Protein Language Model-based Assignment of NMR ^{1}H-^{15}N Chemical Shifts

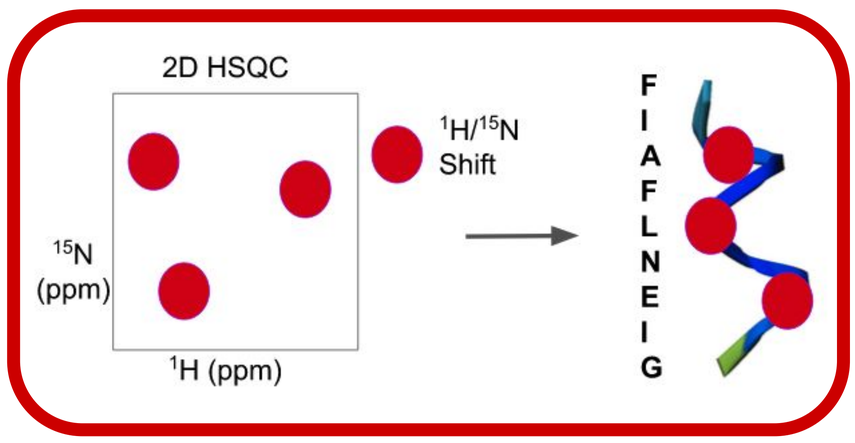
Adel **Schmucklermann***, Markus **Haak**, Tobias **Senoner,** Burkhard **Rost**, Janosch **Hennig**, Michael **Heinzinger***

Department for Bioinformatics, TUM School of Computation, Information and Technology, Technical University of Munich
Biophysical Chemistry, Chair of Biochemie IV, University Bayreuth

Contact: ge39row@tum.de, mheinzinger@rostlab.org, janosch.hennig@uni-bayreuth.de

SHORT SUMMARY

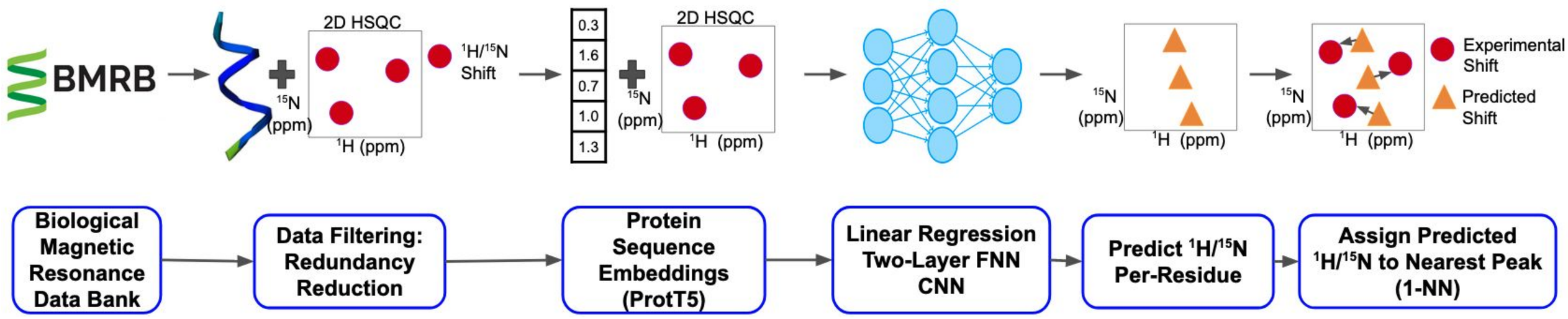
NMR ^{1}H-^{15}N chemical shift assignments with **Deep Learning** based only on protein sequence encodings from the **protein language model** ProtT5 [1]



Traditional Approach:

1) cost expensive, highly concentrated samples (^{13}C required)
2) time-consuming (3-6 d) spectrum acquisition and manual assignment of backbone resonances

New Deep Learning Approach:



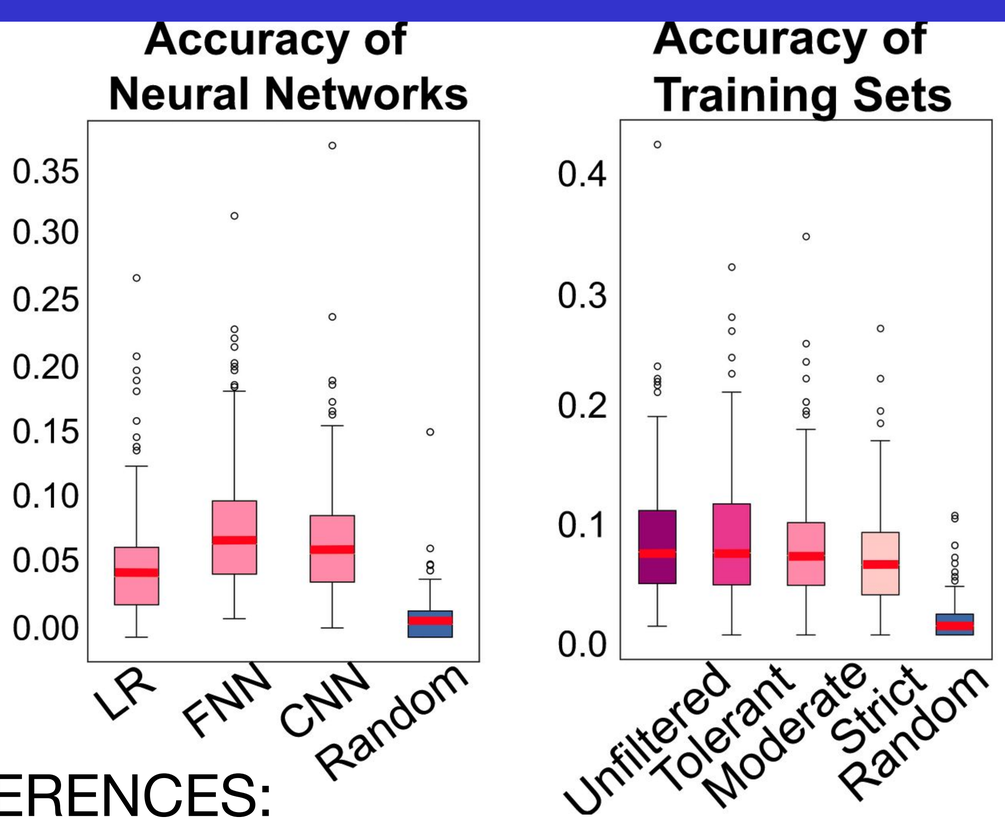
INTRODUCTION

- **Goal:** Assign ^{1}H/^{15}N shifts to the protein sequence without the need to measure additional ^{13}C backbone resonances
- Assignment is only based on the protein sequence encoded with a protein language model ProtT5 [1] and Deep Learning models
- **Benefits**: light-weight, fast alternative reducing the spectrum acquisition and assignment from 3-6 days to 20min-4h by avoiding the dependency on ^{13}C shifts

METHODS

- **Training:** protein sequence embeddings as input for Linear Regression/Two-Layer FNN/CNN
- **Inference:** Assigning predicted ^{1}H/^{15}N shifts to the closest experimentally measured shifts by 1-Nearest Neighbor
- **Data:** ^{1}H/^{15}N shifts from BMRB Database [2] redundancy reduced with mmseqs2 [3] with sequence identity 50% and coverage of 80%

PRELIMINARY RESULTS



OUTLOOK

- Simple linear and non-linear models seem not to be able to pick up the necessary information
- Although including information of protein 3D conformation in form of ProstT5 embeddings [4] did not show a significant improvement, incorporating **structure coordinates from PDB/AlphaFold2** [5,6] directly would be the next step
- **Cross-Attention** suggests to be a promising strategy to improve performance

ACKNOWLEDGMENTS:

Timothy Karl and Tobias Olenyi (TUM) for invaluable help with hardware issues. Experimentalists for providing the data for BMRB and the team maintaining the database.

REFERENCES:

[1]"ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning". A. Elnaggar, M. Heinzinger; B. Rost et al., IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022
[2]"Biological Magnetic Resonance Data Bank",.J C Hoch; M Yokochi et al., Nucleic Acids Research, 2023
[3]"MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets". Steinegger M. and Söding J., Nat Biotechnol, 2017
[4]"ProstT5: Bilingual Language Model for Protein Sequence and Structure". M. Heinzinger; K. Weissenow, M Steinegger, Burkhard Rost et al., bioRxiv, 2023
[5]"The Protein Data Bank". H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucleic Acids Research, 2000
[6]"Highly accurate protein structure prediction with AlphaFold". Jumper, J. et al., Nature, 2021

GitHub