# Curriculum Vitae - Dr. Milot Mirdita

Laboratory of Machine Learning and Bioinformatics
502-407, Seoul National University, 1 Gwanak-ro, Gwanak-gu, 08826 Seoul, Korea
milot@mirdita.de, ORCiD: 0000-0001-8637-6719, USA citizenship

## Research interests

Open-source methods for metagenomic analysis, fast and sensitive homology search, clustering, taxonomy, function and structure prediction.

## Education

| | |
|---|---|
| 07/2017-02/2022 | **Dr. rer. nat.** (summa cum laude). Advisor: Johannes Söding,<br>Max-Planck Institute for Multidisciplinary Sciences, Göttingen, Germany<br>University of Göttingen, Göttingen, Germany |
| 04/2014–08/2016 | **M.Sc.** in Computer Science, LMU, Munich, Germany |
| 10/2010–03/2014 | **B.Sc.** in Bioinformatics, LMU & TUM, Munich, Germany |
| Until 2010 | **Abitur**, Schönbuch Gymnasium, Holzgerlingen, Germany |

## Research experience

| | |
|---|---|
| Since 07/2022 | **Postdoctoral fellow** with Prof. Dr. Martin Steinegger<br>Seoul National University, Seoul, Korea |
| 03/2022-05/2022 | **Researcher** with Dr. Johannes Söding<br>Max-Planck Institute for Biophysical Chemistry, Göttingen, Germany |
| 07/2017-02/2022 | **PhD candidate** with Dr. Johannes Söding<br>Max-Planck Institute for Biophysical Chemistry, Göttingen, Germany |
| 04/2014-06/2017 | **Research assistant** with Dr. Johannes Söding (part time)<br>Max-Planck Institute for Biophysical Chemistry, Göttingen, Germany<br>Gene Center, Ludwig Maximilian University, Munich, Germany |
| 07/2011-05/2012 | **Research assistant** with Prof. Dr. Burkhard Rost (part time)<br>Technical University of Munich, Munich, Germany |

## Industry experience

| | |
|---|---|
| 06/2011-04/2016 | **Software engineer** (part time)<br>SpinSoft IT Solutions GmbH, München, Germany |

## Publications

My work has been cited over 17,300 times and my H-index is 24, according to Google Scholar.
A star indicates equal contributions.

Kallenborn F*, Chacon A*, Hundt C, Sirelkhatim H, Didi K, Dallago C$^\dagger$, **Mirdita M**$^\dagger$, Schmidt B$^\dagger$, Steinegger M$^\dagger$. GPU-accelerated homology search with MMseqs2. *Nature Methods*, 2025.

Akiyama Y, Zhang Z, **Mirdita M**, Steinegger M, Ovchinnikov S. Scaling down protein language modeling with MSA Pairformer. *bioRxiv*, 2025.08.02.668173, 2025.

Kim H, Kim R, **Mirdita M**, Steinegger M. Structural motif search across the protein-universe with Folddisco. *bioRxiv*, 2025.07.06.663357, 2025.

Yeo J, Han Y, Bordin N, Lau AM, Kandathil SM, Kim H, Levy Karin E, **Mirdita M**, Jones DT, Orengo C, Steinegger M. Metagenomic-scale analysis of the predicted protein structure universe. *bioRxiv*, 2025.04.23.650224, 2025.

Lee S, Kim J, **Mirdita M**, Steinegger M. Easy and interactive taxonomic profiling with Metabuli App. *bioRxiv*, 2025.03.10.642298, 2025.

Kim W, **Mirdita M**, Levy Karin E, Gilchrist CLM, Schweke H, Soeding J, Levy E, Steinegger M. Rapid and Sensitive Protein Complex Alignment with Foldseek-Multimer. *Nature Methods*, 469-472, 2025.

Fleming J, Magana P, Nair S, Tsenkov M, Bertoni D, Pidruchna I, Afonso MQL, Midlik A, Paramval U, Žídek A, Laydon A, Kovalevskiy O, Pan J, Cheng J, Avsec Ž, Bycroft C, Wong LH, Last M, **Mirdita M**, Steinegger M, Kohli P, Váradi M, Velankar S. AlphaFold Protein Structure Database and 3D-Beacons: New Data and Capabilities. *Journal of Molecular Biology*, 168967, 2025.

Zhang R, **Mirdita M**, Söding J. De novo discovery of conserved gene clusters in microbial genomes with Spacedust. *bioRxiv*, 2024.10.02.616292, 2024. *Accepted in Nature Methods*.

Kim R, Levy Karin E, **Mirdita M**, Chikhi R, Steinegger M. BFVD—a large repository of predicted viral protein structures. *Nucleic Acids Research*, D340–D347, 2024.

Zelenskaia M, Arangasamy Y, **Mirdita M**, Söding J, Raghavan V. TransAnnot—a fast transcriptome annotation pipeline. *Bioinformatics Advances*, vbae152, 2024.

Kim G*, Lee S*, Levy Karin E*, Kim H, Moriwaki Y, Ovchinnikov S$^\dagger$, Steinegger M$^\dagger$, **Mirdita M**$^\dagger$. Easy and accurate protein structure prediction using ColabFold. *Nature Protocols*, 620-642, 2024.

Cornman A, West-Roberts J, Camargo AP, Roux S, Beracochea M, **Mirdita M**, Ovchinnikov S, Hwang Y. The OMG dataset: An Open MetaGenomic corpus for mixed-modality genomic language modeling. *bioRxiv*, 2024.08.14.607850, 2024.

Gilchrist CLM, **Mirdita M**, Steinegger M. Multiple Protein Structure Alignment at Scale with FoldMason. *bioRxiv*, 2024.08.01.606130, 2024. *Under review in Science.*

Heinzinger M*, Weissenow K*, Gomez Sanchez J, Henkel A, **Mirdita M**, Steinegger M, Rost B. Bilingual Language Model for Protein Sequence and Structure. *bioRxiv*, 2023.07.23.550085, 2024.

Lee S*, Kim G*, Levy Karin E, **Mirdita M**, Park S, Chikhi R, Babaian A, Kryshtafovych A, Steinegger M. Petabase-Scale Homology Search for Structure Prediction. *Cold Spring Harbor Perspectives in Biology*, a041465, 2024.

Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, Tsenkov M, Nair S, **Mirdita M**, Yeo J, Kovalevskiy O, Tunyasuvunakool K, Laydon A, Žídek A, Tomlinson H, Hariharan D, Abrahamson J, Green T, Jumper J, Birney E, Steinegger M, Hassabis D, Velankar S. AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, D368-D375, 2024.

Basu S, Zhao B, Biró B, Faraggi E, Gsponer J, Hu G, Kloczkowski A, Malhis N, **Mirdita M**, Söding J, Steinegger M, Wang D, Wang K, Xu D, Zhang J, Kurgan L. DescribePROT in 2023: more, higher-quality and experimental annotations and improved data download options. *Nucleic Acids Research*, gkad985, 2023.

Barrio-Hernandez I*, Yeo J*, Jänes J, **Mirdita M**, Gilchrist CLM, Wein T, Varadi M, Velankar S, Beltrao P, Steinegger M. Clustering predicted structures at the scale of the known protein universe. *Nature*, 637–645, 2023.

Ruperti F*, Papadopoulos N*, Musser JM, **Mirdita M**, Steinegger M, Arendt D. Cross-phyla protein annotation by structural prediction and alignment. *Genome Biology*, 113, 2023.

van Kempen M*, Kim SS*, Tumescheit C, **Mirdita M**, Lee J, Gilchrist CLM, Söding J, Steinegger M. Fast and accurate protein structure search with Foldseek. *Nature Biotechnology*, 2023.

Kim H, **Mirdita M***, Steinegger M*. Foldcomp: a library and format for compressing and indexing large protein structure sets. *Bioinformatics*, 39, btad153, 2023.

Olenyi T*, Marquet C*, Heinzinger M, Kröger B, Nikolova T, Bernhofer M, Sändig P, Schütze K, Littmann M, **Mirdita M**, Steinegger M, Dallago C, Rost, B. LambdaPP: Fast and accessible protein-specific phenotype predictions. *Protein Science*, 32, e4524, 2023.

**Mirdita M***, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S*, Steinegger M*. ColabFold – Making protein folding accessible to all. Nature Methods, 19, 679–682, 2022.

**Mirdita M**, Steinegger M, Breitwieser F, Söding J, Levy Karin E. Fast and sensitive taxonomic assignment to metagenomic contigs. *Bioinformatics*, 37, 3029–3031, 2021.

Zhang R, **Mirdita M**, Levy Karin E, Norroy C, Galiez C, Söding J. SpacePHARER: Sensitive identification of phages from CRISPR spacers in prokaryotic hosts. *Bioinformatics*, 37, 3364–3366, 2021.

Zhao B, Katuwawala A, Oldfield CJ, Dunker AK, Faraggi E, Gsponer J, Kloczkowski A, Malhis N, **Mirdita M**, Obradovic Z, Söding J, Steinegger M, Zhou Y, Kurgan L. DescribePROT: database of amino acid-level protein structure and function predictions. *Nucleic Acids Research*, 49, D298-D308, 2021.

Bernhofer M, Dallago C, Karl T, Satagopam V, Heinzinger M, Littmann M, Olenyi T, Qiu J, Schütze K, Yachdav G, Ashkenazy H, Ben-Tal N, Bromberg Y, Goldberg T, Kajan L, O'Donoghue S, Sander C, Schafferhans A, Schlessinger A, Vriend G, **Mirdita M**, Gawron P, Gu W, Jarosz Y, Trefois C, Steinegger M, Schneider R, Rost B. PredictProtein - Predicting Protein Structure and Function for 29 Years. *Nucleic Acids Research,* 49, W535–W540, 2021.

Aevarsson A, Kaczorowska AK, [...], **Mirdita M**, et al. Going to extremes – a metagenomic journey into the dark matter of life. *FEMS Microbiology Letters*, 368, fnab067, 2021.

Gabler F, Nam SZ, Till S, **Mirdita M**, Steinegger M, Söding J, Lupas AN, Alva V. Protein sequence analysis using the MPI bioinformatics toolkit. *Current Protocols in Bioinformatics*, 72, e108, 2020.

Levy Karin E, **Mirdita M**, Söding J. MetaEuk – sensitive, high-throughput gene discovery and annotation for large-scale eukaryotic metagenomics. *Microbiome*, 8, 48, 2020.

Steinegger M, Meier M, **Mirdita M**, Vöhringer H, Haunsberger SJ, Söding J. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, 20, 473, 2019.

**Mirdita M**, Steinegger M, Söding J. MMseqs2 desktop and local web server app for fast, interactive sequence searches. *Bioinformatics*, 35, 2856–2858, 2019.

Steinegger M, **Mirdita M**, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. *Nature Methods*, 16, 603–606, 2019.

Keasar C, McGuffin LJ, Wallner B, Chopra G, Crivelli SN, [...], **Mirdita M**, et al. An analysis and evaluation of the WeFold collaborative for protein structure prediction and its pipelines in CASP11 and CASP12. *Scientific Reports*, 8, 1-18, 2018.

**Mirdita M***, von den Driesch L*, Galiez C, Martin M, Söding J, Steinegger M. Uniclust databases of clustered and deeply annotated protein sequences and alignments. *Nucleic Acids Research*, 45, D170-D176, 2017.

## Talks

Jun 2025 Protein Analysis in the AlphaFold Era.
Max-Planck Institute for Multidisciplinary Sciences, Göttingen, Germany.

Jun 2025 Protein Analysis in the AlphaFold Era. 17th Seoul International New Drug Forum, Korea.

Apr 2025 GPU-accelerated homology search with MMseqs2. RECOMB-SEQ, Seoul, Korea.

Apr 2025 Protein Analysis in the AlphaFold Era. ITACA.SB Workshop, Bari, Italy.

Jan 2024 Protein Analysis in the AlphaFold Era. ISCB-SC RSG-India, Virtual.

Dec 2024 Challenges to metagenomic annotation with Foldseek and protein language models, Invited speaker, MLSB @ NeurIPS 2024, Vancouver, Canada.

Oct 2024 Protein Analysis in the AlphaFold Era. CSHL Microbiome, Cold Spring Harbor, USA.

Jun 2024 Artificial intelligence and machine learning based software development for protein structure prediction. British Neuropathological Society Summer School, Cirencester, UK.

Dec 2023 Petasearch: Efficient and Sensitive Sequence Comparison at Scale.
DTMBIO 2023, Okinawa, Japan.

Aug 2023 Compressing huge protein structure databases with Foldcomp.
ICBP 2023, Seoul, Korea.

July 2023 Petasearch: Efficient and Sensitive Sequence Comparison at Scale.
ISMB/ECCB 2023, Lyon, France.

Nov 2021 Fast and sensitive taxonomic assignment to metagenomic contigs.
SNU Online Symposium on Bioinformatics for metagenomic analysis, Virtual.

July 2021 Fast and sensitive taxonomic assignment to metagenomic contigs.
ISMB/ECCB 2021, Virtual.

## Selected posters

Kallenborn F, Chacon A, Hundt C, Sirelkhatim H, Didi K, Dallago C, **Mirdita M**, Schmidt B, Steinegger M.
GPU-accelerated homology search with MMseqs2. RECOMB & RECOMB-SEQ 2025, April 24-29, 2024, Seoul, Korea.
*+ co-authored 8 other posters at RECOMB & RECOMB-SEQ 2024.*

**Mirdita M**, Mihaila V, Bouras G, Heinzinger M, Steinegger M. (Poster award)
Metagenome-scale structural homology detection with Foldseek-ProstT5. APBJC 2024, October 22-25, 2024, Okinawa, Japan.
*+ co-authored 3 other posters at APBJC 2024.*

**Mirdita M**, Li M, Hügel J, Söding J, Steinegger M.
Petasearch: Efficient and Sensitive Sequence Comparison at Scale. BIOINFO 2023, November 13-15, 2023, Yeosu, Korea.
*+ co-authored 1 other poster at BIOINFO 2023.*

**Mirdita M**, Li M, Hügel J, Söding J, Steinegger M.
Petasearch: Efficient and Sensitive Sequence Comparison at Scale. ISMB/ECCB 2023, July 23-27, 2023, Lyon, France.
*+ co-authored 5 other posters at ISMB/ECCB 2023.*

Hyunjoo J, **Mirdita M**, Sommer HG, Galiez C, Söding J, Steinegger M.

MMseqs2 profile/profile: fast and ultra sensitive searches beyond the twilight zone. ISMB/ECCB 2021, July 26-30, 2021, Virtual.

*+ co-authored 2 other posters at ISMB/ECCB 2021.*

Steinegger M, **Mirdita M**, Söding J.

New algorithms and tools for large-scale sequence analysis of metagenomics data. Genome Informatics, November 6-9, 2019, Cold Spring Harbor, USA.

Levy Karin E, **Mirdita M**, Söding J.

MetaEuk – sensitive, high-throughput gene discovery and annotation for large-scale eukaryotic metagenomics. Microbiome COSI. ISMB/ECCB, July 21-25, 2019, Basel, Switzerland.

**Mirdita M**, Steinegger M, Söding J. (Poster award)

MMseqs2 desktop and local web server app for fast, interactive sequence searches. September 7-12, 2018, ECCB, Athens, Greece.

Steinegger M, **Mirdita M**, von den Driesch L, Söding J. (Poster award)

Sensitive protein sequence searching for the analysis of massive data sets. December 10-13, 2016, CASP12, Gaeta, Italy.

## Teaching & workshops

Spring 2025 Introduction to Bioinformatics, taught w. Prof. Steinegger, undergraduate, SNU, Korea.

Apr 2024 ColabFold and Foldseek, KSBB Workshop, Changwon, Korea.

Fall 2023 Advanced Bioinformatics, co-taught w. Prof. Steinegger, graduate course, SNU, Korea.

2022-2023 (Co-)supervised multiple interns, B.Sc., M.Sc. and Ph.D. students at SNU, Korea.

2017-2022 (Co-)supervised multiple interns, B.Sc. and M.Sc. students at MPI-BPC, Germany.

Sept. 2020 Workshop Deep dive into metagenomics using MMseqs2. ECCB 2020, Virtual.

Sept. 2018 Workshop Modern and scalable tools for the efficient analysis of very large metagenomic datasets. ECCB 2018, Athens, Greece.

2017-2020 Yearly 1.5-day tutorial on protein structure prediction, homology search and metagenomics analysis in the International M.Sc. Curriculum for Molecular Biology, University of Göttingen, Germany.

## Funding

2023-2026 National Research Foundation of Korea (RS-2023-00250470)
Unraveling the Global Virome through Petabase-scale Sequence and Structural Analysis.
KRW 210.000.000 (~150.000€).

## Awards, fellowships and achievements

| | |
|---|---|
| 2024 | Best poster award, APBJC 2024, Okinawa, Japan |
| 2018 | Best poster award, ECCB 2018, Athens, Greece |
| 2016 | Best poster award, CASP12 2016, Gaeta, Italy |
| 2016 | Stefan-Hell scholarship, Max-Planck-Gesellschaft |
| 2012 | Hardware donations for HPC: NVIDIA and AMD, ~$4,000 in GPUs |
| 2011 | Amazon AWS in Education, cloud research grant, $10,000 AWS credits |

## Open source software

| | |
|---|---|
| ColabFold | github.com/sokrypton/ColabFold |
| Foldseek | github.com/steineggerlab/foldseek |
| MMseqs2 | github.com/soedinglab/mmseqs2 |
| MMseqs2-App | github.com/soedinglab/mmseqs2-app |
| Foldcomp | github.com/steineggerlab/foldcomp |
| Metaeuk | github.com/soedinglab/metaeuk |
| SpacePHARER | github.com/soedinglab/spacepharer |
| Plass | github.com/soedinglab/plass |
| HH-suite | github.com/soedinglab/hh-suite |

## Languages

| | |
|---|---|
| Native | English, German |
| Fluent | Albanian |
| Basic | Korean, Vietnamese |

## Peer review

I have peer reviewed publications in Bioinformatics, GigaScience, Scientific Reports, NAR Genomics and Bioinformatics, Nature Methods, Nature Communications and Protein Science.