



Efektywne przewidywanie własności peptydów w projektowaniu leków

Jakub Adamczyk, Piotr Ludynia

Opiekun: dr inż. Wojciech Czech

Akademia Górniczo-Hutnicza im. Stanisława Staszica w Krakowie

- **peptydy:**

- małe białka, do 50 aminokwasów
- wiążą się z dużymi białkami w naszym ciele, regulując ich działanie
- stosowane w farmacji jako leki (*peptide therapeutics*)

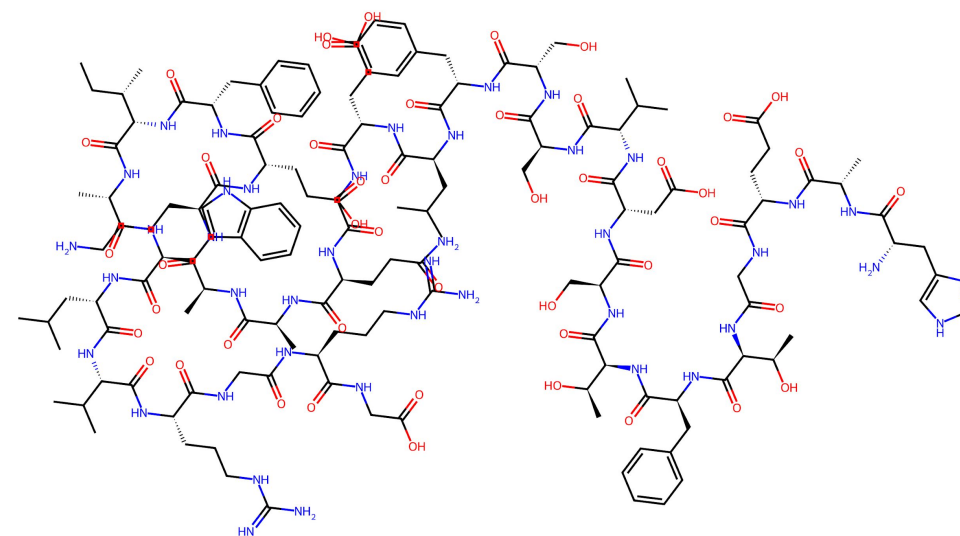
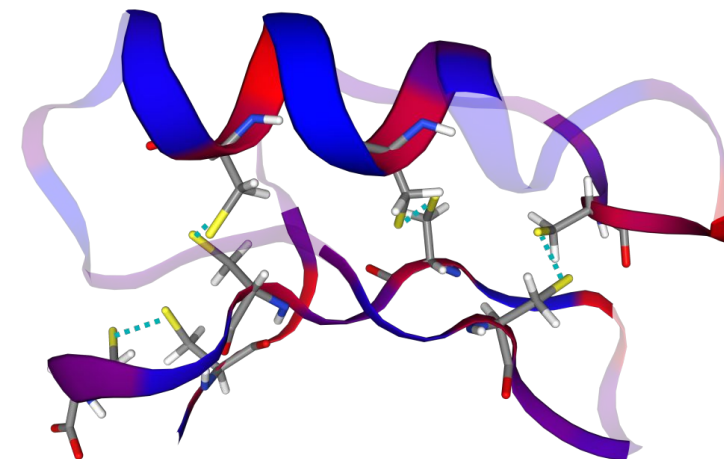
- **zastosowania jako nowoczesne leki peptydowe, np.:**

- antybiotyki (przeciwko bakteriom lekoopornym)
- leki antynowotworowe
- leki na otyłość i cukrzycę

- **reprezentowane jako:**

- sekwencje aminokwasów (tekst)
- zwinięte białka w 3D (konformacje przestrzenne)
- "płaskie" cząsteczki chemiczne (molekuła, atomy i wiązania)

HAEGTFTSDVSSYLEGQAAKEEFIAW
LVRGRG



- **przewidywanie własności peptydów:**

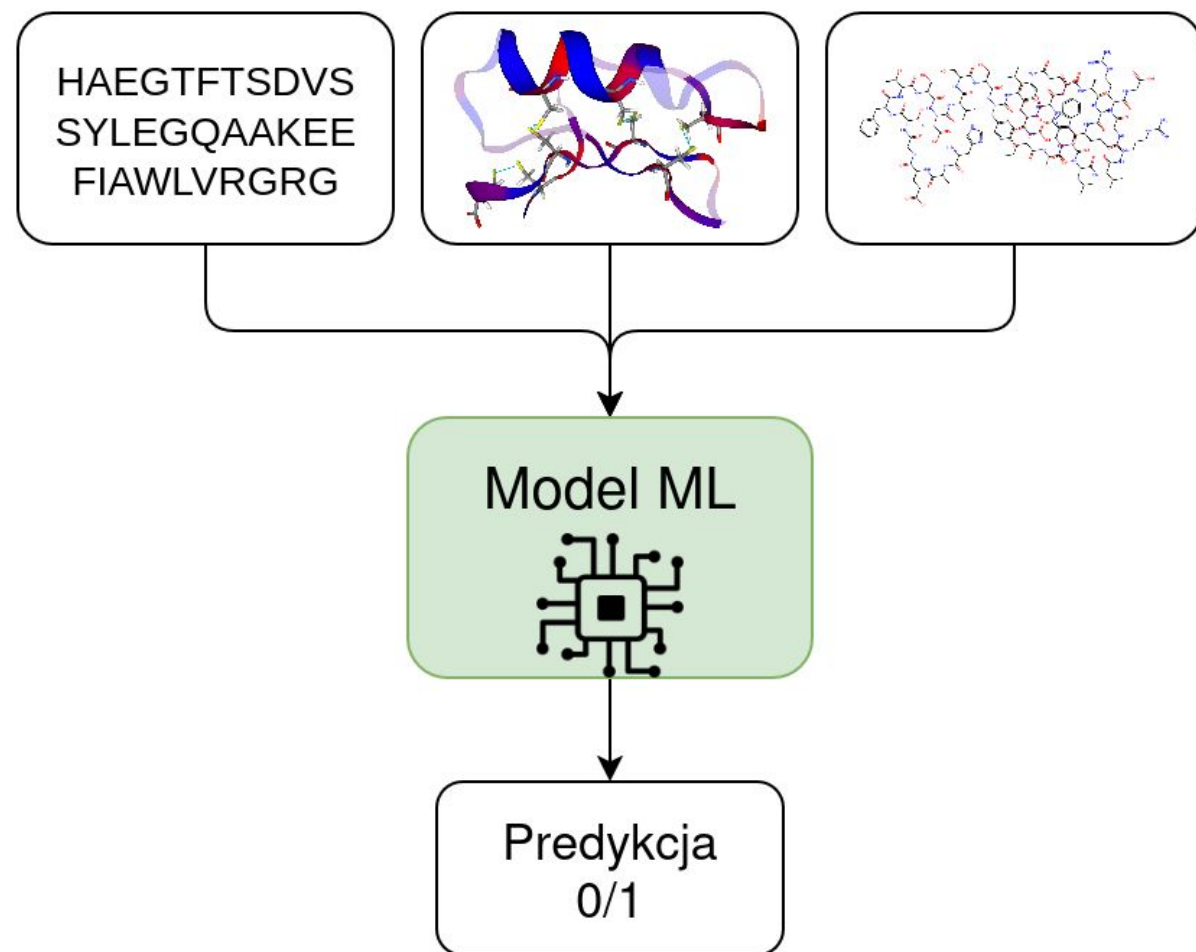
- czy peptyd ma pożądaną własność?
- np. antybakteryjny, antynowotworowy
- kluczowe pytanie w projektowaniu leków

- **chemoinformatyka + uczenie maszynowe (ML):**

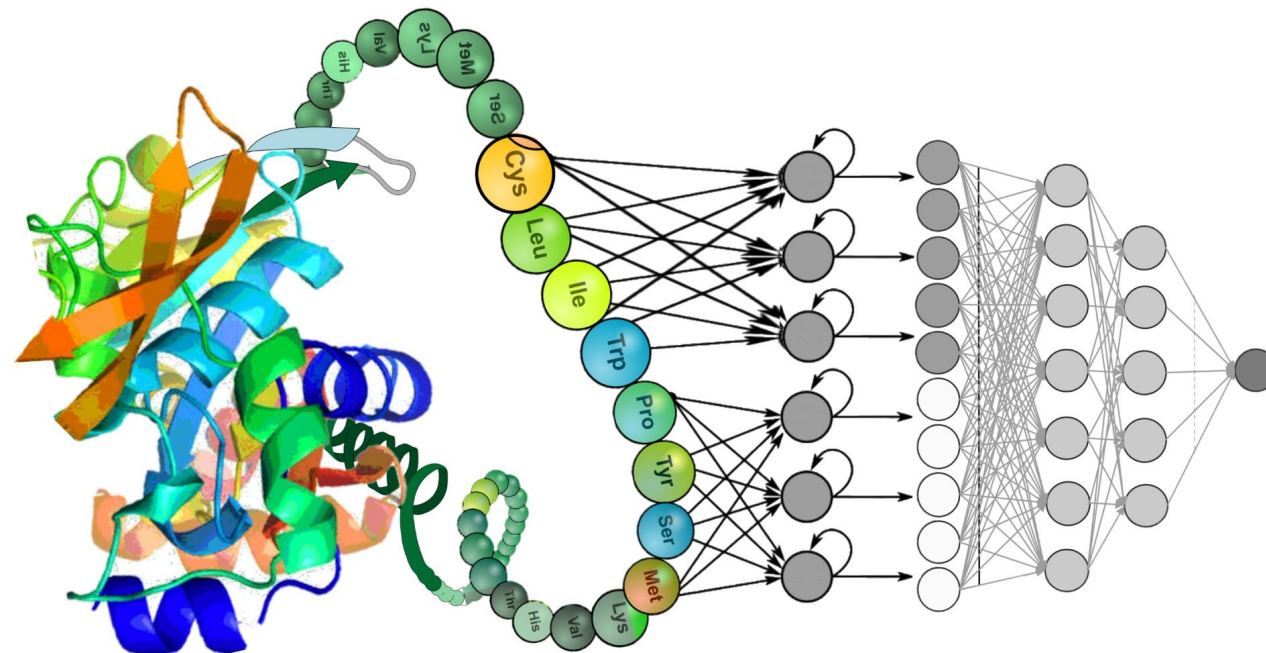
- badania w laboratorium są długie, trudne i drogie
- budujemy za to model ML do predykcji
- klasyfikacja binarna, 0/1

- **ekstrakcja cech:**

- musimy zakodować peptyd jako wektor liczb
- można wykorzystać: sekwencję aminokwasów, strukturę 3D, graf molekularny
- lub też wiele z tych (multimodalność)



- **Protein Language Models (PLMs)**
 - transformery tekstowe dla białek
 - duże modele językowe (LLMs)
 - używają sekwencji aminokwasów
- **grafowe sieci neuronowe (GNNs):**
 - operują na grafie atomów i wiązań
 - uczone od zera do konkretnych zadań
- **multimodalna inżynieria cech:**
 - ekstrakcja cech z pomocą wiedzy biologów
 - wykorzystują wiele różnych modalności
 - tradycyjne, ale skomplikowane podejście



- **fingerprinty molekularne:**

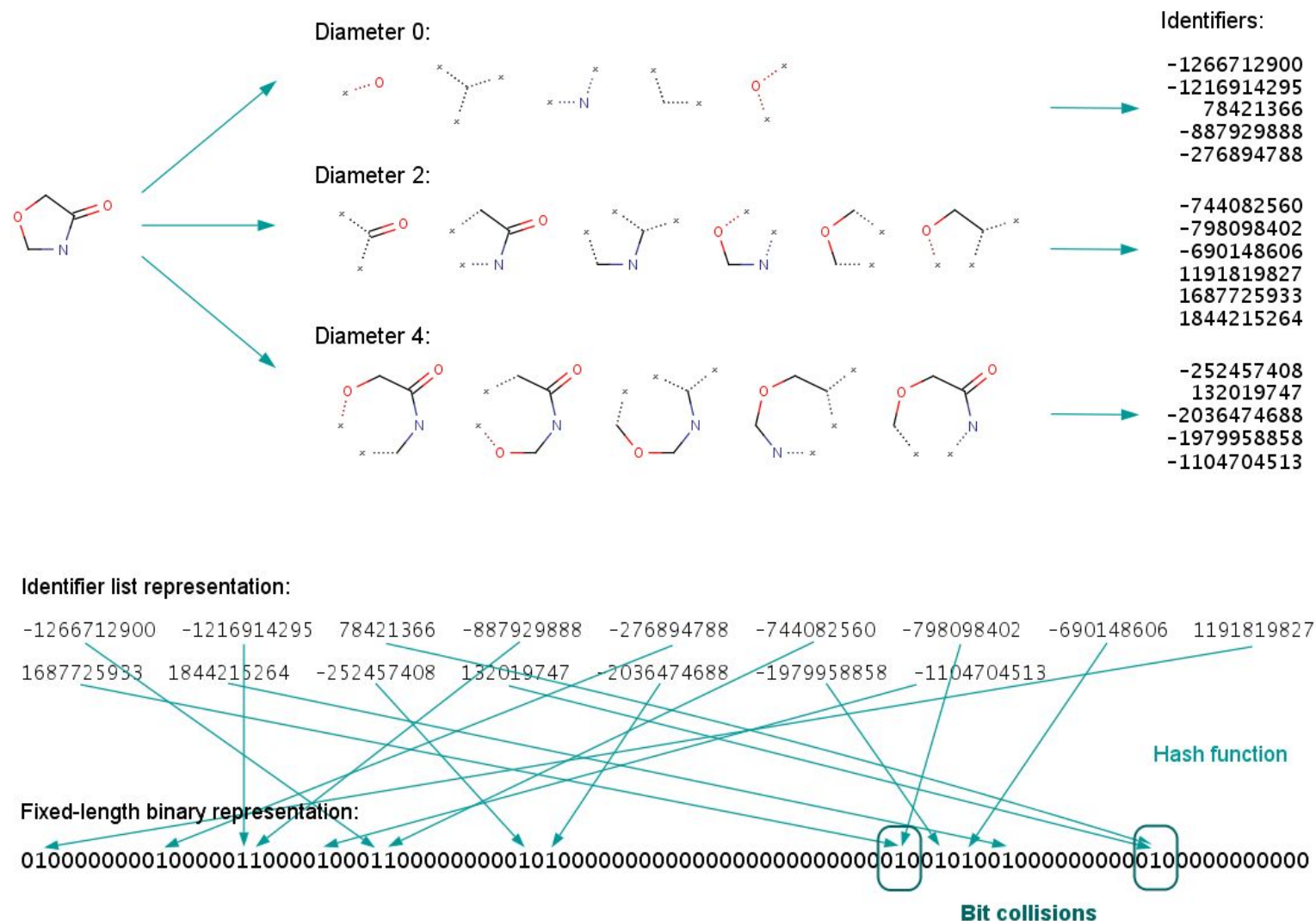
- metoda ekstrakcji cech
- typowa dla małych cząsteczek
- tradycyjna w chemoinformatyce

- **działanie:**

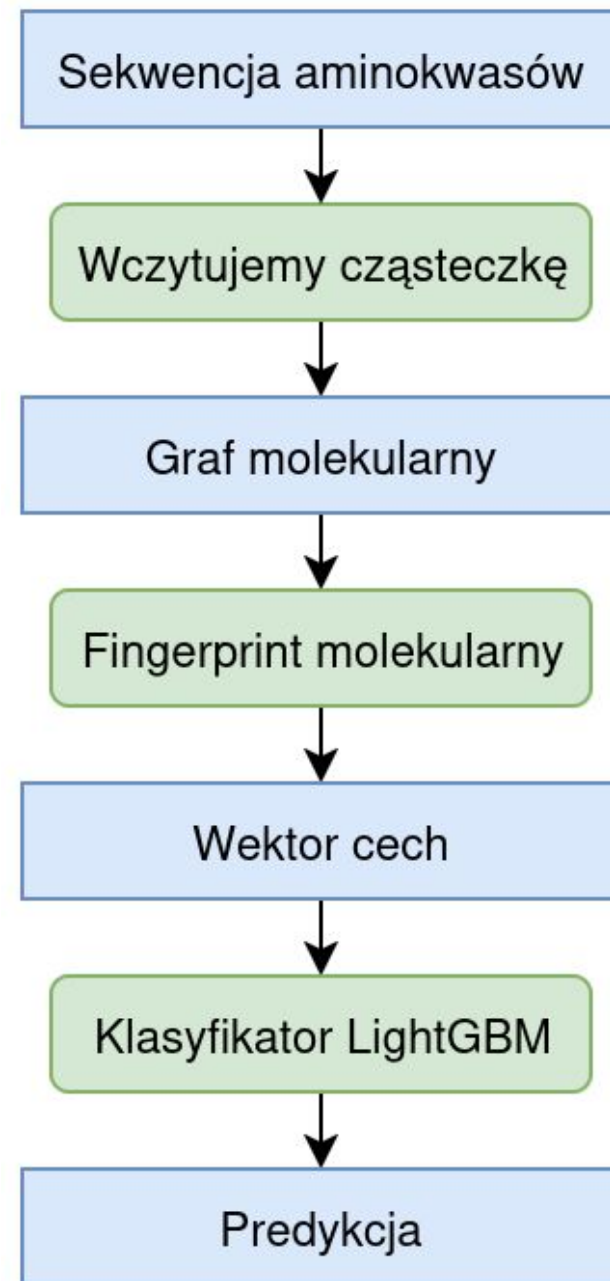
- graf molekularny
- listujemy wszystkie małe subgrafy
- każdy dostaje identyfikator
- haszujemy je na wektor cech

- **rodzaje fingerprintów:**

- różne kształty subgrafów, np. ECFP
- binarne vs zliczające (count), biorą pod uwagę rozmiar



- **obserwacja:** peptyd = pomiędzy małą cząsteczką a białkiem
- **idea:**
 - zignorujemy biologię, aminokwasy, białka etc.
 - traktujemy peptyd jak czystą chemię, dużą molekułę
 - przewidywanie własności molekuł - typowe w chemoinformatyce
- **proponowane podejście:**
 - tworzymy graf molekularny z peptydu, na poziomie atomów i wiązań
 - fingerprint molekularny do ekstrakcji cech
 - klasyfikator LightGBM
 - zaimplementowane z pomocą biblioteki *scikit-fingerprints*
- **zalety:**
 - prostota, szybkość
 - doskonałe wyniki, state-of-the-art



- **state-of-the-art (SOTA)**
- 6 zróżnicowanych benchmarków, 126 zbiorów danych, >215k peptydów
- 3 fingerprinty: ECFP, RDKit, Topological Torsion (TT)
- wygrywamy ze wszystkimi innymi podejściami
- **największe** takie porównanie w literaturze

Model	Average F1
AMP-BERT	75.3
BERT-Protein	75.1
cAMPs_pred	64.1
LM_pred	63.8
LM_pred (BFD)	70.3
RDKit	87.2
TT	87.8
ECFP	87.6

Benchmark BERT-like (NLP)

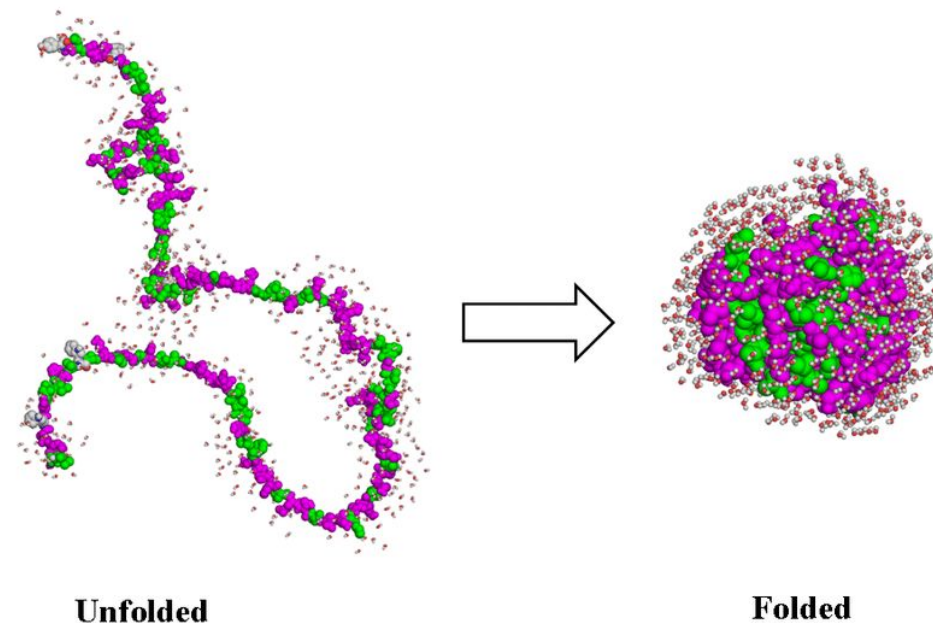
Method	AUROC	MCC
AMPscannerV2	58.5	0.137
iAMP-2L	59.2	0.261
ADAM-SVM	61.2	0.264
ampir	61.9	0.156
MLAMP	62.9	0.194
ADAM-HMM	68.4	0.39
AMPlify	69.7	0.381
AMPEP	72.7	0.425
AMPfun	73.5	0.414
RDKit	74.5	0.404
TT	73.1	0.385
ECFP	75.0	0.429

Benchmark XUAMP (ekstrakcja cech)

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 ² GCN	73.11 ± 0.66	0.2447 ± 0.0032
RDKit	73.11	0.2459
TT	73.18	0.2438
ECFP	74.60	0.2432

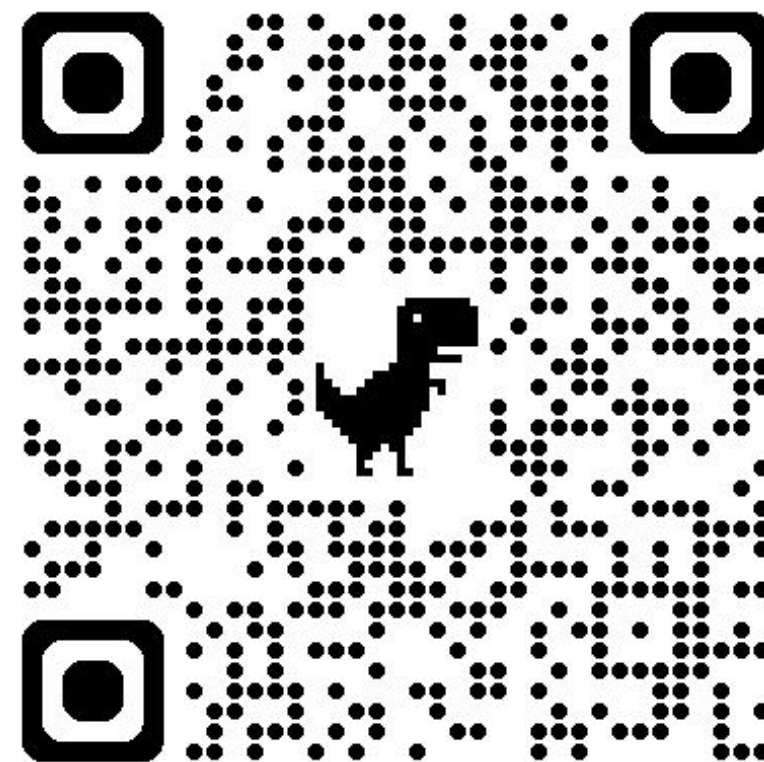
Benchmark LRGB (GNNy)

- **odkrycie biologiczne:** wyniki wskazują, że w peptydach nie zachodzą dalekie zależności!
- **zależności w białkach:**
 - aminokwasy odległe w sekwencji mogą leżeć blisko w strukturze 3D - **zwijanie (folding)**
 - kluczowe w przewidywaniu funkcji białek
 - peptydy to małe białka - różnice fizyczne i chemiczne
- **fingerprinty molekularne:**
 - ignorują kolejność aminokwasów
 - tylko zliczamy subgrafy, nieważne gdzie wystąpiły
 - patrzą na bardzo krótkie zależności - małe subgrafy
- SOTA przy tych ograniczeniach = w peptydach dominują **krótkie zależności**



Dataset	Variant	ECFP	TT	RDKit
Peptides-func	binary	70.57	66.18	63.88
Peptides-func	count	74.60	73.18	73.11
AUPRC gain ↑		+4.03	+7.00	+9.23
Peptides-struct	binary	0.3049	0.3298	0.3331
Peptides-struct	count	0.2432	0.2438	0.2459
MAE gain ↓		-0.0617	-0.0860	-0.0872

- **"Molecular Fingerprints Are Strong Models for Peptide Function Prediction"** J. Adamczyk, P. Ludynia, W. Czech
- <https://arxiv.org/abs/2501.17901>
- zachęcamy do przeczytania:
 - rozszerzone wyniki
 - pomiary czasu
 - dodatkowe analizy
- obecnie w recenzji w Bioinformatics
- publicznie dostępny kod open source na licencji MIT



- **problem:**
 - przewidywanie własności peptydów
- **proponowane podejście:**
 - graf molekularny + fingerprinty molekularne
- **zalety:**
 - prostota, szybkość, doskonałe wyniki
- **wyniki:**
 - SOTA na 6 benchmarkach i 126 datasetach
 - brak dalekich zależności - istotne biologicznie
 - problem opiera się o zliczanie substruktur
- **artykuł:**
 - *"Molecular Fingerprints Are Strong Models for Peptide Function Prediction"* J. Adamczyk, P. Ludynia, W. Czech

