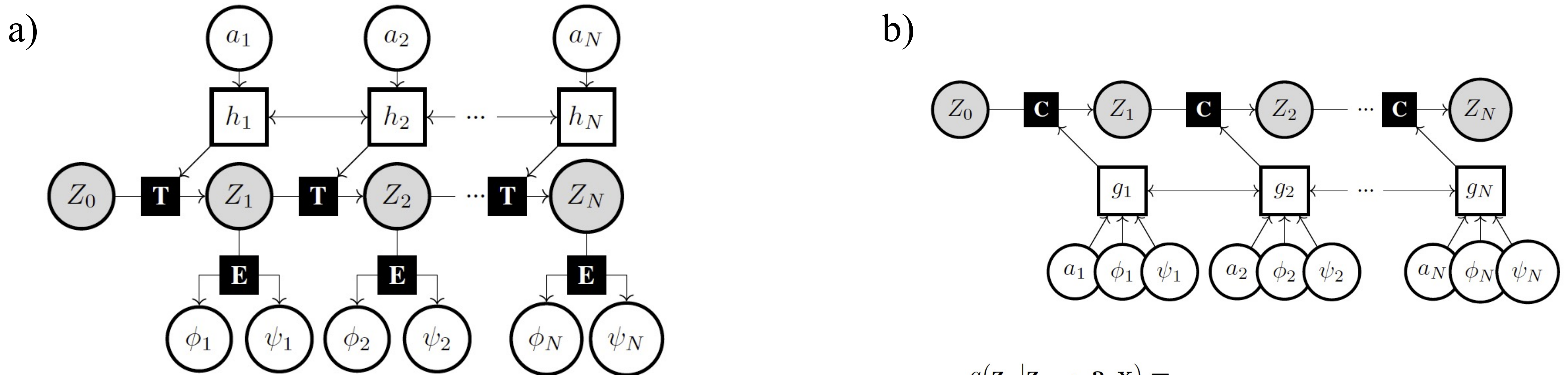


Figure 1: Schematic of the three dihedral angles (ϕ, ψ, and ω) that parameterise the protein backbone. R represents the side chain.

MATERIALS & METHODS

BIFROST consists of a DMM with an architecture similar to an Input-Output HMM (IO-HMM) [10]. The model employs the Markovian structure of an HMM, but with continuous latent states (z) and with neural networks parameterising the transition and emission densities. The model was extended with a bidirectional recurrent neural network processing the amino acid sequence. This processed information is passed to the transition (T) network to transform the previous latent state. The periodic angle values are modelled through a wrapped student t distribution parameterised by an emitter (E) neural network. The structure of the model is shown in figure 2a along with the factorised joint distribution over latent states and the sequence of ϕ/ψ angle pairs, denoted x, given the amino acid sequence, denoted a (equation 1-3). The model is trained through stochastic variational inference with a variational distribution (guide). The guide has a similar structure to the model (figure 2b) but parses observed angles along with the amino acid sequence to infer distributions over latent states given the amino acid sequence and angles (equation 4).



$$p(\mathbf{z},\mathbf{x}|\mathbf{a}) = \prod_{n=1}^{N} p(\mathbf{z}_n|\mathbf{z}_{n-1},\mathbf{h}_n(\mathbf{a}))p(\mathbf{x}_n|\mathbf{z}_n) \qquad (1)$$

$$p(\mathbf{z}_n|\mathbf{z}_{n-1},\mathbf{h}_n(\mathbf{a})) = \mathcal{N}(\boldsymbol{\mu}_T(\mathbf{z}_{n-1},\mathbf{h}_n(\mathbf{a})),\boldsymbol{\Sigma}_T(\mathbf{z}_{n-1},\mathbf{h}_n(\mathbf{a}))) \qquad (2)$$

$$p(\mathbf{x}_n|\mathbf{z}_n) = \mathcal{T}(\mathbf{x}_n|\nu_E(\mathbf{z}_n),\boldsymbol{\mu}_E(\mathbf{z}_n),\boldsymbol{\Sigma}_E(\mathbf{z}_n)) \qquad (3)$$

$$q(\mathbf{z}_n|\mathbf{z}_{n-1},\mathbf{a},\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_C(\mathbf{z}_{n-1},\mathbf{g}_n(\mathbf{a},\mathbf{x})),\boldsymbol{\Sigma}_C(\mathbf{z}_{n-1},\mathbf{g}_n(\mathbf{a},\mathbf{x}))) \qquad (4)$$

Figure 2: The BIFROST model (a) and variational distribution (b). Grey nodes are latent random variables, white circular nodes are observed variables, white rectangular nodes represent hidden states from bidirectional Recurrent Neural Networks (RNNs), and black squares represent neural networks. E and T denote the emitter and the transition networks, respectively, while C denotes the combiner network.

Implemented in Pyro

RESULTS

The model was able to recreate the observed Ramachandran plots, of 5000 previously unseen sequences, with minimal added noise (figure 3a). Additionally, BIFROST captures individual amino acid properties (figure 3b), as evidenced by the individual Ramachandran plots of the flexible residue glycine, the rigid residue proline, and leucine to represent the general behavior of the remaining amino acids. Finally, BIFROST can model sequence dependency as evidenced in figure 3c, showing superimposed cartoon representations of 100 backbone samples from BIFROST (blue) conditioned on fragments that were observed to be either α-helix, β-strand, or coil (yellow).

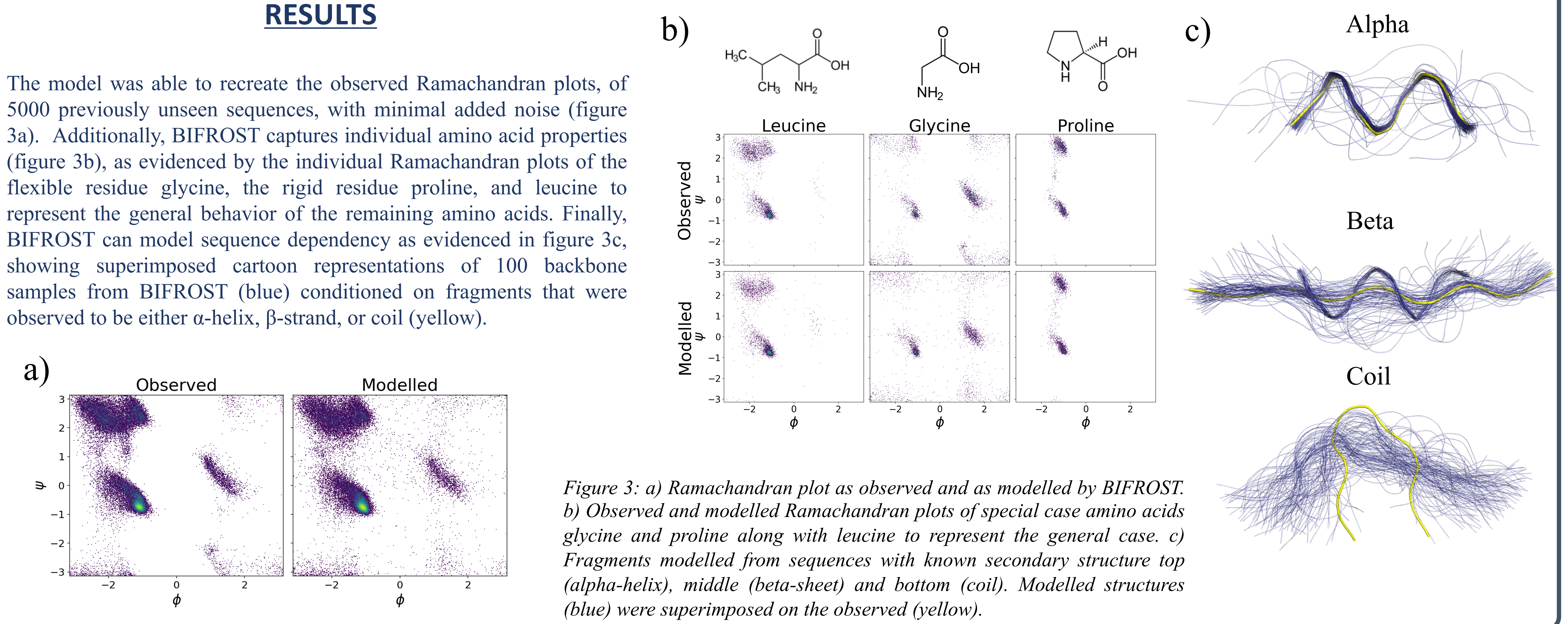


Figure 3: a) Ramachandran plot as observed and as modelled by BIFROST. b) Observed and modelled Ramachandran plots of special case amino acids glycine and proline along with leucine to represent the general case. c) Fragments modelled from sequences with known secondary structure top (alpha-helix), middle (beta-sheet) and bottom (coil). Modelled structures (blue) were superimposed on the observed (yellow).

Benchmarking

BIFROST was benchmarked against Rosetta's fragment picker, which selects fragments from a database of experimentally determined structures based on auxiliary information such as secondary structure predictions.

BIFROST performs on par with the fragment picker and shows similar RMSD distributions stratified by secondary structure, despite relying only on sequence, while doing so with an improved runtime (figure 4).

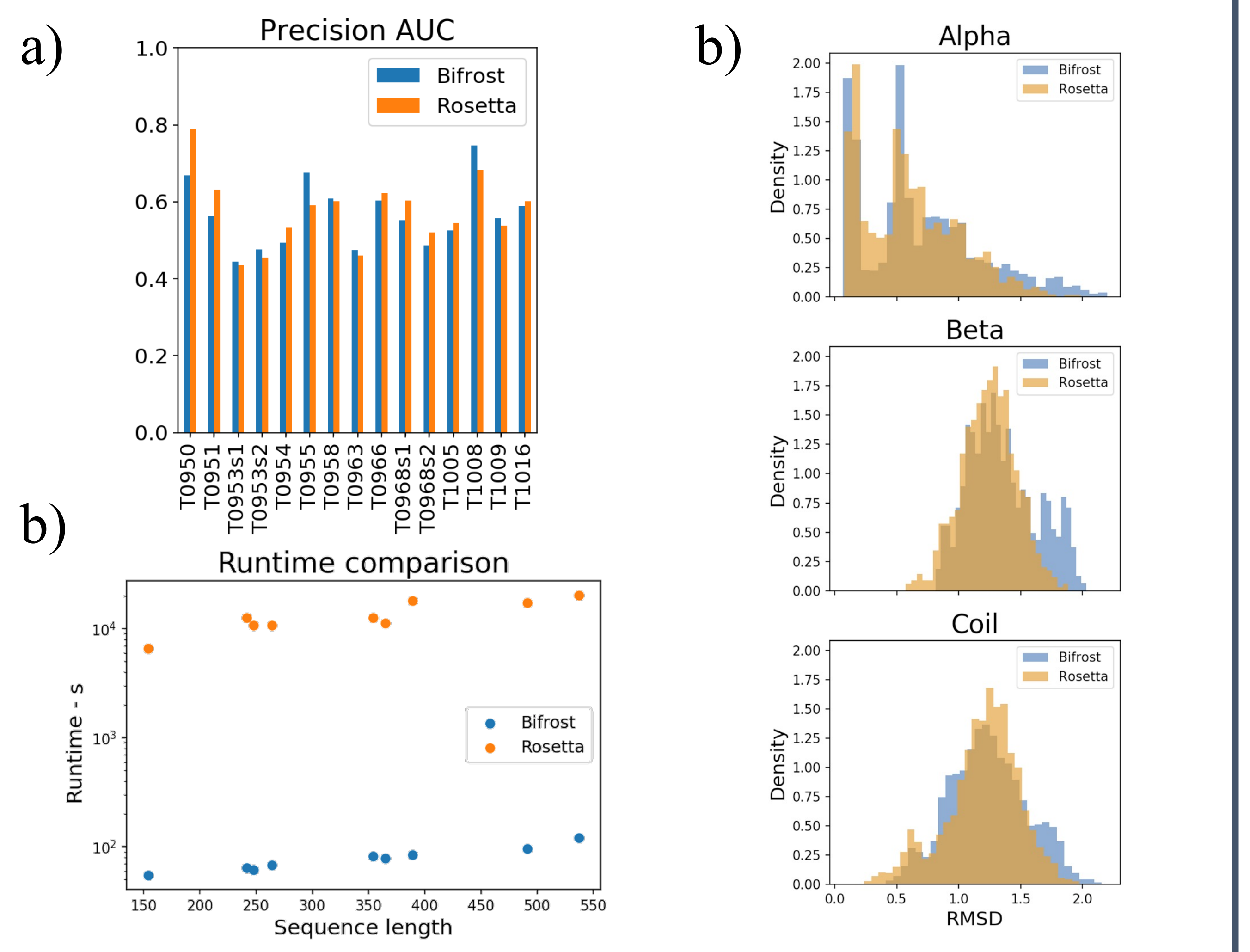


Figure 4 a): Precision AUC of fragment libraries for CASP13 regular (T) targets generated by BIFROST on and Rosetta's fragment picker. b): Runtime of BIFROST and Rosetta for generating fragment libraries for proteins of varying sequence length. c): RMSD distributions stratified by secondary structure.

REFERENCES

[1] Jones & Thirup, The EMBO journal, 1986

[2] Trevizani et al, PLoS ONE, 2017

[3] Chikenji et al, Proceedings of the National Academy of Sciences, 2006

[4] Boomsma et al, Springer, 2012

[5] Rohl et al, Methods in Enzymology, 2004

[6] Roy et al, Nature Protocols, 2010

[7] Senior et al, Proteins: Structure, Function and Bioinformatics, 2019

[8] Krishnan et al, 31st AAAI Conference on Artificial Intelligence, 2017.

[9] Mardia & Jupp, Directional Statistics, Wiley Series in Probability and Statistics, 2008

[10] Bengio & Frasconi. Neural Information Processing Systems 1995

Acknowledgements

We acknowledge funding from the Innovation Fund Denmark under the grant "Accelerating vaccine development through a deep learning and probabilistic programming approach to protein structure prediction".