iCluto v2

Self-Distillation na break-junction datech

Jak naučit neuronovou síť číst vodivostní křivky, aniž bychom jí kdy řekli, co znamenají.

Oliver Klimt — Obhajoba DP — **FEL ČVUT & ÚOCHB AV ČR**

Praha — 15. červen 2026

Vedoucí: **Ing. Ladislav Sieger, CSc.** Konzultanti: **RNDr. Jaroslav Vacek, Ph.D., RNDr. Jindřich Nejedlý, Ph.D.**

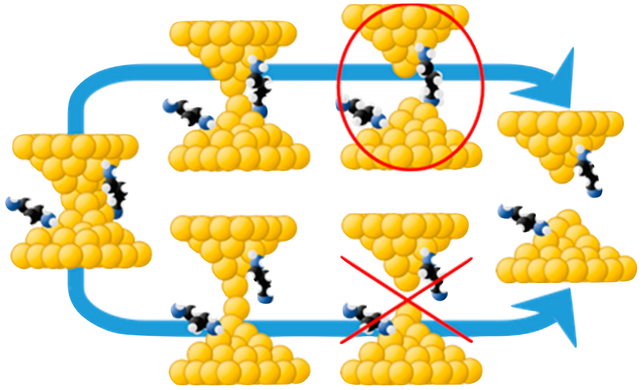
Handout, prezentace i diplomová práce ke stažení na…

https://defence.icluto.oklimt.com/

Problém: vzácný signál ve velkém datasetu

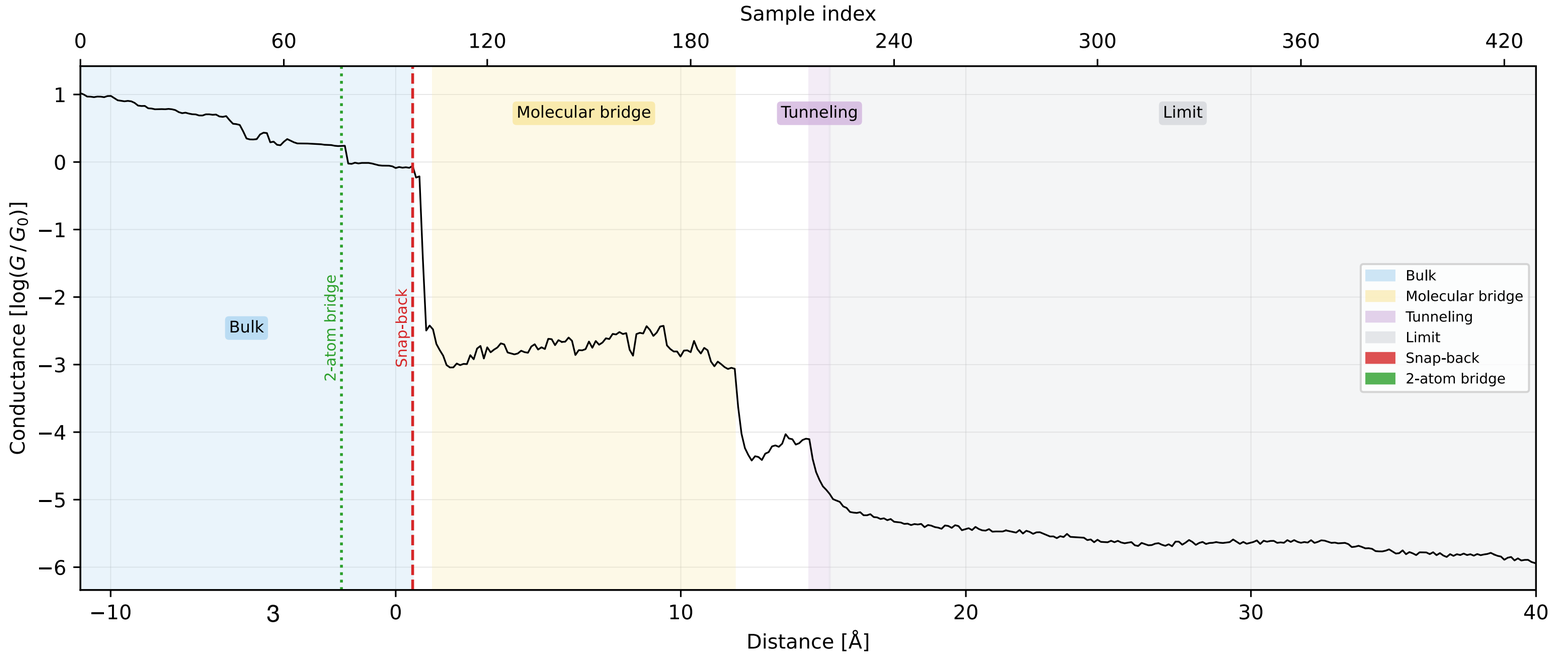
Klasické metody nestačí.

- desetitisíce vodivostních křivek na jeden experiment
- většinou pouze tunelovací proud, málokdy molekula
- Cíl: najít je automaticky a bez ruční anotace dat.



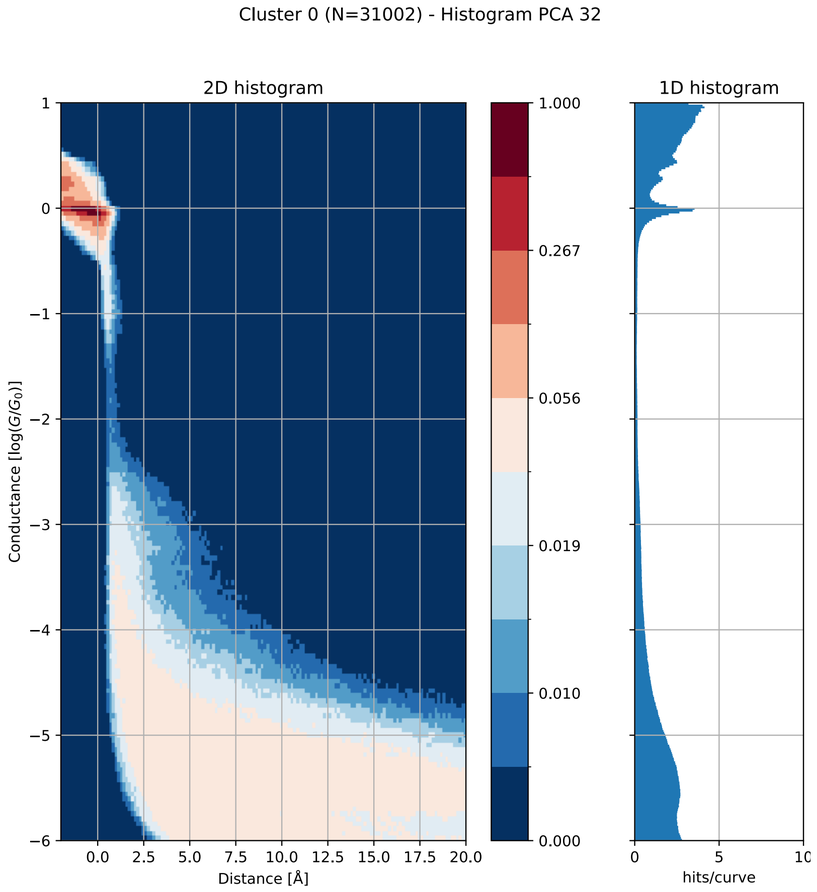
J. Phys. Chem. Lett. 2021, 12, 44, 10802–10807

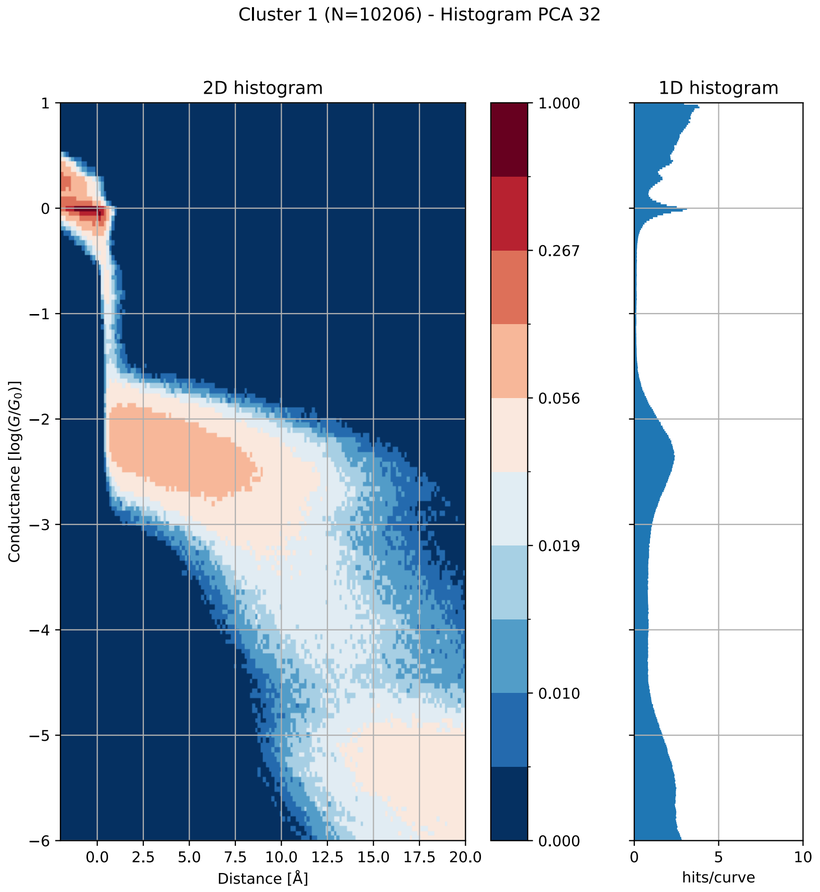
Molekulární elektronika:
- stavíme přístroje z organických molekul
- velmi malý footprint, lepší stericita
- molekula jako:
  - drát, dioda, switch či senzor
- alternativa ke křemíku



iCluto v1

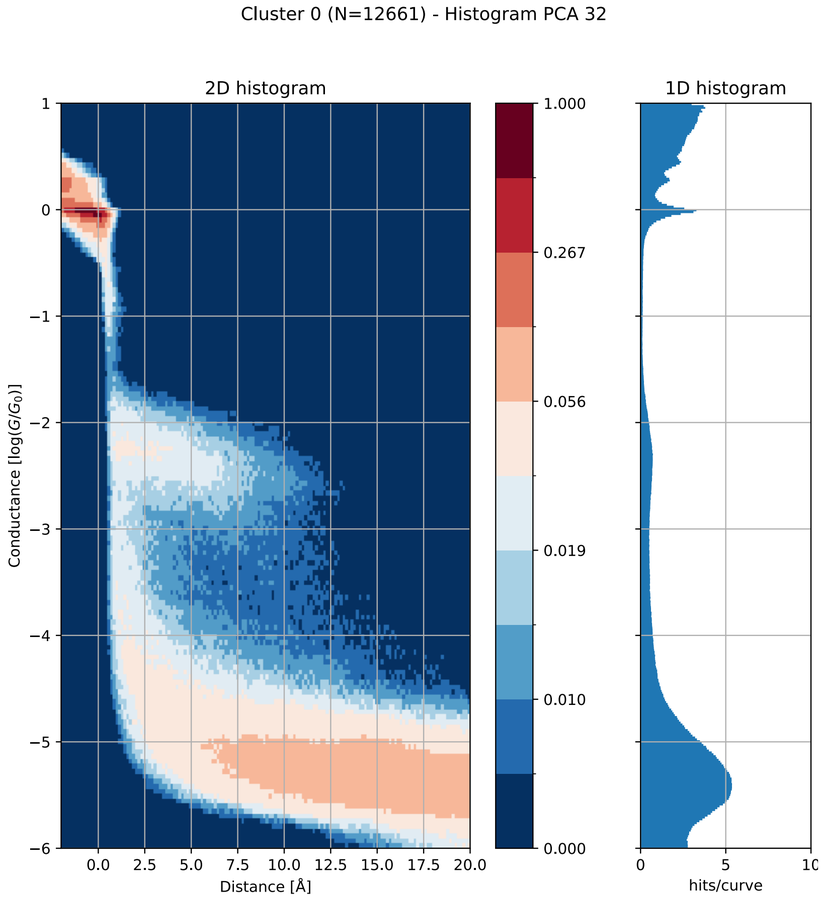
Předchozí verze byla křehká.

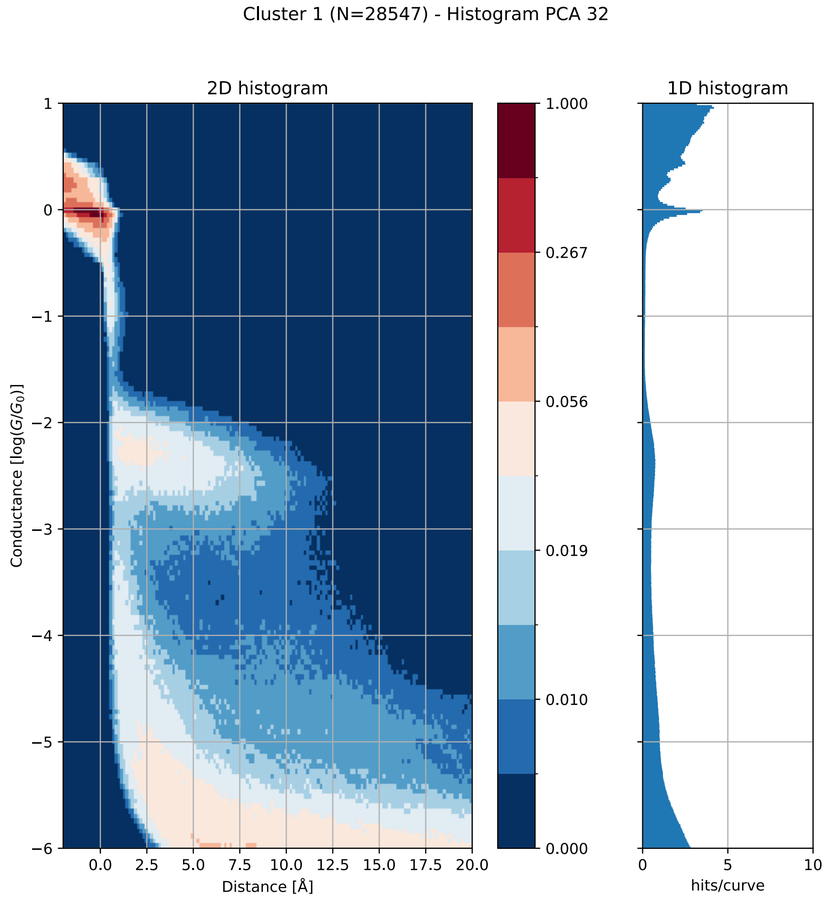




iCluto v1

Předchozí verze byla křehká.





☹️

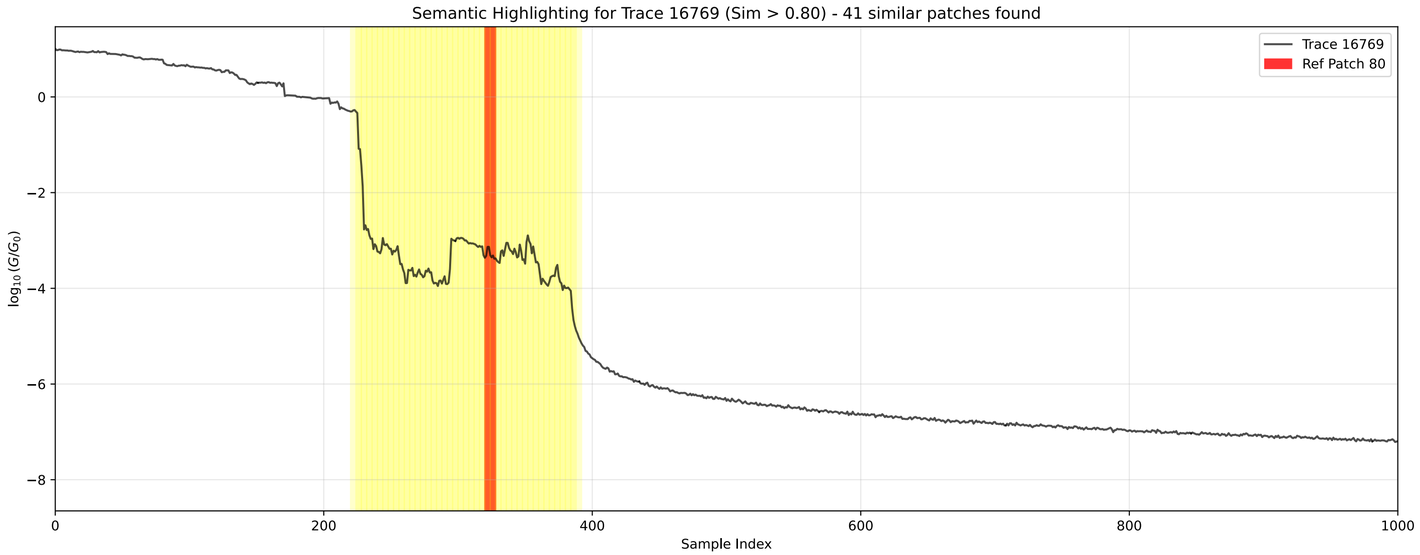
Inspirace z počítačového vidění 👀

Dříve jsme popisovali celý obrázek, nyní segmenty (patches)

Siméoni, Oriane, et al. "Dinov3." *arXiv preprint arXiv:2508.10104* (2025).

Inspirace z počítačového vidění 👀

Dříve jsme popisovali celý obrázek, nyní segmenty (patches)



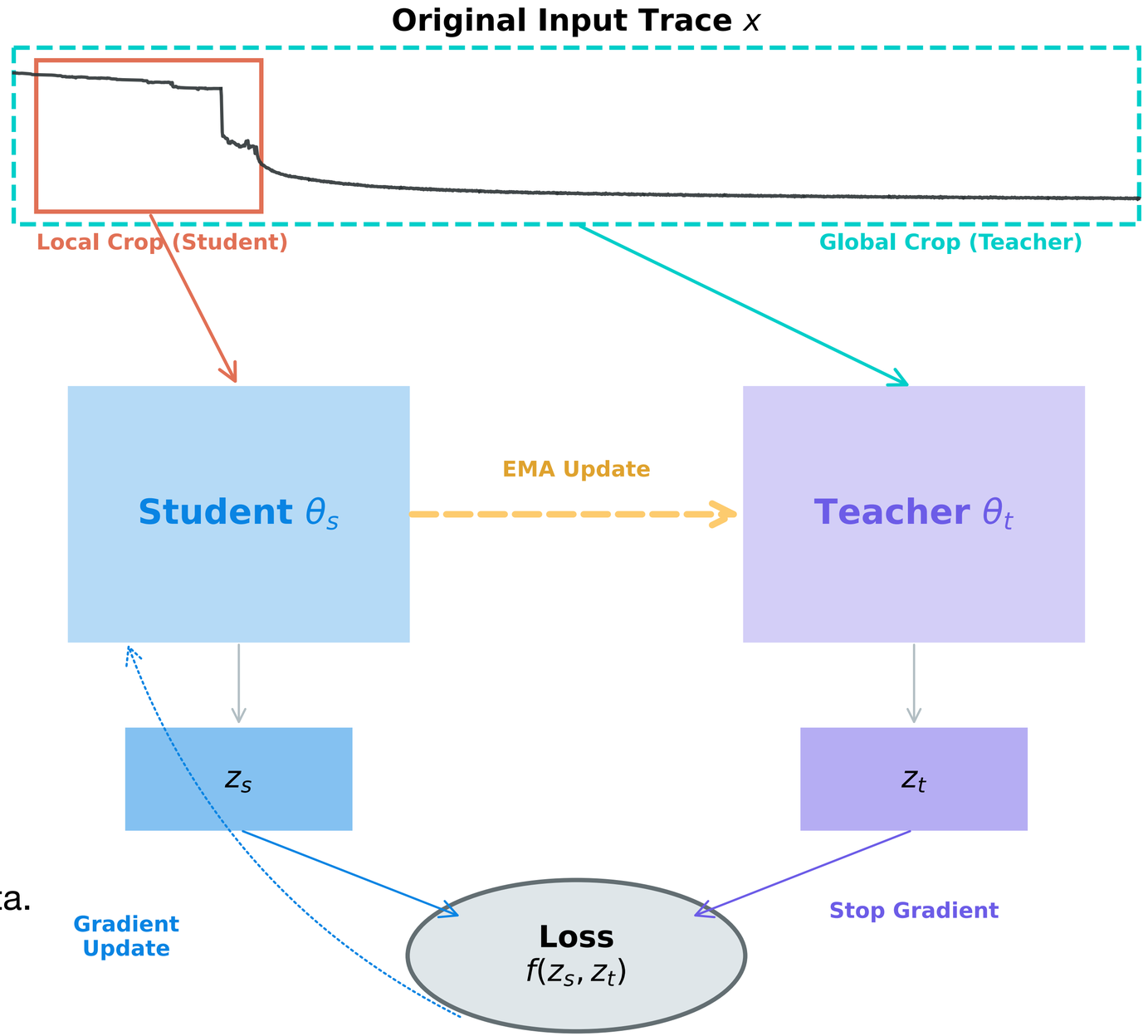
Siméoni, Oriane, et al. "Dinov3." *arXiv preprint arXiv:2508.10104* (2025).

Student & učitel

Jak DINO funguje.

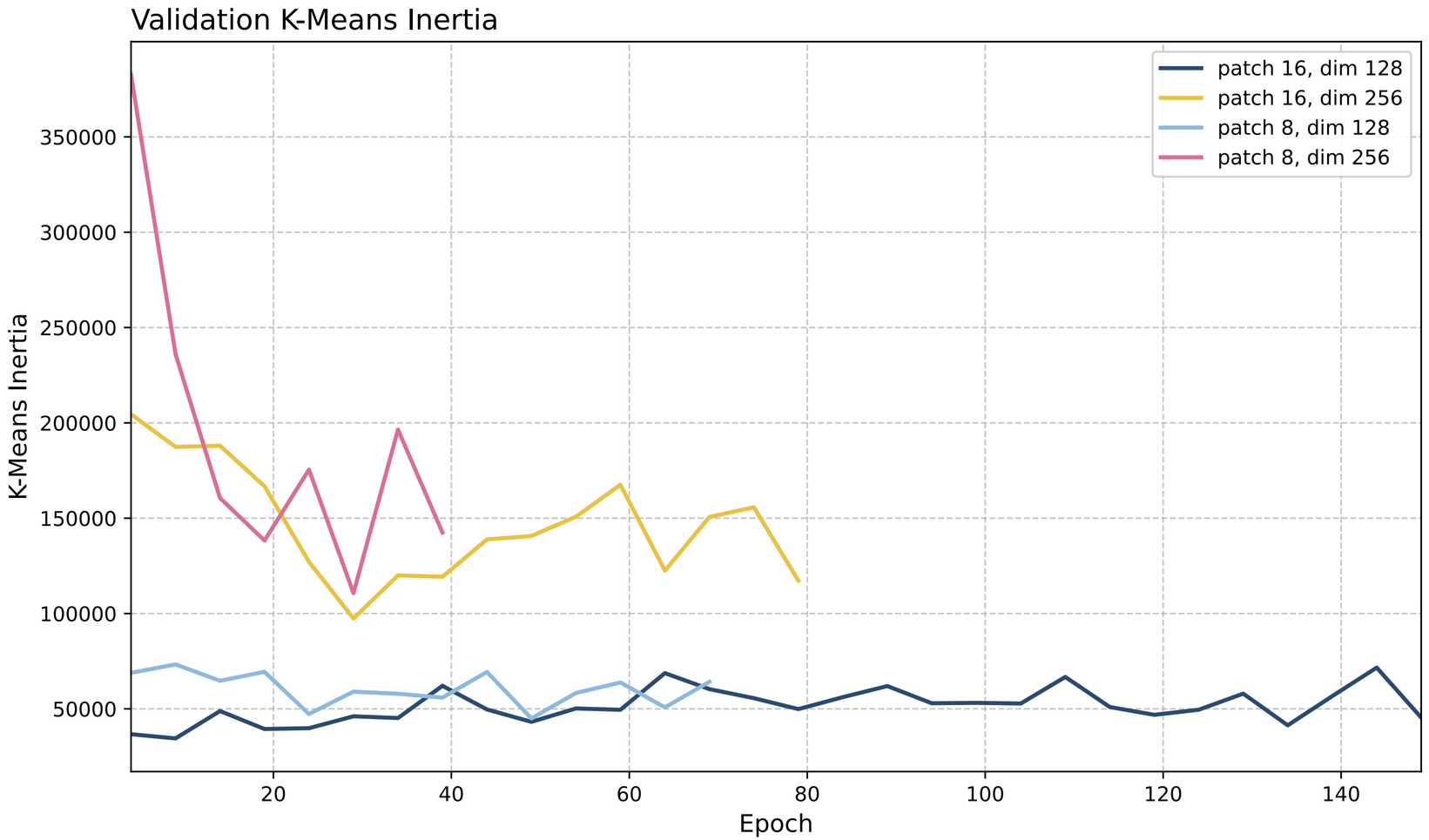
založeno na Vision Transformer architektuře

1. Vezmeme křivku
2. Studentovi poskytneme malý (lokální) úsek.
3. Učitel dostane větší (globální) část křivky.
4. Trénujeme studenta tak, aby se shodl s učitelem.
5. Váhy učitele jsou průměrem minulých vah studenta.
   Bez anotací, bez jediné supervize.



Trénink

… a výběr modelu



400 000+
trénovacích křivek

20
GPU dní

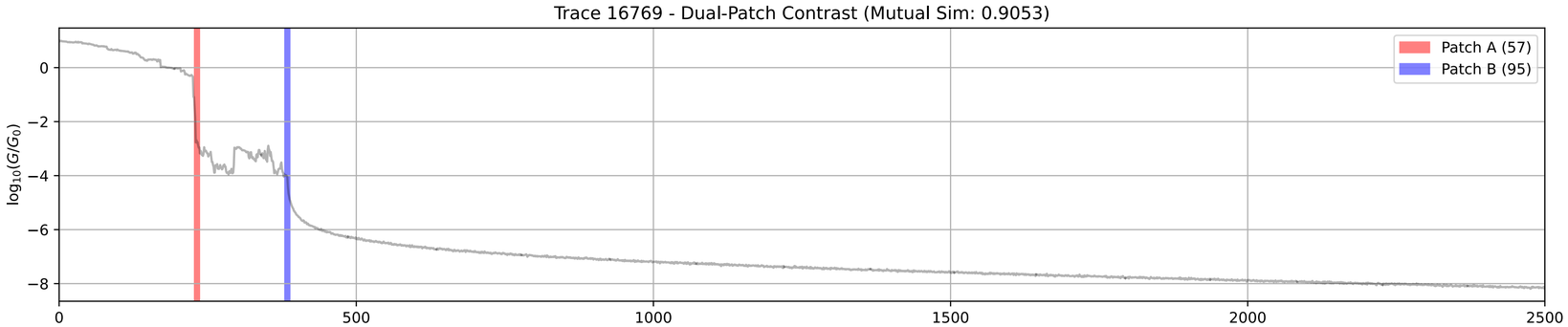
Finální model

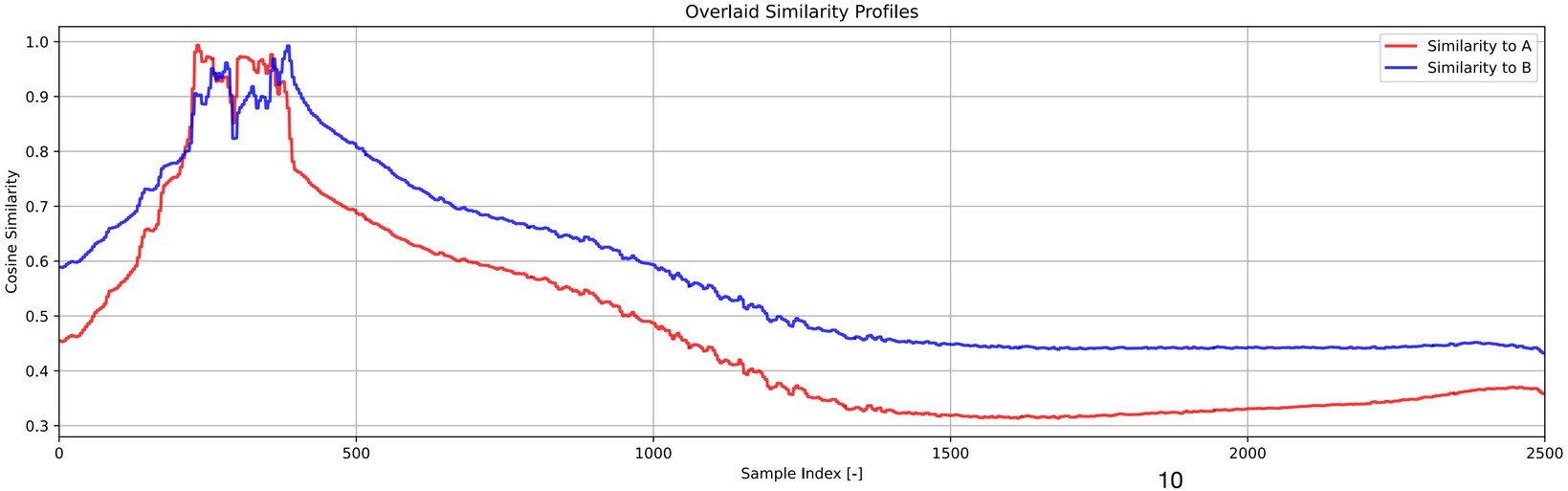
patch 8
dim 256

Co jsme tréninkem získali?

Naučenou reprezentaci každého *patche*!

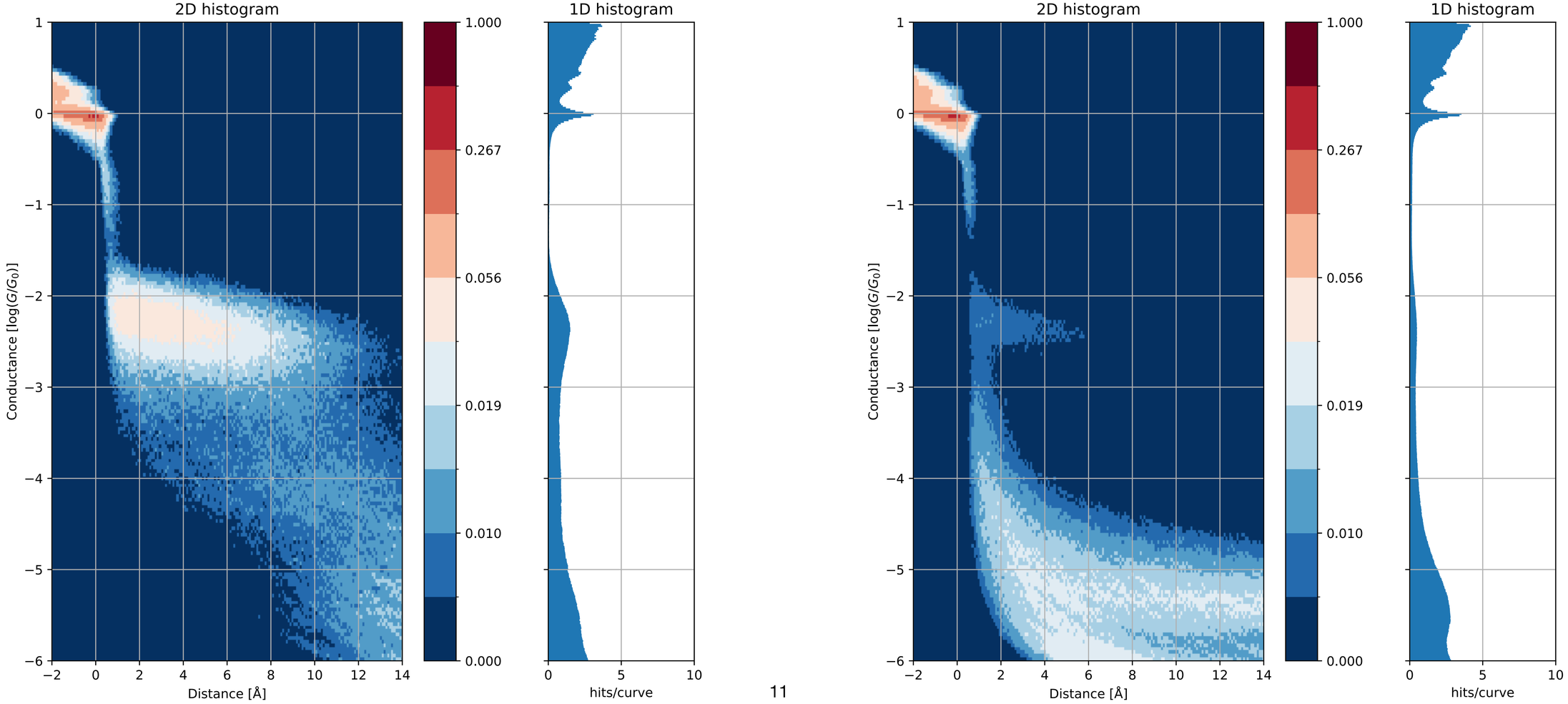
$$\text{Cosine Similarity}(\mathbf{a}, \mathbf{b}) = \frac{\mathbf{a}^\top \cdot \mathbf{b}}{\|\mathbf{a}\| \cdot \|\mathbf{b}\|}$$





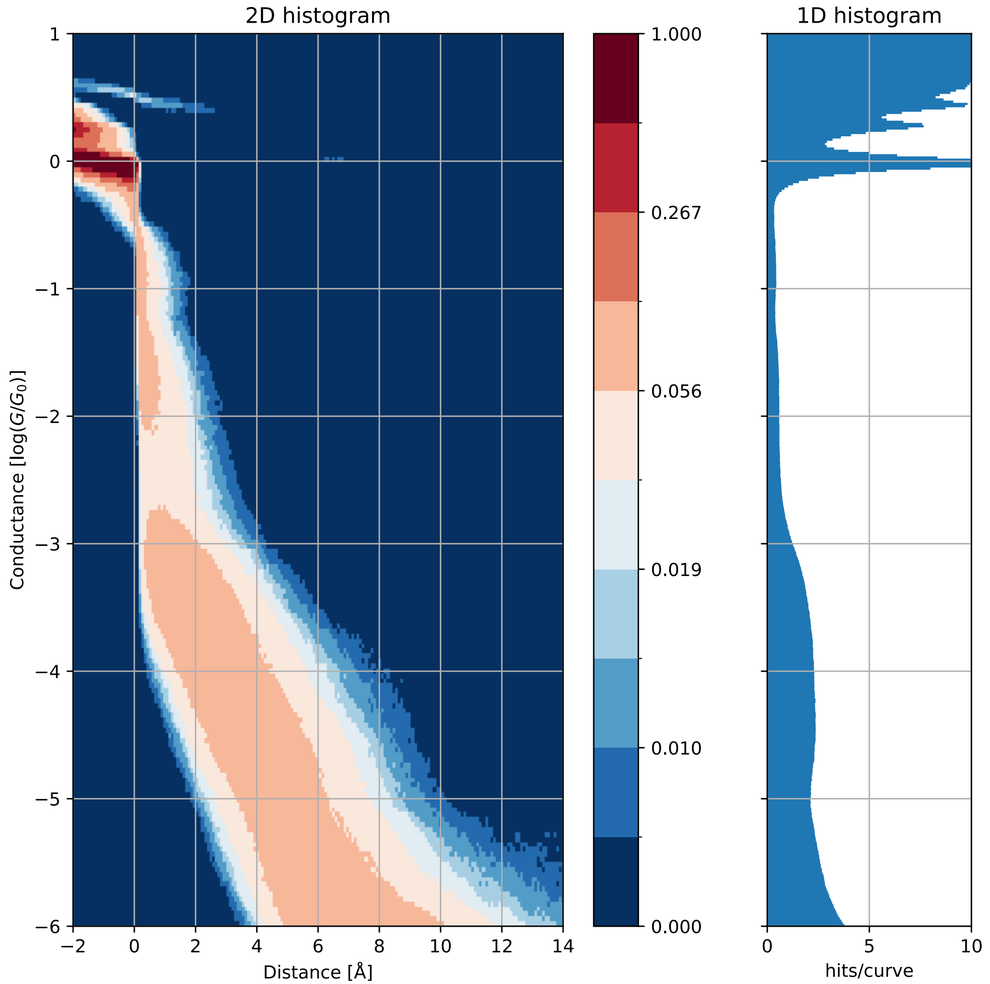
Similarity search

Jeden referenční patch → 9 766 molekulárních křivek

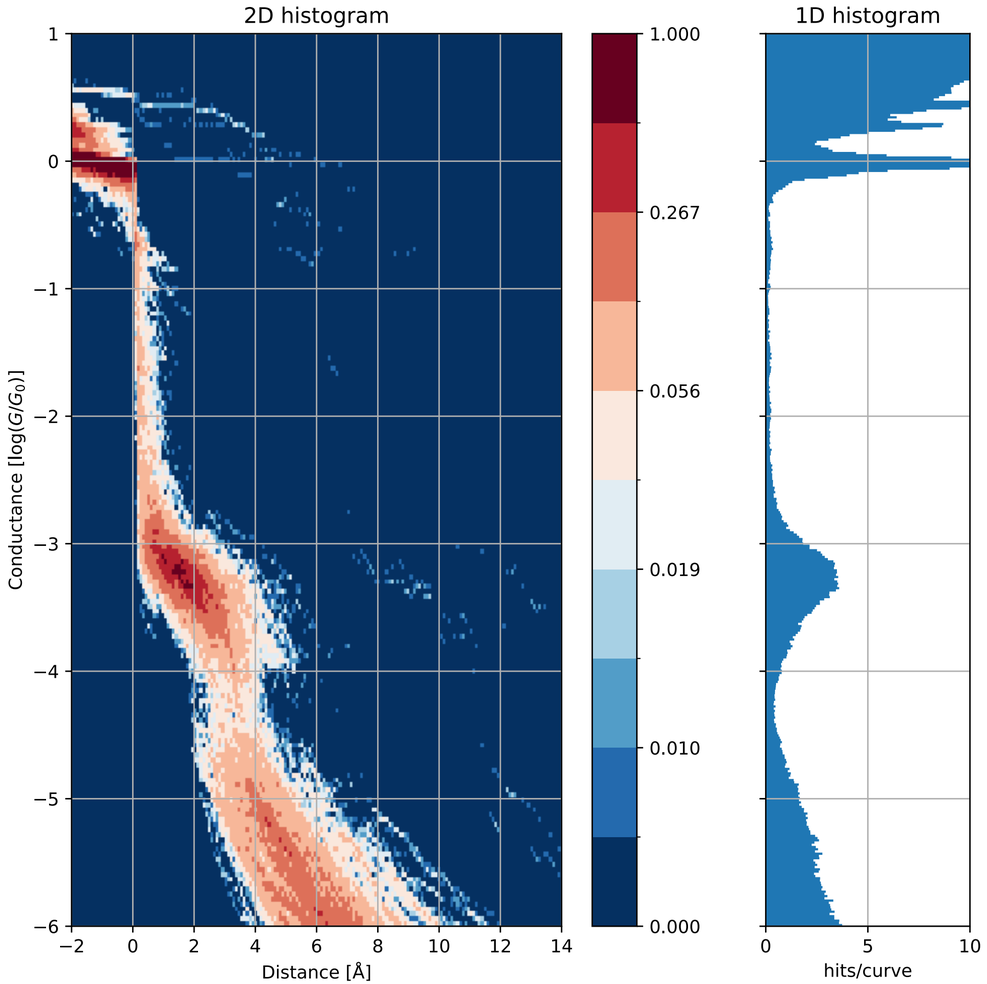


Obtížný dataset

Remaining Traces (N=33752)



Matched Traces (N=316)



Kam jsme BJ analýze pokročili?

Žádná anotace

Neuronová síť se učí ze samotných dat.

Žádné parametry aparatury

Stačí referenční patch a podobnostní práh.

„Vhled“ do křivek

Neuronová síť rozliší, co je podstatné.

V čem budeme pokračovat?:

- Hledání pomocí více patchů, jeden není dostatečný.
- Anotace pomocí DINO ➡️ získáme benchmark jako je ImageNet.

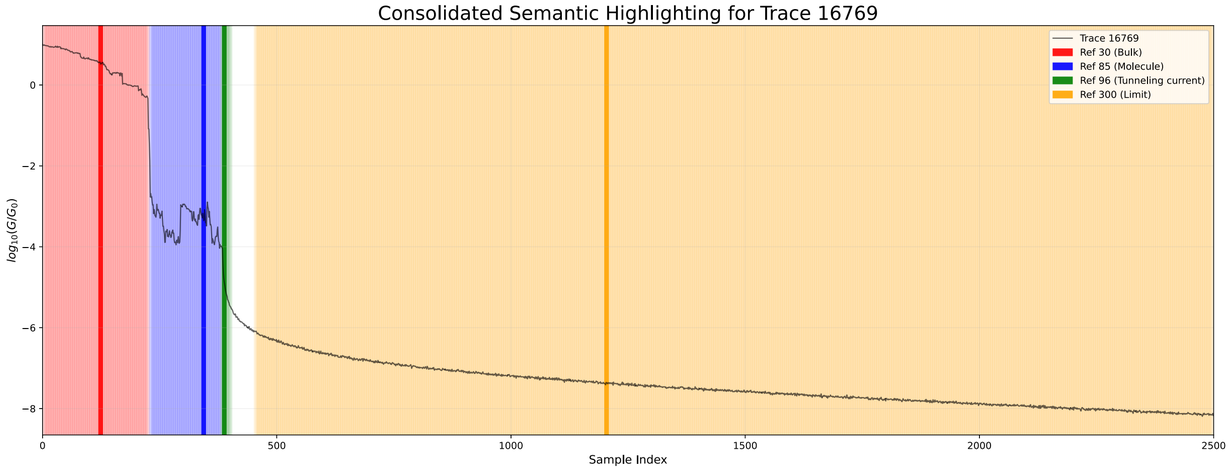
Děkuji za pozornost.

... velké díky Dr. Siegerovi, Dr. Vackovi & Dr. Nejedlému 🙌

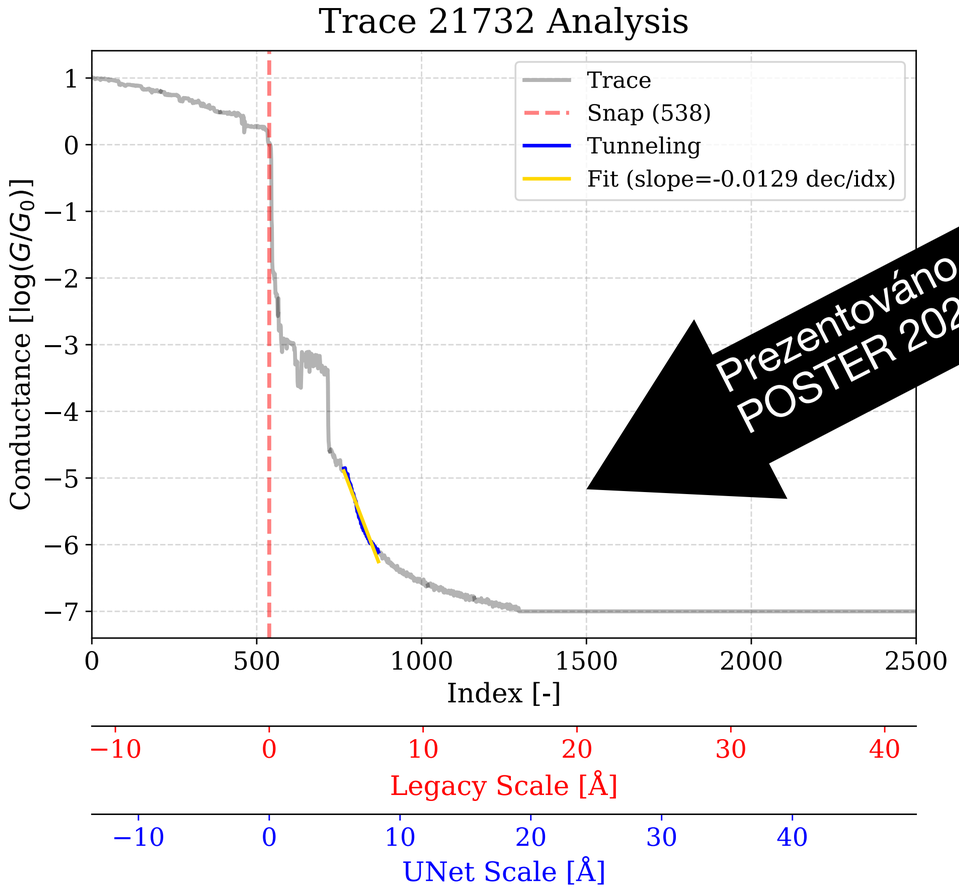
FAKULTA
ELEKTROTECHNICKÁ
ČVUT V PRAZE

ÚOCHB AV ČR
IOCB PRAGUE

Segmentace pomocí iCluto & DINO

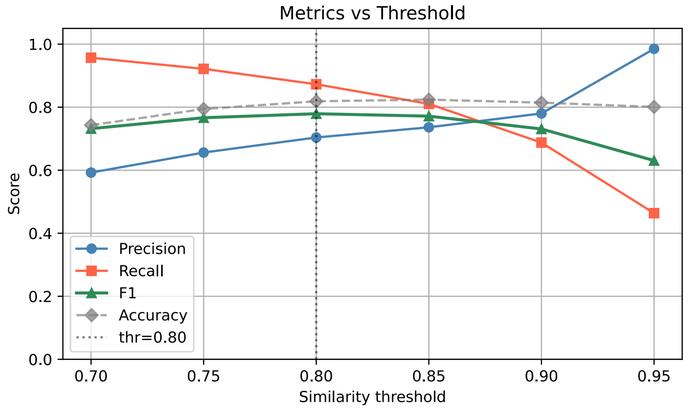


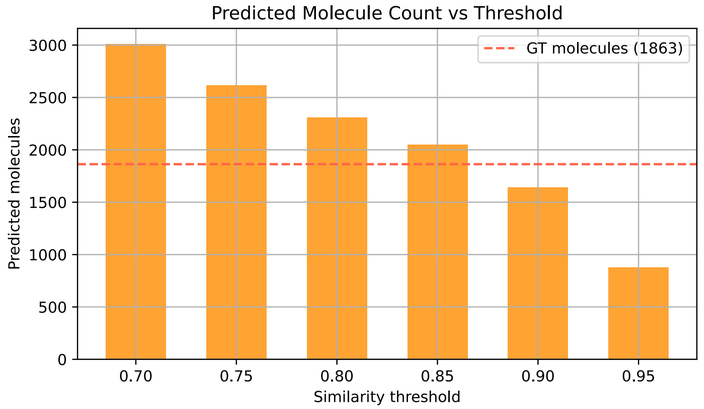
Odhad rozteče elektrod pomocí UNet + RANSAC



Prezentováno na POSTER 2026

Bro-Jørgensen, William, et al.
"Trusting our machines: validating machine learning models for single-molecule transport experiments." 2022





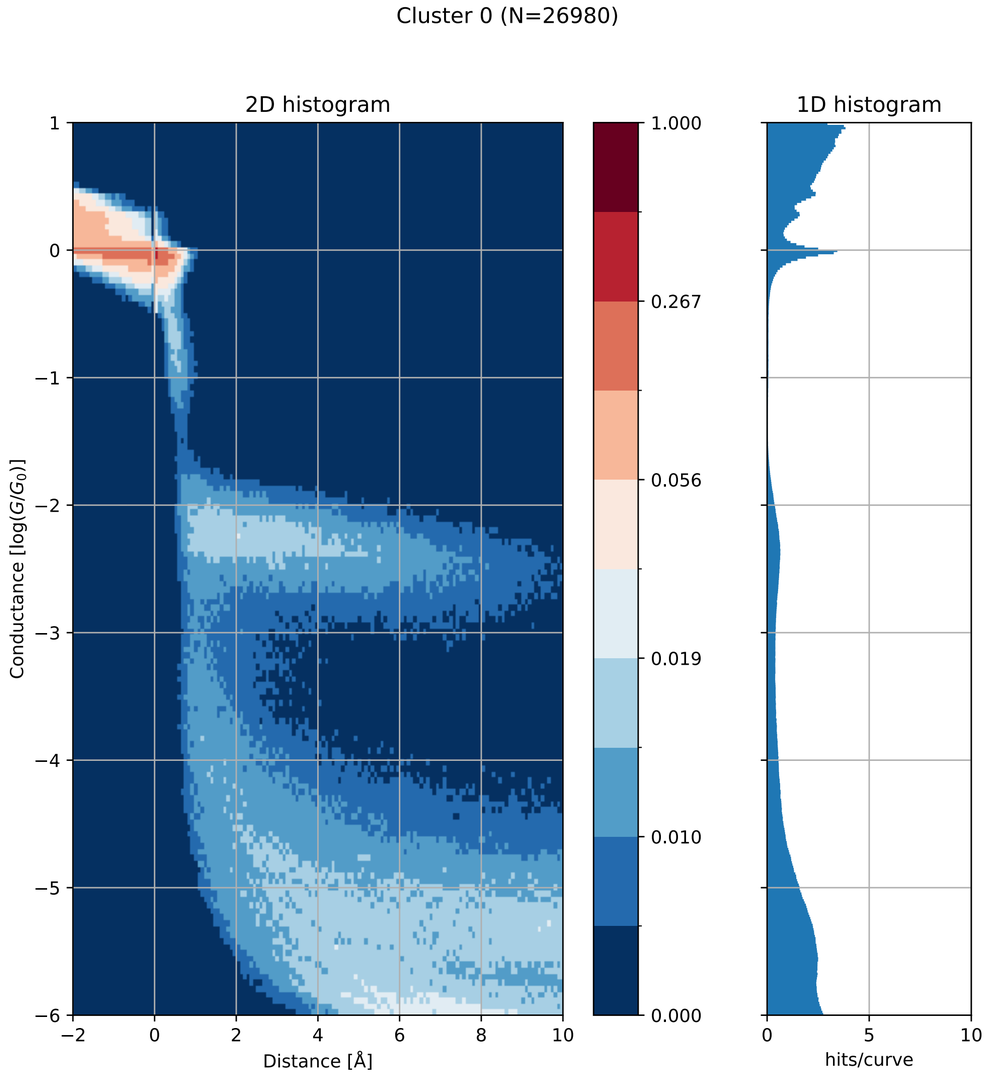
Validace na Kodaňském datasetu

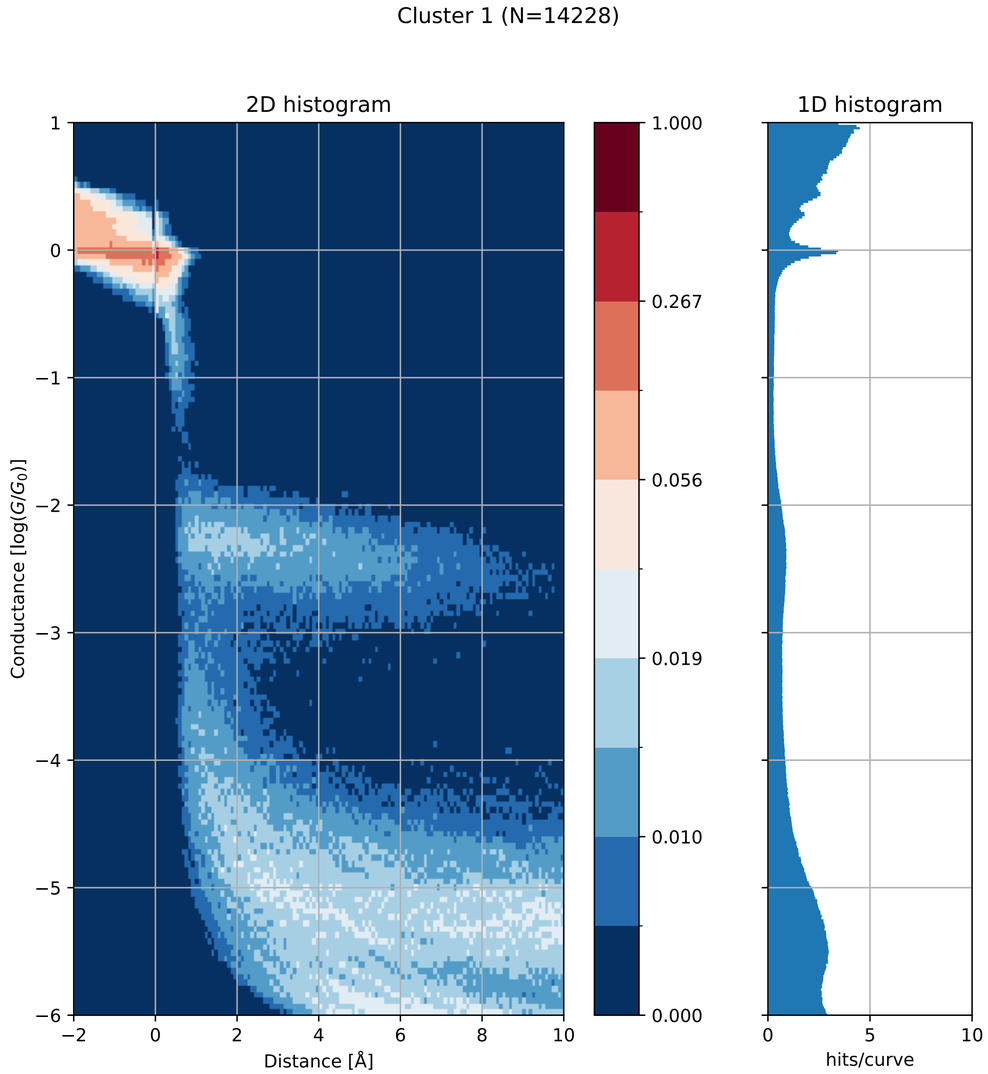
- teplota 4 K, jiný sampling rate
- pouze resampling, žádný finetuning

Q&A

CLS token I

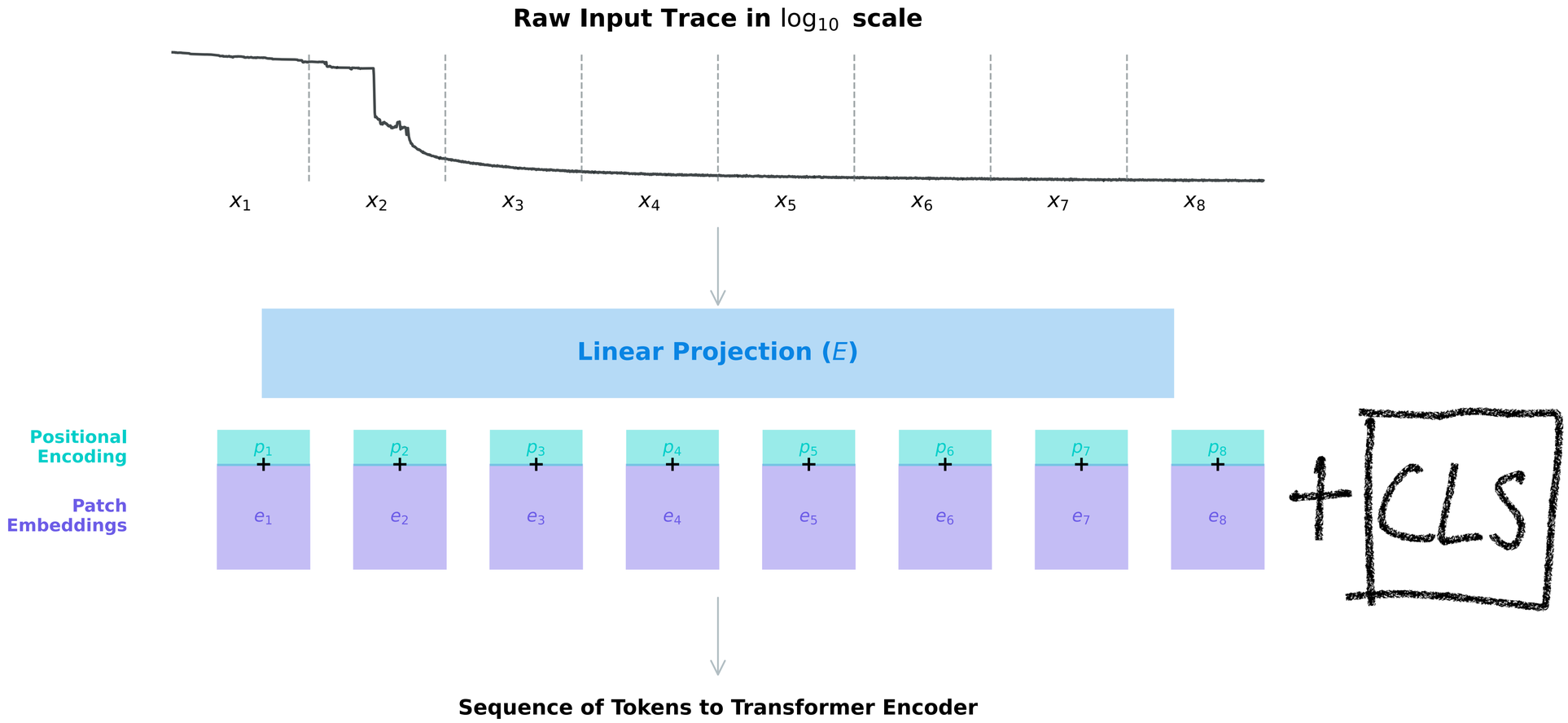
Odpověď: K-Means při K=2 selhalo i u CLS.





CLS token II

Odpověď: K-Means při K=2 selhalo i u CLS.



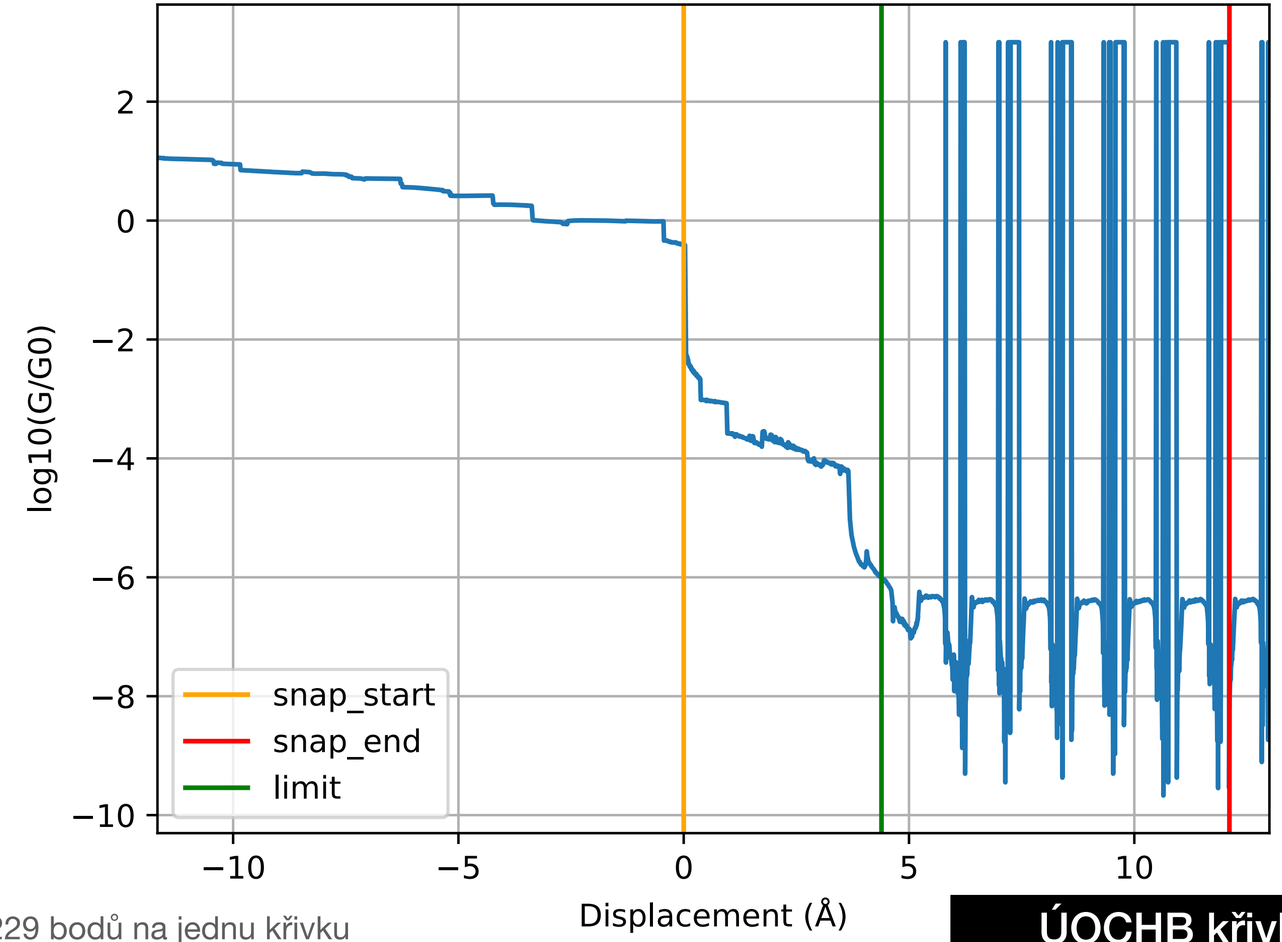
Externí validace I

Odpověď: TL;DR ➡️ Je to složité…

- Absence anotací (jak na ÚOCHB, tak obecně)
- Jiná aparatura, ale hlavně jiné prostředí získávání křivek
  - pokojová teplota ×téměř 0 K (paper zmiňuje 4 K, nezmiňuje tlak)
  - Naprosto jiná vzorkovací frekvence (paper neuvádí jaká) a artefakty.
- Z dat špatně čitelný tunelovací segment ➡️ Obtížné určit indifaktor.
  - Velmi odlišné okno záběru (hodně bulk, hodně molecular bridge).
  - Na ÚOCHB dominuje limit přístroje, ale záznamové okno dáno spíše historicky.

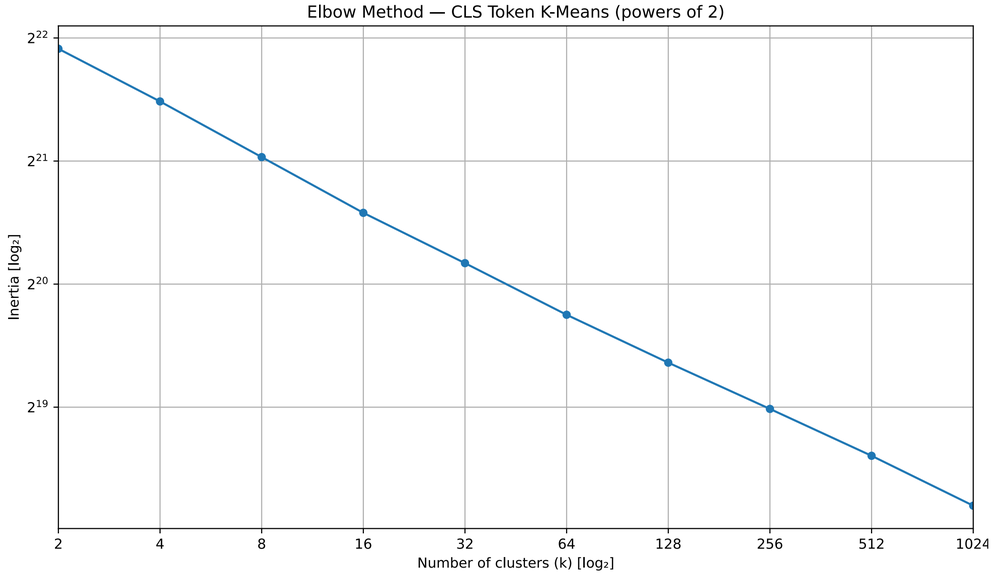
Externí validace II

Odpověď: TL;DR ➡️ Je to složité

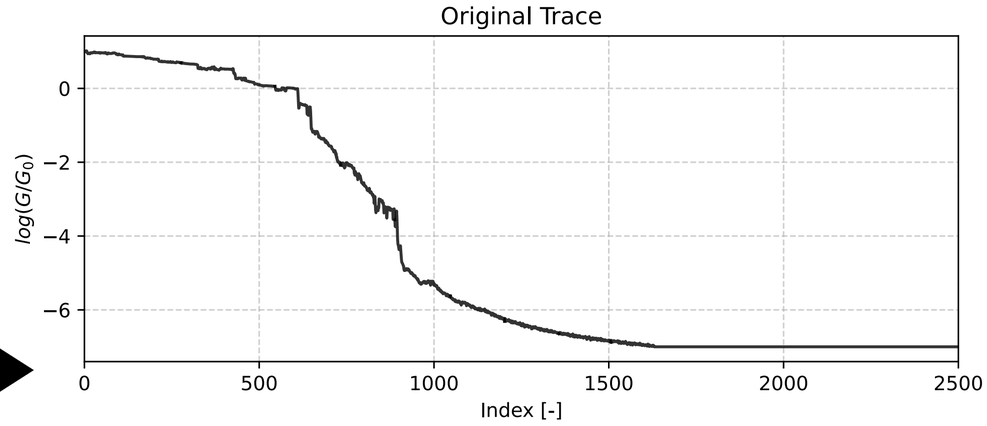


4229 bodů na jednu křivku

ÚOCHB křivka

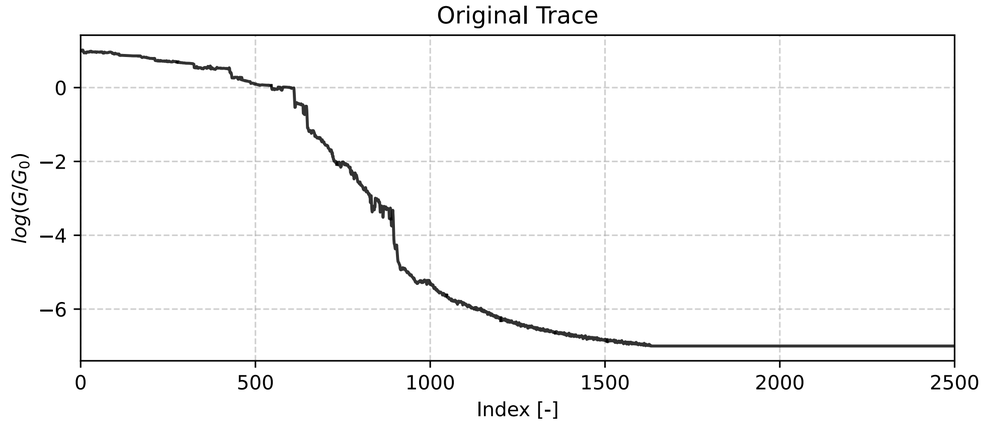


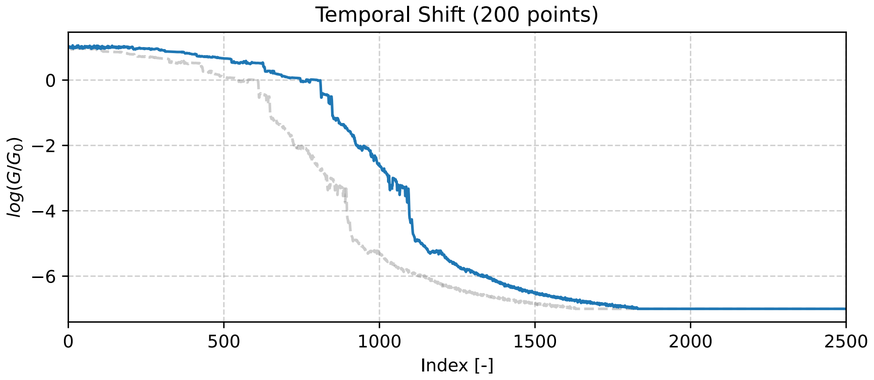
Resampling nutný, případně přetrénovat DINO s ‘agresivnější’ augmentací.

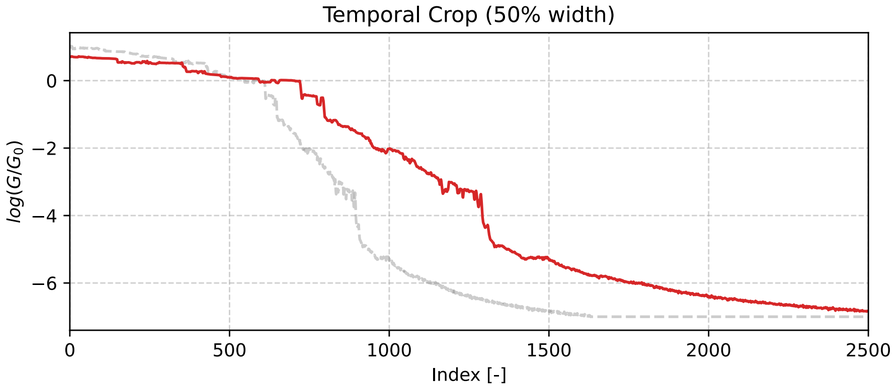


Externí validace III

Příklad augmentace.

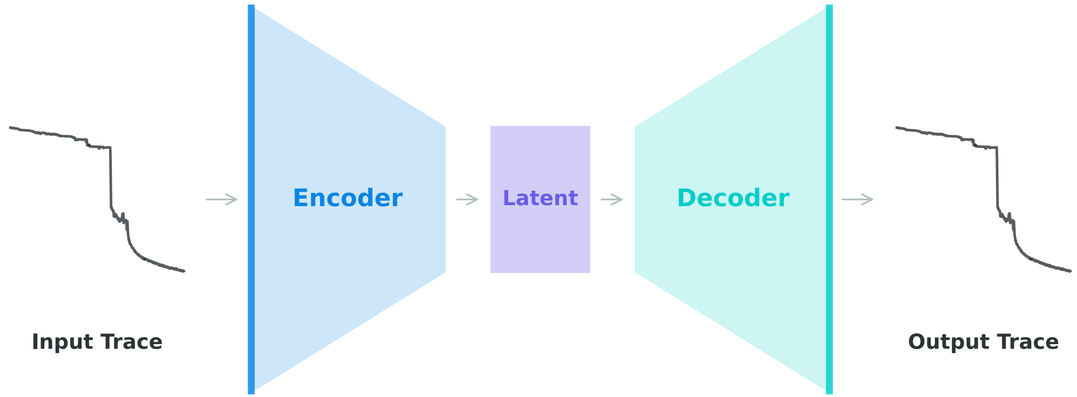




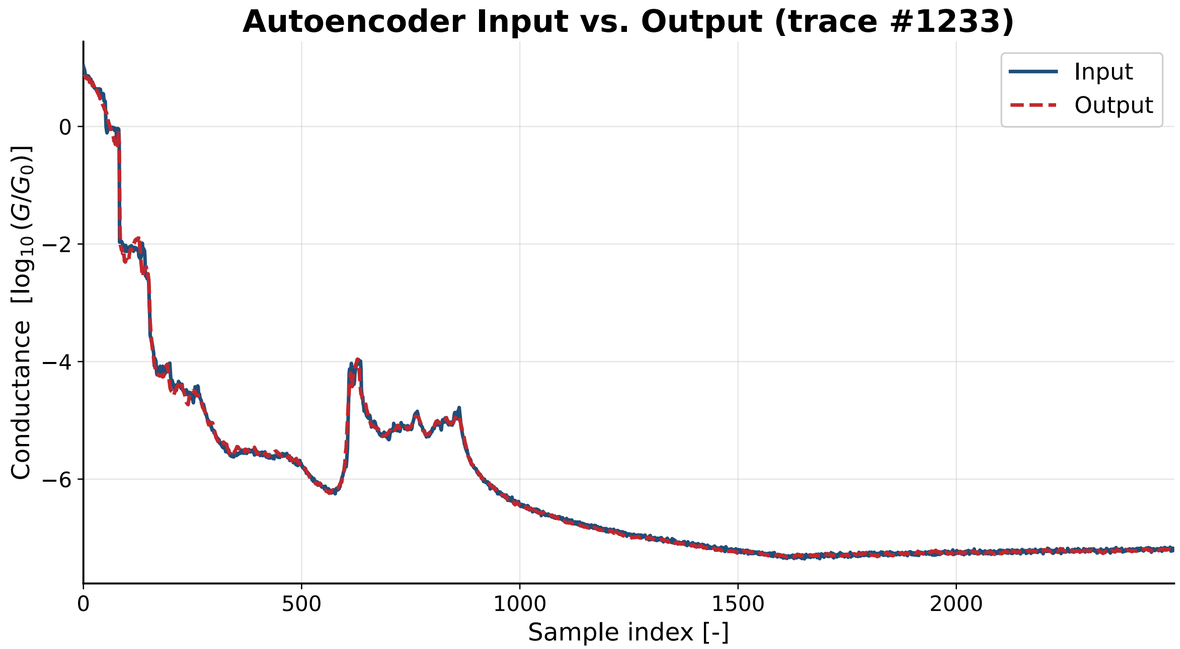


Auto-encoder I

Bere v potaz celou křivku.

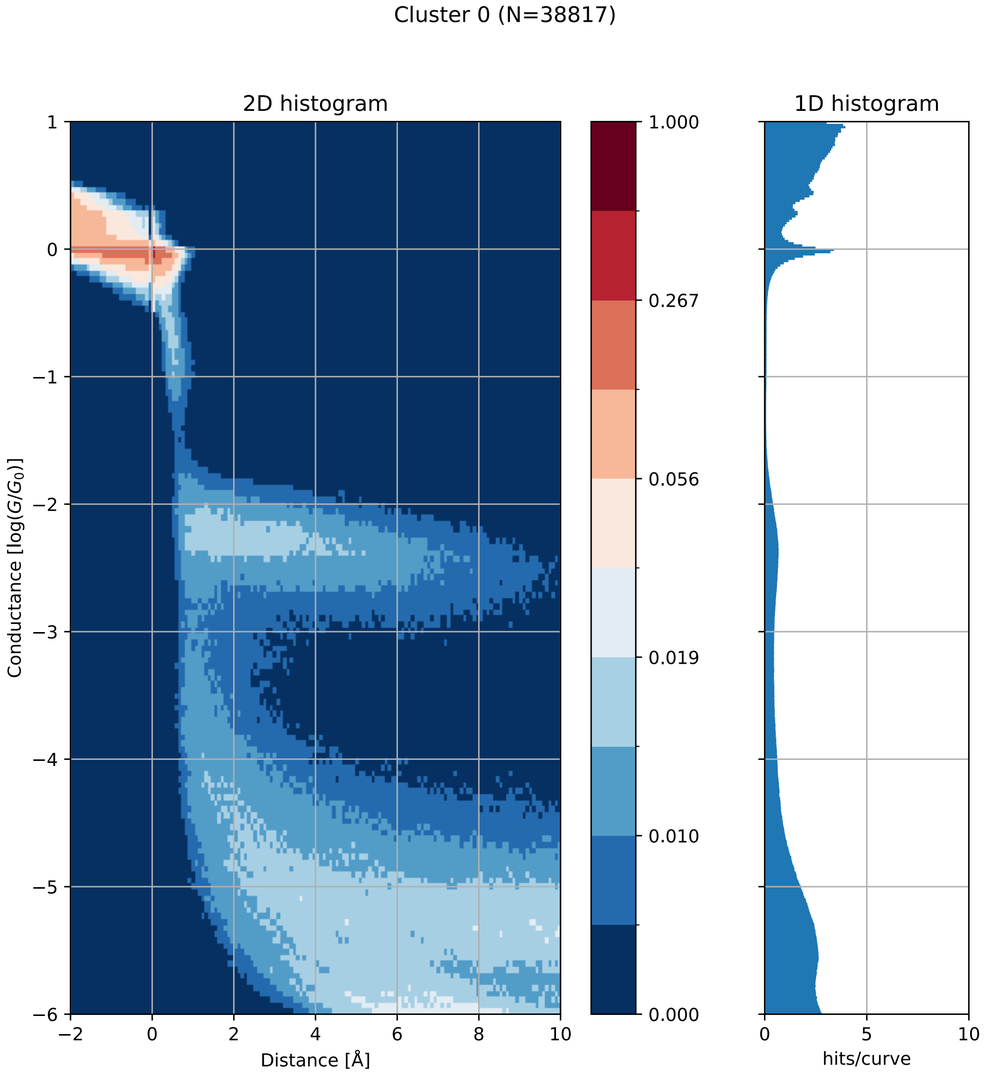


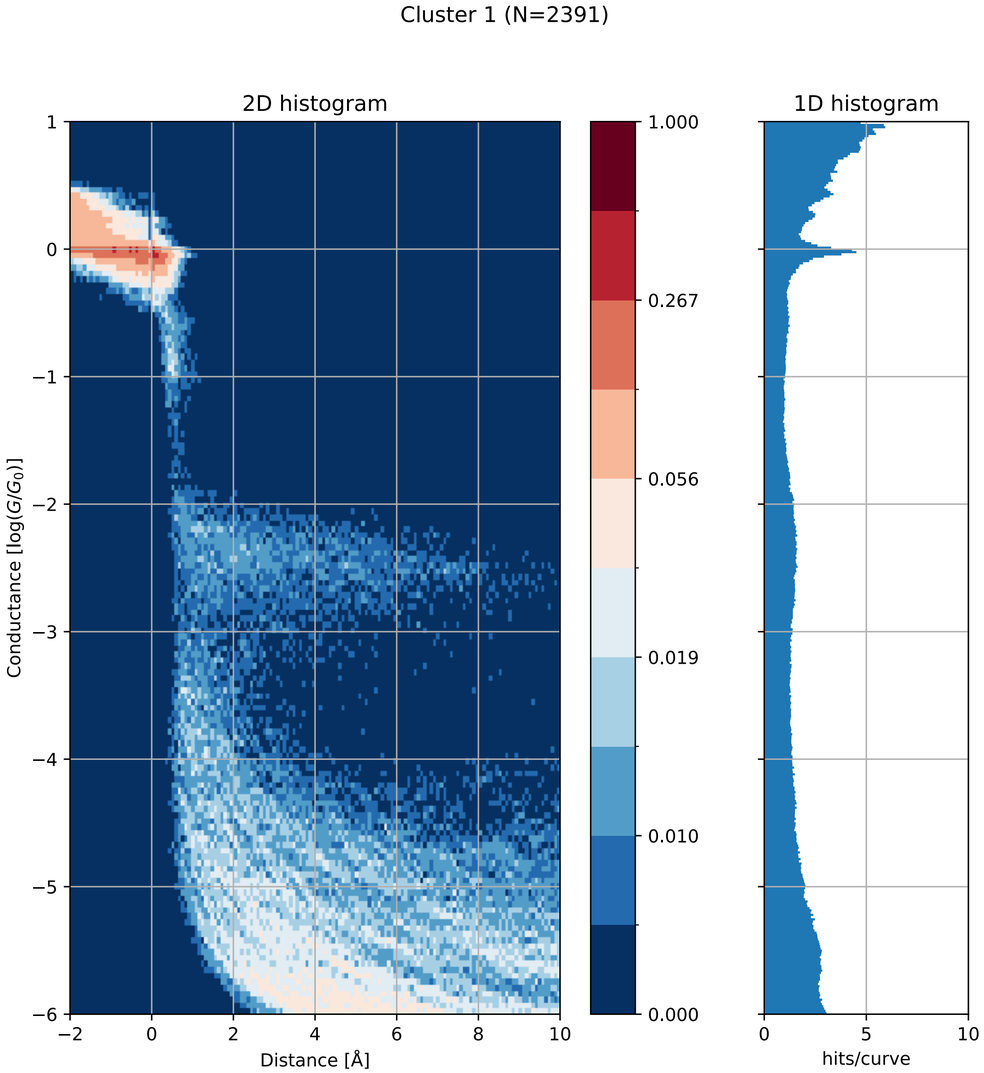
- Jak vymezit vodivost molekuly?
  - podobné úskalí jako iCluto v1
- Natrénováno v řádu minut.
  - … a výsledky jsou neuspokojující.
- Molekulární režim má různé délky
  - -> různě dlouhé input vektory
  - lze vyřešit paddingem či fixním výřezem.
  - ale molekuly mají od ~ 4 Å do ~ 13 Å



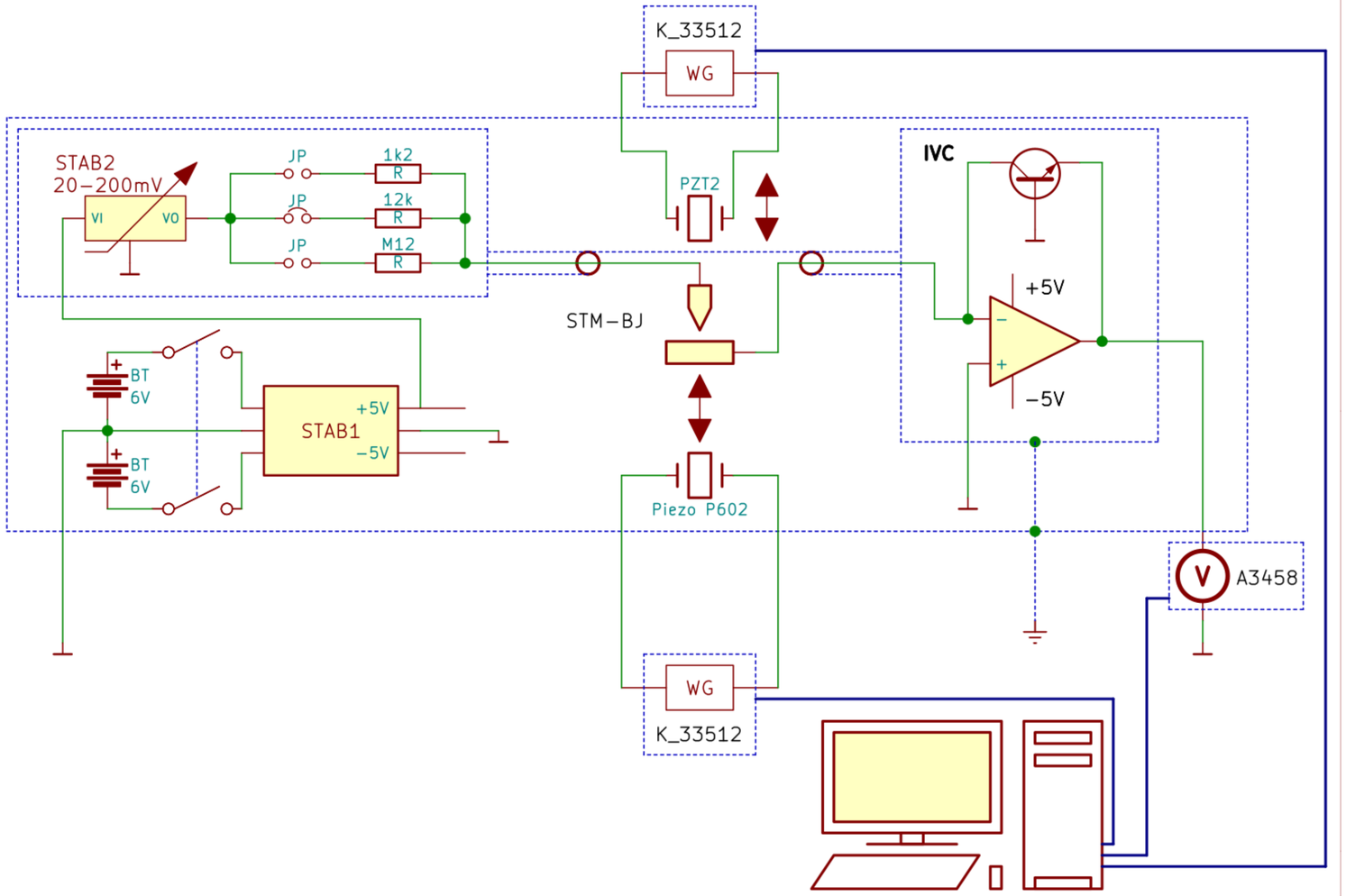
Auto-encoder II

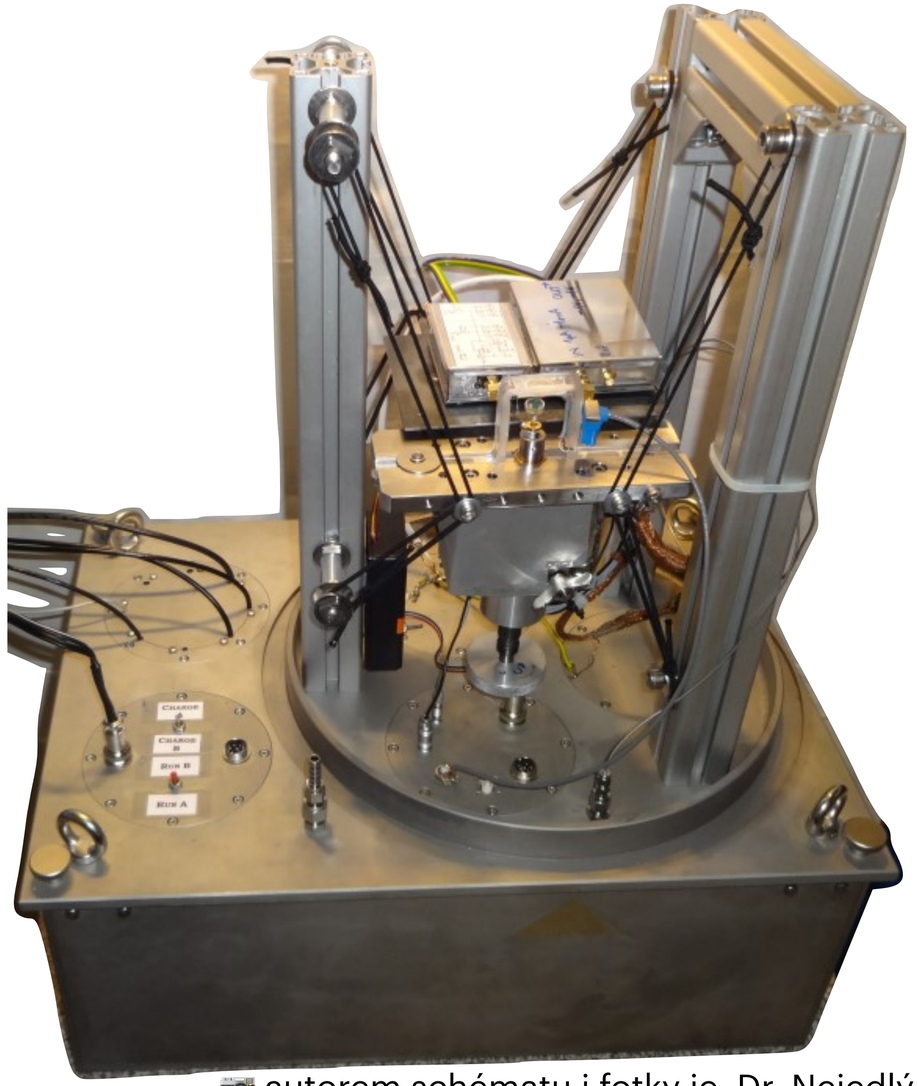
Výsledek





Aparatura





autorem schématu i fotky je Dr. Nejedlý

Reprezentace a embedding prostor

- 256-dim vektory promítnuty do RGB prostoru.
- Bulk patches, Plateau patches a Tunnelling patches se shlukují k sobě.