NeurIPS 2024 - Predict New Medicines with BELKA

Predict small molecule-protein interactions using the Big Encoded Library for Chemical Assessment

Background

🏢 20+ years of corporate finance experience. BS in economics, MS in management.

🏢 Ex-CFO of a Russian version of Fannie Mae ($20bn assets).

🎓 MSc candidate in Data Science and AI at Northwestern (2025).

💊 Primary focus: medical imaging, chemistry, pharmaceuticals, and [just on the edge of the radar] finance.

Predict small molecule-protein interactions using the Big Encoded Library for Chemical Assessment: **1st out of 1950 teams**

Hacking the Human Vasculature in 3D. Segment vasculature in 3D scans of human kidney: **18th out of 1149 teams**

Jane Street Real-Time Market Data Forecasting. Predict financial market responders using real-world data: **7th out of 2280 teams****

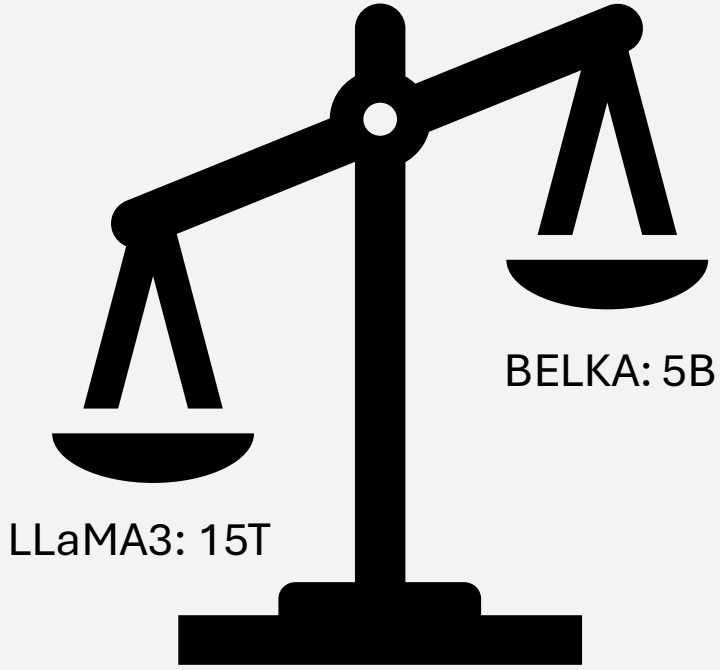
Challenge: a needle in a haystack

Drug-like-space: **10^60** molecules

BELKA: **133M** small molecule drug candidates, 3 target proteins

FDA-registered drugs: **4471**

NLP vs. Chemistry: 3000-0 to Poets…



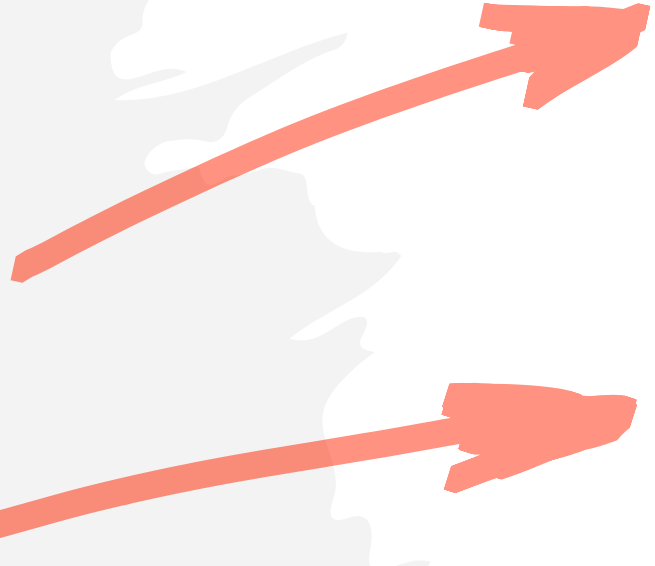
	NLP	Chem
Vocabulary	170K words, 50K+ tokens	118 periodic elements
Rules / structure	Vague and evolving	Defined and universal
Targets	Vague and context-driven	Clear
Uncertainty	High: polysemy, ambiguity, cultural context	Moderate: fundametally deterministic

Architecture [finally]

Chem-aware tokenization.
43 tokens: Br, C, [C++], etc.
Also tried but failed: n-grams, Wordpiece, Bert.

Embedding / latent space
dimension: 32

ChemBERTa-2[2]: 591 tokens, embedding size 768.



Tokenizer: atomInSmiles[1]

Embeddings / Positional Encodings

4 x Encoder (Self-Attention -> FFN -> Add)

Neck: GRU

Head: Dense(sigmoid)

1) https://pypi.org/project/atomInSmiles/
2) https://arxiv.org/abs/2209.01712

Training schedule

MLM: classics from “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”.
Mini-batch adaptive Categorical Focal Cross Entropy.

ECFP and Affinity: GRU feature extractor on top of Self-Attention Encoder.
Binary Focal Cross Entropy

	MLM	ECFP	Affinity
Train data	✓	✓	✓
Public test data	✓	✓	
Extra data[1]	✓	✓	
Epochs (100M each)	10+	5+	10+
Validation	None	9m non-shared	

1) "Building Block-Based Binding Predictions for DNA-Encoded Libraries", sited by @hengck23 and processed by @chemdatafarmer

Brilliant ideas [that failed]

Pretrain to predict outcomes of chemical reactions

GAN-based augmentations

SMILES-to 3D fingerprints pre-training

Read ”Chemistry for Dummies”

- 💊 ZINC pre-training
- 💊 MTR pre-training
- 💊 Fixed-position fingerprints (MACCS, PubChem)
- 💊 ECFP- or multi-input models
- 💊 High-dimensional models
- 💊 Gated Fusion / Cross-attention with blocks

Contacts:

vshlepov@gmail.com

https://www.linkedin.com/in/victor-shlepov/