Simple Models Work Best in Peptide Function Prediction

Piotr Ludynia • ML & Chemoinformatics Lab (MLCIL), Faculty of Computer Science, AGH

J. Adamczyk, P. Ludynia, W. Czech *"Molecular Fingerprints Are Strong Models for Peptide Function Prediction"*, ArXiv preprint

Peptide Function Prediction

Peptides:

- small proteins, up to 50 amino acids
- potential novel therapeutics, e.g. antibiotics or anticancer drugs
- SMILES, amino-acid sequences, graphs

Peptide function prediction:

- binary classification
- does a peptide have a certain property, e.g. antibacterial?

Popular Approaches

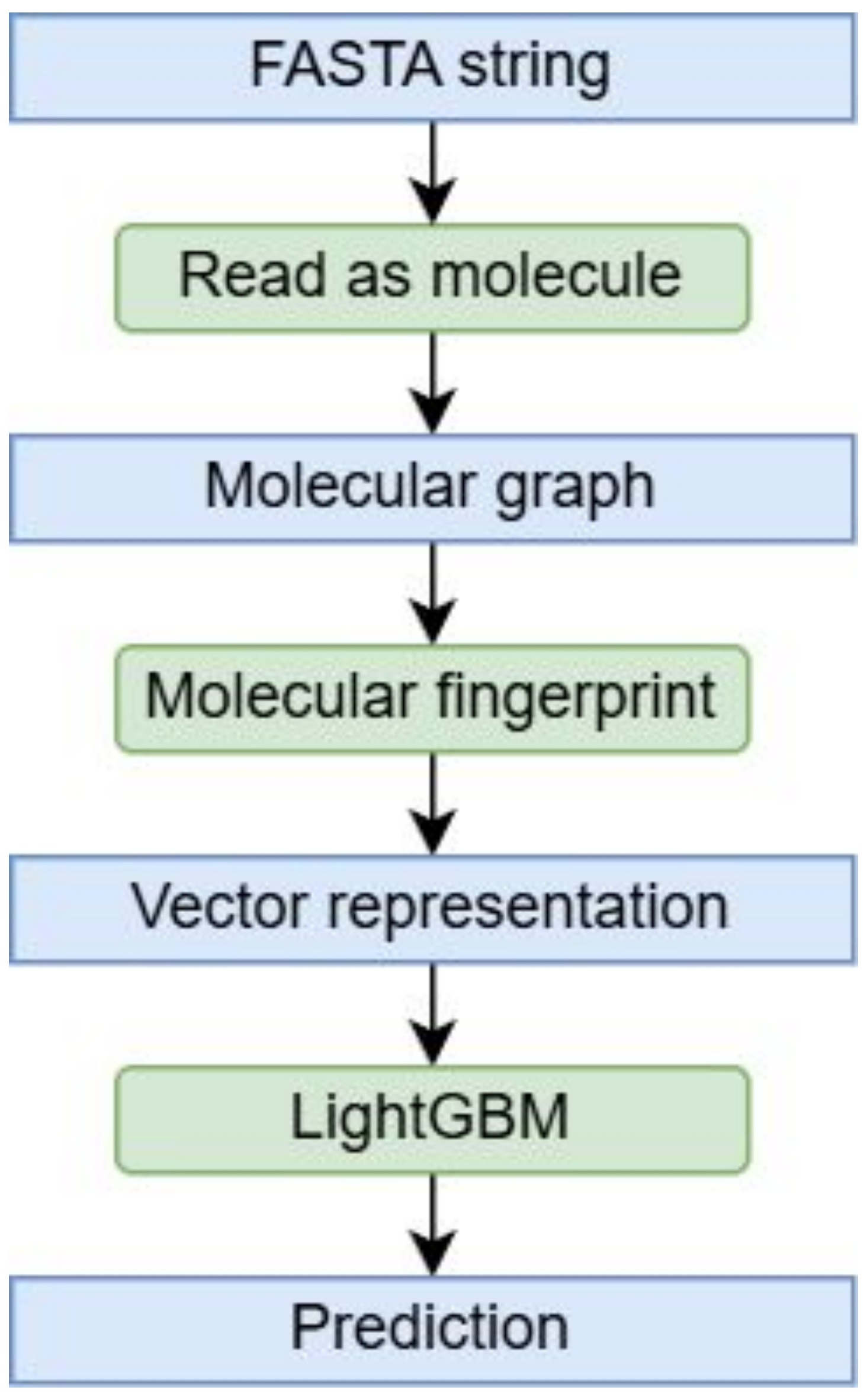
- protein language models (**PLMs**)
- graph neural networks (**GNNs**)
- complex multimodal feature engineering

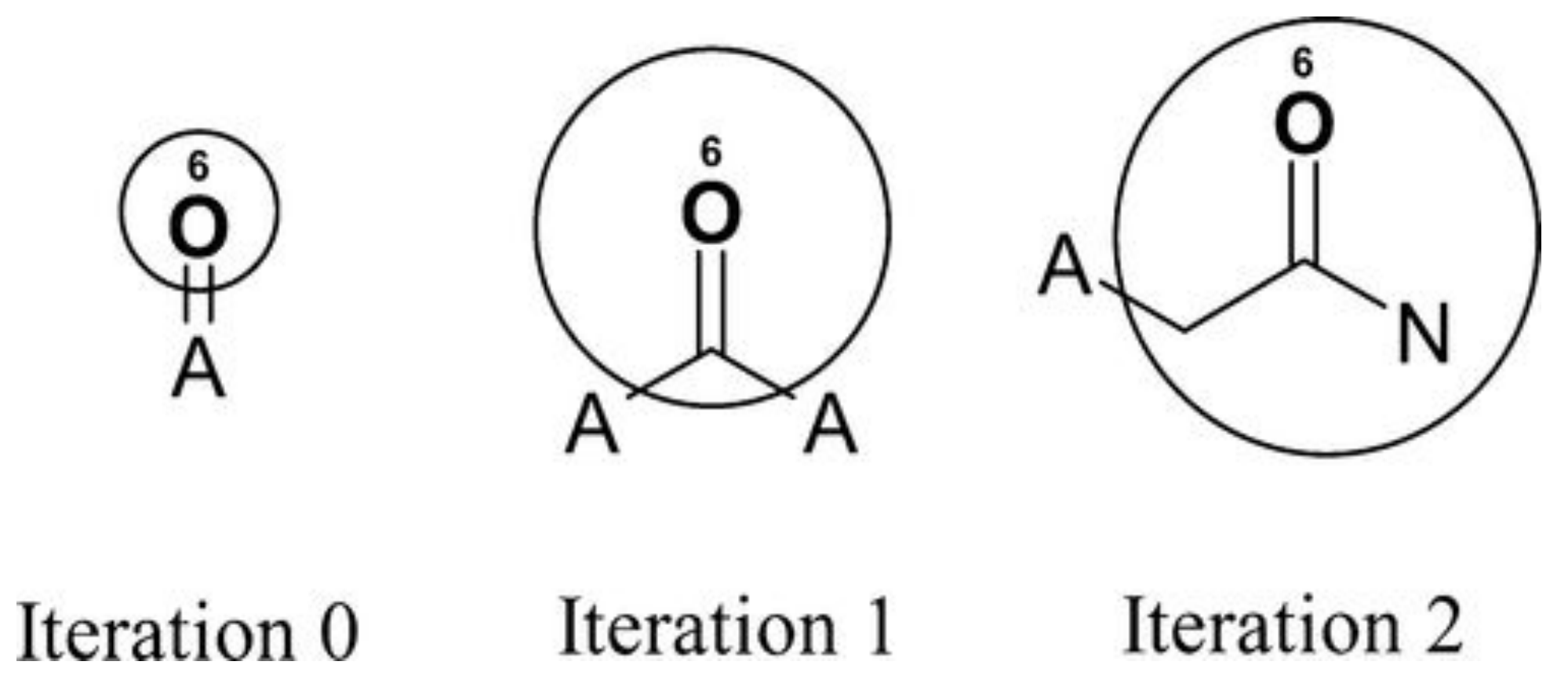
Long Range Graph Benchmark (LRGB)

- peptide benchmark for **graph learning**
- designed to require learning **long-range dependencies**
- benchmark mostly for graph neural networks (GNNs)
- **important assumption** - "proteins require long-range dependencies so peptides must require them too" - **but is it correct?**

Methods

- **molecular fingerprints**: ECFP, RDKit fp, Topological Torsion (TT)
- **vectorization algorithms** for molecules
- **hashed fingerprints** detect subgraphs of given shape and hash results onto feature vector
- **count fingerprints** count occurrences of subgraphs (binary would just detect if it exists)
- implemented using our own library for efficient computation of molecular fingerprints, **scikit-fingerprints**
- **LightGBM** classifier
- quickly trainable model with minimal number of parameters


FASTA string
Read as molecule
Molecular graph
Molecular fingerprint
Vector representation
LightGBM
Prediction



Example subgraph extraction in consecutive iterations of ECFP algorithm.

Reprinted with permission from [7]. Copyright 2010 American Chemical Society.

Large Scale Benchnmarking

- **6** benchmarks, **126** datasets, **>215 thousand** unique peptides
- SOTA on 5 benchmarks:
  - LRGB
  - BERT-based models benchmark
  - XUAMP
  - AMPBenchmark
  - PeptideReactor
- AutoPeptideML - near SOTA but much faster and lighter

LRGB results:

Model	Peptides-func AUPRC ↑	Peptides-struct MAE ↓
Transformer	63.26 ± 1.26	0.2529 ± 0.0016
SAN	64.39 ± 0.75	0.2545 ± 0.0012
MOLTOP	64.59 ± 0.05	-
GraphGPS	65.35 ± 0.41	0.2500 ± 0.0005
GINE	66.21 ± 0.67	0.2473 ± 0.0017
GatedGCN	67.65 ± 0.47	0.2477 ± 0.0009
GCN	68.60 ± 0.50	0.2460 ± 0.0007
GraphViT	69.42 ± 0.75	0.2449 ± 0.0016
GRIT	69.88 ± 0.82	0.2460 ± 0.0012
CRaWl	70.74 ± 0.32	0.2506 ± 0.0022
GRED	71.33 ± 0.11	0.2455 ± 0.0013
DRew	71.50 ± 0.44	0.2536 ± 0.0015
HDSE	71.56 ± 0.58	0.2457 ± 0.0013
S^2GCN	73.11 ± 0.66	0.2447 ± 0.0032
RDKit	73.11	0.2459
TT	73.18	0.2438
ECFP	**74.60**	**0.2432**

- short-range model outperform the complex ones
  - conclusion: **peptides don't require long range dependencies**
  - short-range features contain enough information for accurate peptide function prediction
- why so many papers haven't used this simple method?
  - their fingerprint baselines used **binary** fingerprints instead of **count** variants
  - many scientists don't compare their models against **strong baselines!**

Conclusions and Lessons

- conclusions
  - molecular fingerprints are effective methods for peptide function prediction problem
  - simple vectorization + tree ensemble is able to **outperform deep-learning models**
  - assumption that peptides need long-range dependencies was false
- lessons
  - complex models might not always be the best
  - don't underestimate simple feature extraction methods
  - always compare your models against simple but strong baselines
  - design your baselines very carefully, read documentation and learn about the methods

References

1. Adamczyk, Jakub, Piotr Ludynia, and Wojciech Czech. "Molecular Fingerprints Are Strong Models for Peptide Function Prediction." arXiv preprint arXiv:2501.17901 (2025).
2. Goles, Montserrat, et al. "Peptide-based drug discovery through artificial intelligence: towards an autonomous design of therapeutic peptides." Briefings in Bioinformatics 25.4 (2024): bbae275.
3. Dwivedi, Vijay Prakash, et al. "Long range graph benchmark." Advances in Neural Information Processing Systems 35 (2022): 22326-22340.
4. Rogers, David, and Mathew Hahn. "Extended-connectivity fingerprints." Journal of Chemical Information and Modeling 50.5 (2010): 742-754.
5. Adamczyk, Jakub, and Piotr Ludynia. "Scikit-fingerprints: Easy and efficient computation of molecular fingerprints in Python." SoftwareX 28 (2024): 101944.
6. Gao, Wanling, et al. "Comprehensive Assessment of BERT-Based Methods for Predicting Antimicrobial Peptides." Journal of Chemical Information and Modeling 64.19 (2024): 7772-7785.
7. Fernández-Díaz, Raúl, et al. "AutoPeptideML: a study on how to build more trustworthy peptide bioactivity predictors." Bioinformatics 40.9 (2024): btae555.