

NOVÉ TECHNIKY UČENIA SA BEZ UČITEĽA PRE KLASIFIKÁCIU A PREDIKCIU MOLEKULÁRNYCH VLASTNOSTÍ

Bc. Samuel Baran

Vedúci práce:

RNDr. Ľubomír Antoni, PhD.

Konzultant:

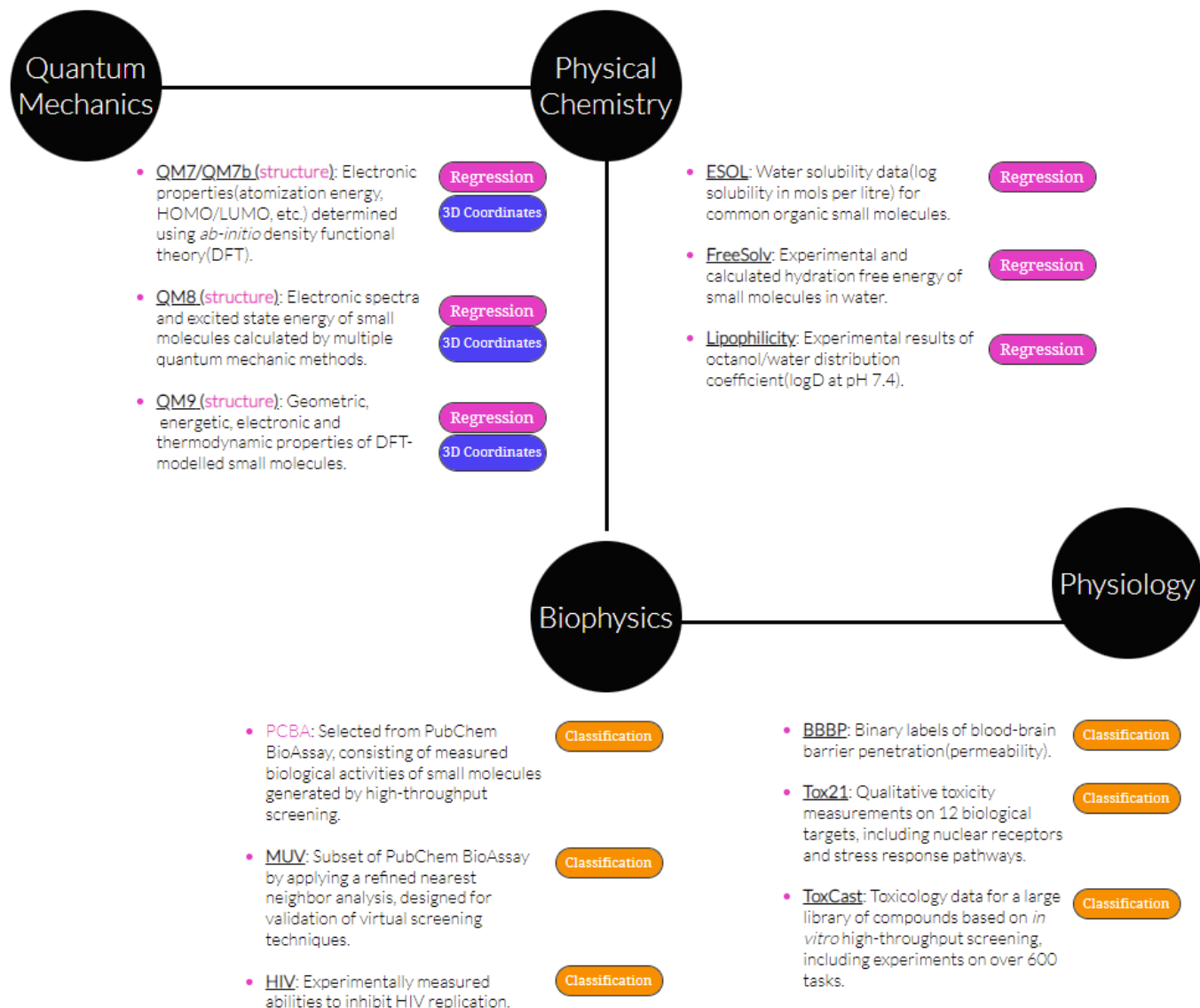
RNDr. René Derian, PhD. (Tachyum s.r.o.)

MOLEKULÁRNE VLASTNOSTI

MoleculeNet (2018)

- dataseť
- metódy práce s dátami
- modely
- metriky

<https://moleculenet.org/>
<https://arxiv.org/abs/1703.00564>



QM8

	A	B	C	D	E
1	Column1	Column2	Column3	Column4	Column5
2	smiles	E1-CC2	E2-CC2	f1-CC2	f2-CC2
3	[H]C([H])([H])[H]	0.43295186	0.43295958	0.24972825	0.24973648
4	[H]N([H])[H]	0.26521952	0.35008064	0.06701544	0.03004918
5	[H]O[H]	0.28653735	0.363579	0.03775532	0
6	[H]C#C[H]	0.35862867	0.35862867	0	0
7	[H]C#N	0.31995762	0.33607406	0	0
8	[H]C([H])=O	0.15391355	0.29123378	0	0.09102332
9	[H]C([H])([H])C([H])([H])[H]	0.37613753	0.37614568	0	0
10	[H]OC([H])([H])[H]	0.26669063	0.33319051	0.0009443	0.07160772
11	[H]C#CC([H])([H])[H]	0.27338914	0.2857496	0	0.0011942
12	[H]C([H])([H])C#N	0.31965511	0.3344136	0.00000001	0.002747
13	[H]C(=O)C([H])([H])[H]	0.16572695	0.28555558	0.00004483	0.04090803
14	[H]C(=O)N([H])[H]	0.21631598	0.28221985	0.00093894	0.01318951
15	[H]C([H])([H])C([H])([H])C([H])([H])[H]	0.35880714	0.36925461	0.0015726	0.00298485
16	[H]OC([H])([H])C([H])([H])[H]	0.26669876	0.32436	0.00021511	0.02683589
17	[H]C([H])([H])OC([H])([H])[H]	0.2741024	0.31468983	0.04407212	0.06006052
18	[H]C1([H])C([H])([H])C1([H])[H]	0.32815231	0.3281839	0.00723734	0.00724217
19	[H]C1([H])OC1([H])[H]	0.29635574	0.29981734	0.00682283	0.06824817
20	[H]C([H])([H])C(=O)C([H])([H])[H]	0.17028194	0.2585645	0	0.03421965

MOTIVÁCIA

Výpočet hodnôt základného a excitovaného stavu molekuly

- Schroedingerova rovnica
 - diferenciálna rovnica kt. popisuje vývoj fyzikálneho systému
 - nevieme presne riešiť -> aproximácie

Metódy:

- Coupled Cluster Singles Doubles Triples (CCSDT) – nevhodná pre väčšie molekuly
- Kvantové Monte Carlo (QMC) – výpočty rádovo 1M CPU hodín
- Density functional theory (DFT) – menej náročná, menej presná

MODELY

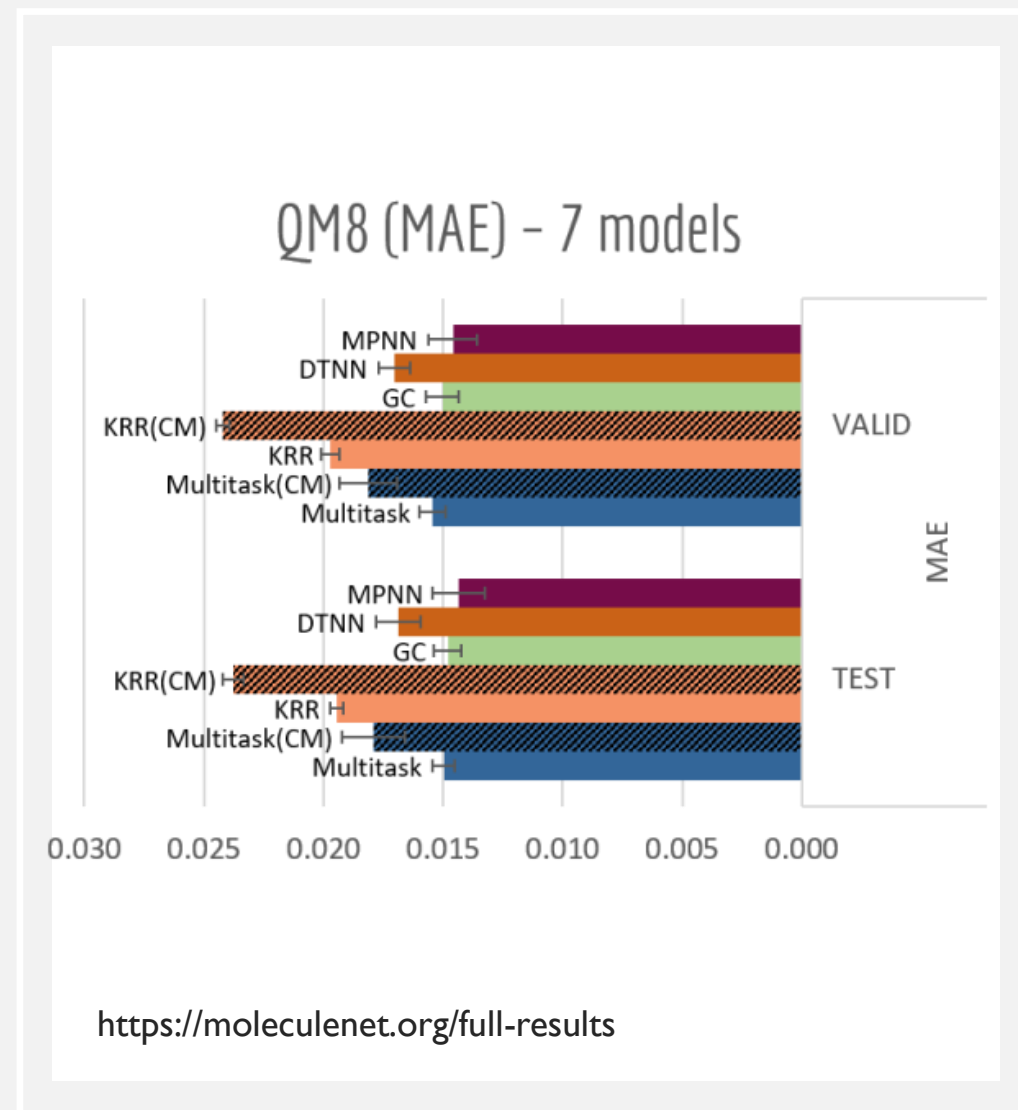
Multitask regressor

Kernel ridge regressor (KRR)

Graph convolution (GC)

Message-passing neural network (MPNN) typ GNN

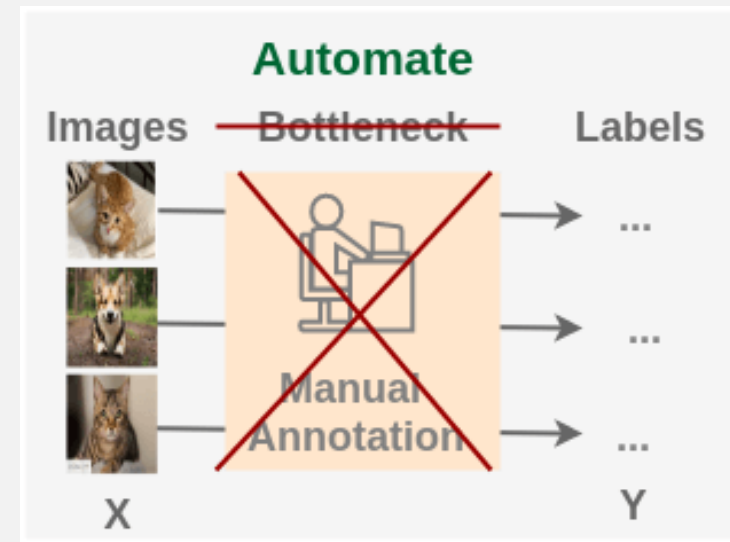
Deep tensor neural networks (DTNN)



MOTIVÁCIA

Metódy učenia bez učiteľa

- využívajú neoznačené dáta
- hľadajú štruktúry, vzory, podobnosti v dátach
- výstupom sú klastre dát bez anotácií
- dotrénovanie anotácií klastrov
 - tréning s učiteľom
 - menšie množstvo dát



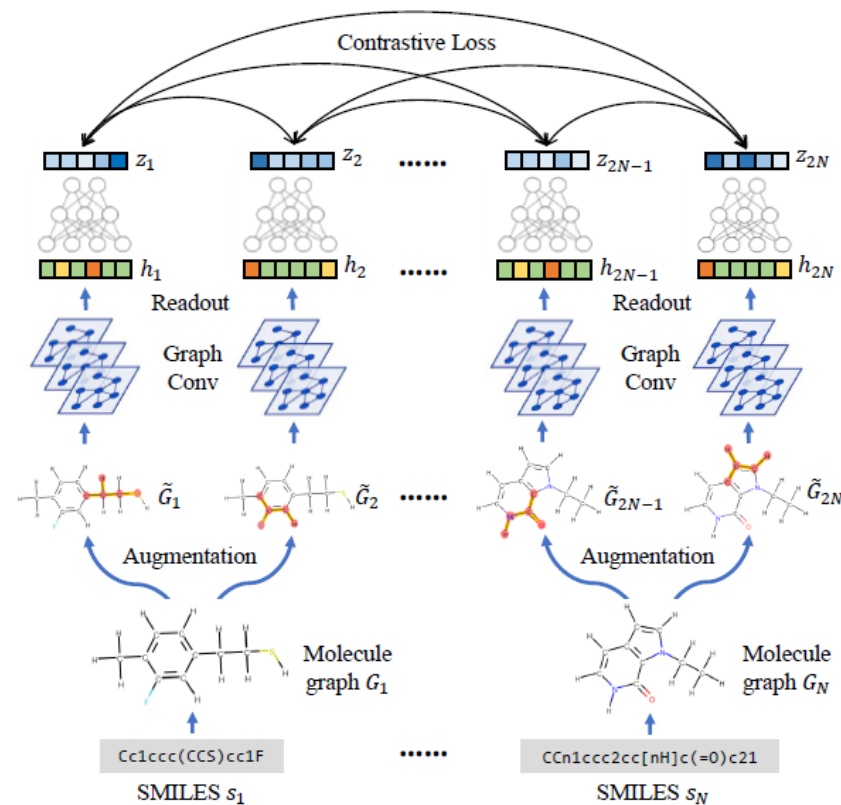
UNSUPERVISED MODELY

MolCLR

- kontrastívne učenie (SimCLR)
- GNN

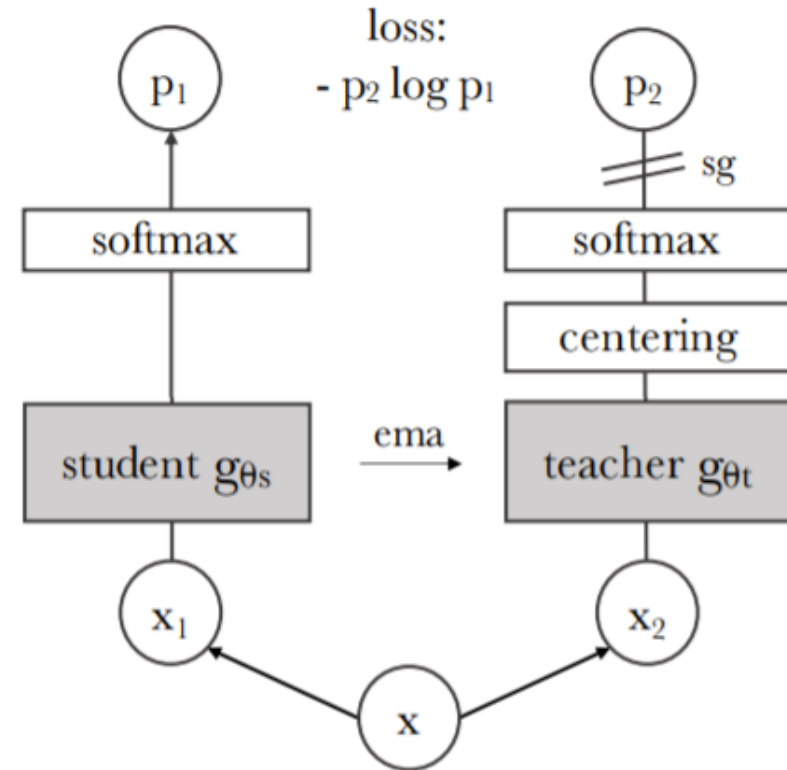
ChemBERTa

- RoBERTa →
BERT →
language representation model



DINO ARCHITEKTÚRA

- DINO: Self-**D**istillation with **no** labels
 - samokontrolovaná metóda
- študent a učiteľ – rovnaká architektúra
- vstupy – transformácie obrázkov
- chyba:
 - contrastive loss
 - spätne šírená len v študentskej sieti
- váhy učiteľa – ema (exponential moving average) $W_t = \lambda W_t + (1 - \lambda)W_s$



CIELE

- Spracovať prehľad metód klasifikácie a predikcie vlastností molekúl a najnovších techník učenia sa bez učiteľa alebo čiastočného učenia bez učiteľa, ktoré sú využívané v oblasti počítačového videnia.
- Navrhnuť novú metódu klasifikácie a predikcie vlastností molekúl, ktorá bude inšpirovaná existujúcimi technikami učenia sa bez učiteľa alebo čiastočného učenia bez učiteľa z oblasti počítačového videnia.
- Implementovať a experimentálne overiť navrhnuté riešenie testovaním na dátach pre klasifikáciu a predikciu molekulárnych vlastností.
- Porovnať dosiahnuté výsledky navrhnutého riešenia s doteraz najúspešnejšími technikami vyvinutými v oblasti molekulárnych vlastností

ĎAKUJEM ZA POZORNOST

ZDROJE A UŽITOČNÉ ODKAZY

- <https://arxiv.org/pdf/1703.00564.pdf>
- <https://arxiv.org/abs/2104.14294>
- <https://arxiv.org/pdf/1706.03762.pdf>
- <https://arxiv.org/pdf/2103.03404.pdf>
- <https://towardsdatascience.com/on-dino-self-distillation-with-no-labels-c29e9365e382>
- https://www.youtube.com/watch?v=h3ij3F3cPlk&ab_channel=YannicKilcher
- <https://towardsdatascience.com/transformers-an-exciting-revolution-from-text-to-videos-dc70a15e617b>
- https://www.youtube.com/watch?v=iDulhoQ2pro&t=10s&ab_channel=YannicKilcher