Extrakce příznaků z křivek experimentu break-junction metodou self-supervised learningu pro následné klastrování

Self-supervised feature extraction from break junction traces for enhanced clustering

Bc. Oliver Klimt — obhajoba diplomové práce, 15. 6. 2026

FEL ČVUT v Praze, katedra měření · vedoucí: Ing. Ladislav Sieger, CSc.
konzultanti (ÚOCHB): RNDr. Jaroslav Vacek, Ph.D. · RNDr. Jindřich Nejedlý, Ph.D.
icluto.oklimt.com · archiv příloh: thesis.icluto.oklimt.com

Problém a přínos

Experimenty break-junction (BJ) generují statisíce stochastických vodivostních křivek, v nichž jsou molekulární události vzácné a obtížně izolovatelné. Dosavadní metody buď agregují data do histogramů (a zahazují informaci na úrovni jednotlivých křivek), nebo spoléhají na expertně laděné prahy vodivosti, případně vyžadují nákladné anotace. Tato práce adaptuje **DINO** — self-supervised vision transformer trénovaný metodou self-distillation — na 1D vodivostní křivky. Model se učí embeddingy jednotlivých patchů přímo z neoznačených dat; nepotřebuje žádnou kalibraci rozsahu vodivosti ani anotované příklady. Vyhledávání kosinovou podobností v embedding prostoru čistě odděluje režimy bulk / molekulární spoj / tunelovací proud a umožňuje vyhledávání molekulárních kandidátů bez anotací — a to i napříč přístroji: na cizím datasetu bp4k funguje model bez přetrénování.

Klíčová čísla

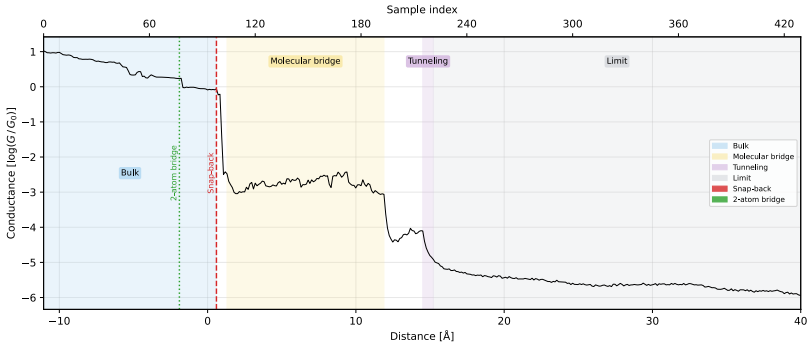
- **Data: 400 000+** trénovacích křivek (JIN206/466/467/536/537 v MesH); validace **41 208** křivek (R296, plató 13 Å).
- **Trénink:** 4 konfigurace × **~5 dní** (NVIDIA L40S, HPCC); finální model **patch 8 / dim 256**, epocha 30.
- **Similarity search: 9 935 / 41 208** křivek (**24,1 %**) při $S \geq 0,80$; celý průchod 652,8 s ⟶ **15,8 ms/křivka** na běžném notebooku.
- **Cross-instrument (bp4k, Kodaň, bez přetrénování): F1 ≈ 0,78**, accuracy ≈ 0,82 (práh 0,80); precision **0,99** (práh 0,95).

Co je v práci (mapa handoutu)

Experiment a data

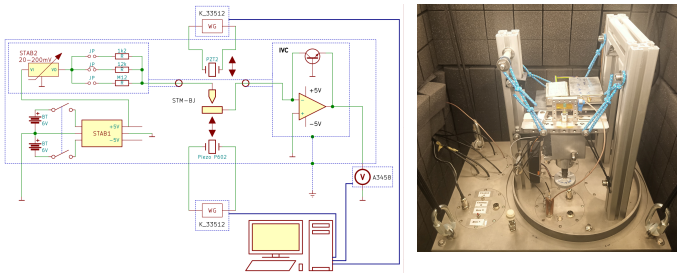
Motivace. Molekulární elektronika staví obvody „zdola nahoru„ z jednotlivých molekul (vodič, dioda, tranzistor) — alternativa k miniaturizaci křemíku narážející na atomární limity. Základní technikou je měření vodivosti jedné molekuly v break-junction experimentu: zlatý kontakt se opakovaně vytváří a trhá; při zúžení na jediný atom je vodivost kvantována v násobcích kvanta vodivosti $G_0 = 2e^2/h$ (e — náboj elektronu, h — Planckova konstanta). Po prasknutí kontaktu může mezeru přemostit molekula — vznikne nízkovodivostní plató, jehož statistika napříč tisíci křivkami určuje nejpravděpodobnější molekulární vodivost.

Anatomie křivky (Obr. 1): bulk (kovový kontakt) → plató N-atomového můstku → **snap-back** (prasknutí posledního atomového můstku) → **molekulární plató** → **tunelovací proud** (exponenciální pokles) → limit (šum aparatury, 10^{-6} –10^{-8} G/G_0).

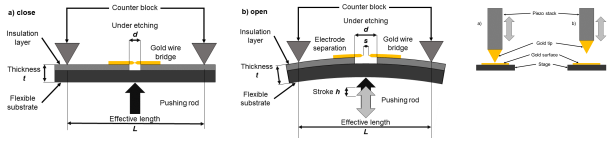


Obr. 1 — Anotovaná vodivostní křivka s vyznačenými oblastmi: bulk, 2-atomový můstek, snap-back, molekulární plató, tunelování (DP, obr. 2.1).

Aparatura. Komerční zařízení neexistuje; měření probíhá na zakázkové aparatuře (návrh prof. Zicha, ČVUT; J. Miletín, BMD) na ÚOCHB, podporující konfigurace STM-BJ i MCBJ se zlatými elektrodami (Obr. 2). Citlivost **100 fA**, vzorkování až **200 kSa/s**, logaritmický převodník proud–napětí (IVC), multimetr Agilent 3458, řízení v LabView; piezokrystal řízen generátorem Keysight 33512B.

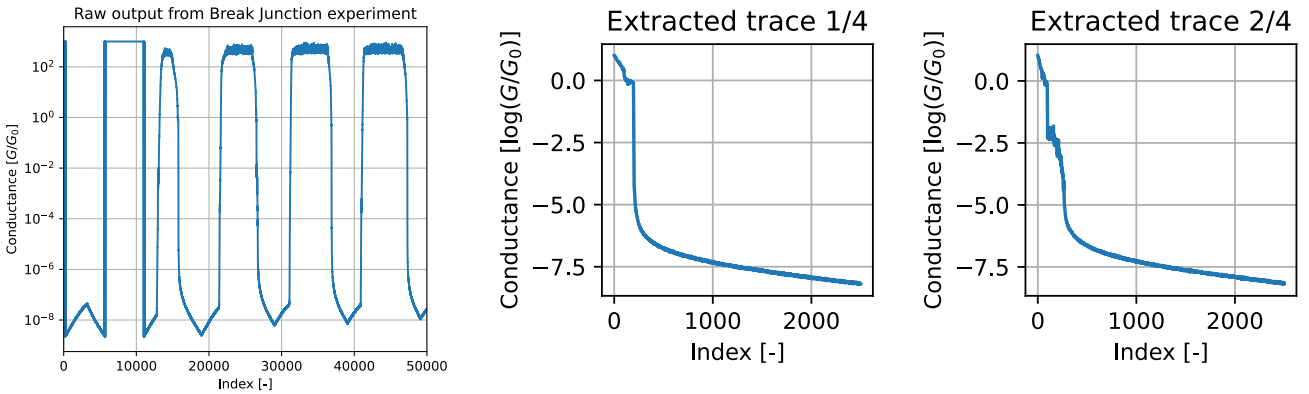


Obr. 2 — Schéma zapojení a fotografie zakázkové aparatury na ÚOCHB (DP, obr. 2.2).



Obr. 3 — Princip MCBJ (sevřený a rozevřený spoj) vs. STM-BJ (vpravo) (DP, obr. 2.3).

Extrakce křivek (Obr. 4). Záznam z aparatury je souvislý proud vzorků z tisíců cyklů (pilový pohyb piezo). Jednotlivé křivky se extrahují ve dvou krocích: (1) naučený detektor snap-backu + Non-Maximum Suppression najde charakteristický prudký pokles; (2) začátek křivky se ukotví v bodě $\log(G/G_0) = 1$. Výsledkem je segment fixní délky **2 500 bodů** obsahující bulk, atomový režim, snap-back a molekulární/tunelovací oblast.



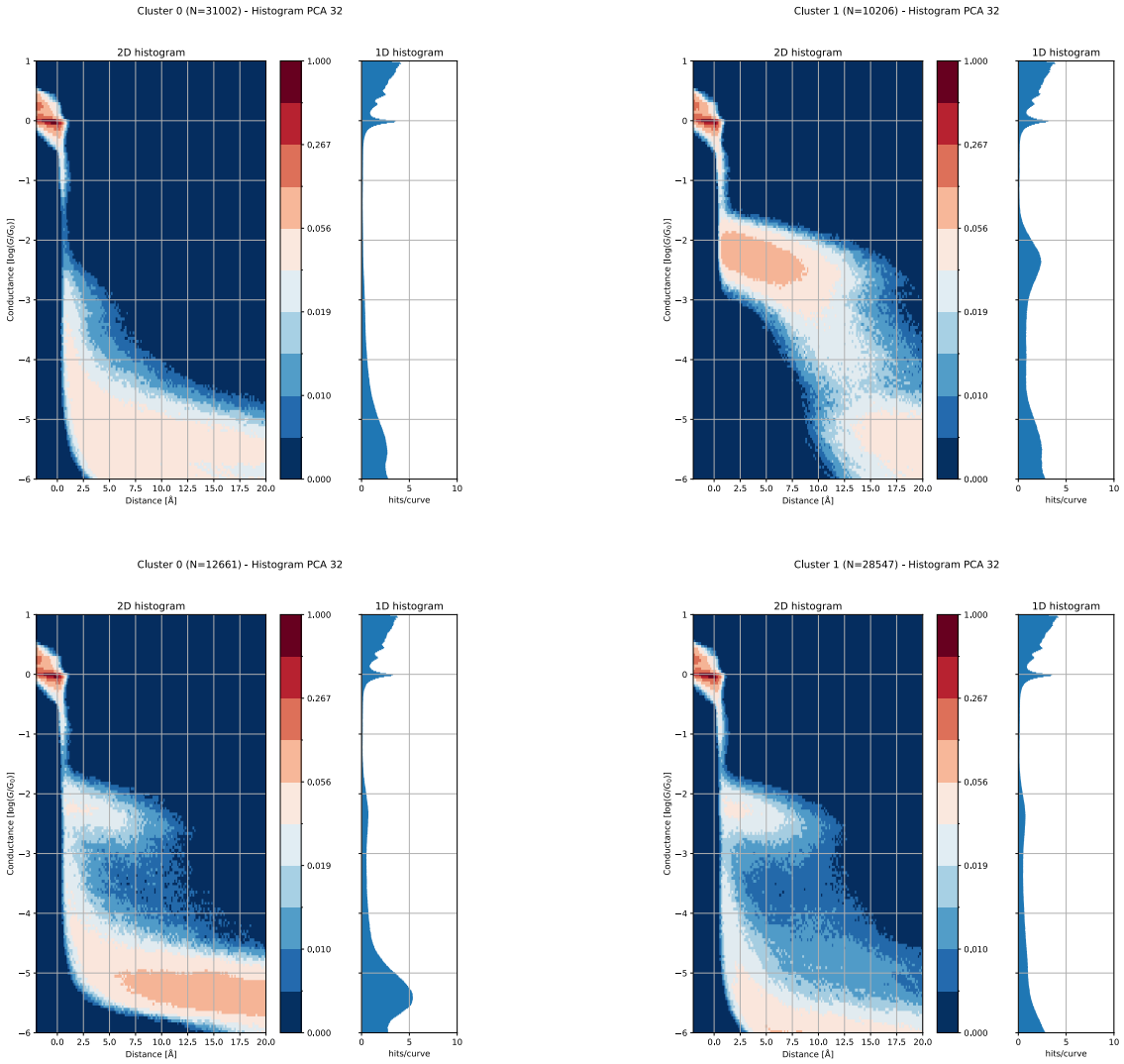
Obr. 4 — Surový datový proud (vlevo) a příklady extrahovaných křivek (DP, obr. 2.4).

Datasety. Trénink: 400 000+ křivek pěti nově syntetizovaných molekul (JIN206, JIN466, JIN467, JIN536, JIN537, Dr. Nejedlý) v mesitylenu (MesH — pomalé odpařování, stabilní prostředí). Validace: **41 208** křivek molekuly R296 (Ing. Seidler) — jednoduchá struktura a známý posuv (13 Å) z ní činí vhodný referenční dataset. SSL silně závisí na rozložení trénovacích dat: dominují-li irelevantní rysy, reprezentace nezachytí molekulární signál.

Motivace: limity iCluto v1

Pipeline v1 (Klimt 2024): filtrace křivek → vodivostní histogram jako vektor příznaků → PCA → K-means/DBSCAN. Rozlišila molekulární vs. tunelovací křivky a hrubé rozdíly ve vazbě molekula–elektroda, ale **selhává při zarušeném tunelovacím segmentu**, její schopnosti rozlišit další stavy je značně omezená.

Citlivost na rozsah vodivosti (Obr. 5). Při pečlivě zvoleném rozsahu 10^{-4}–10^{-1} G/G_0 klastry čistě oddělily molekulu od blanku (horní dvojice); s výchozím rozsahem 10^{-6}–10^{-1} však klastrování degradovalo — oba klastry ovládla limitní oblast (dolní dvojice). Rozlišení tedy stálo na zvolených mezích, ne na rysech naučených z dat.



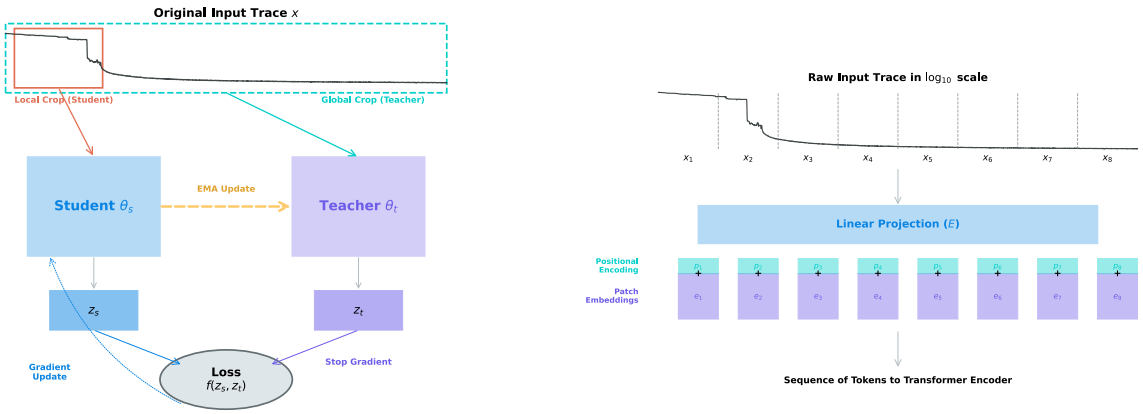
Obr. 5 — Křehkost v1 — tentýž algoritmus, jiný rozsah vodivosti. Nahoře: laděný rozsah 10^{-4}–10^{-1} G/G_0 odděluje blank (vlevo) od molekuly (vpravo) (DP, obr. 1.2). Dole: výchozí rozsah 10^{-6}–10^{-1} klastrování zhroutí — oběma klastrům dominuje limitní oblast (DP, obr. 1.3).

Autoencoder alternativa (DP, §3.1). Autoencoder či MAE musí uchovat informaci o celém vstupu — jenže bulk a limit tvoří ~80 % každé křivky, takže rekonstrukci dominují právě tyto neinformativní segmenty a molekulární plató přispívá příliš málo. K tomu proměnlivá délka plat (4–13 Å) špatně sedí na vektory fixní délky. Experimenty s autoenkodéry proto vedly k neuspokojivým výsledkům a směr byl opuštěn ve prospěch self-distillation.

Metoda: DINO pro 1D křivky

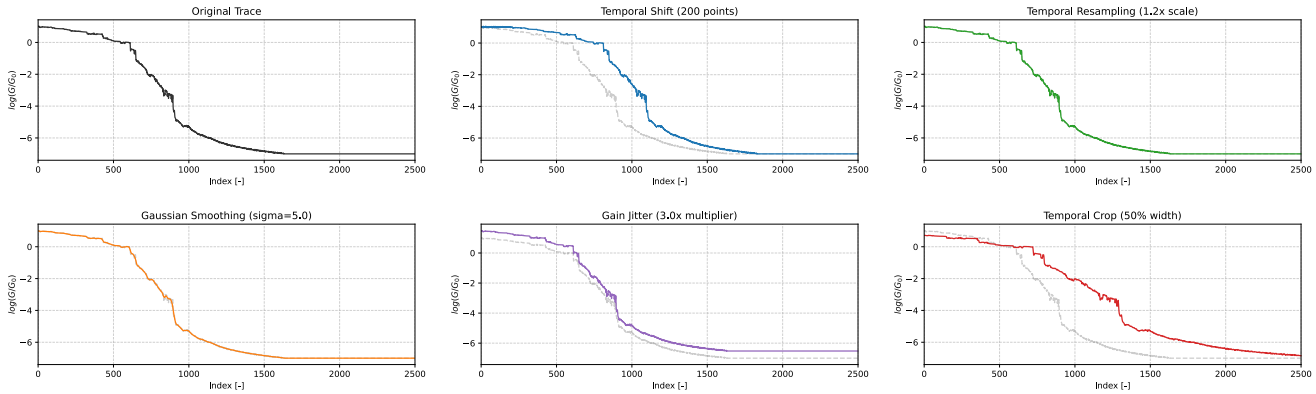
Tři rodiny SSL. *Rekonstrukční* (autoenkodér, MAE) — reprodukce vstupu; u BJ křivek dominuje bulk/limit (str. 3). *Kontrastivní* (SimCLR) — přitahuje augmentované pohledy téhož vzorku, odpuzuje ostatní; vyžaduje explicitní negativní páry. *Self-distillation* (DINO) — student se učí predikovat výstup pomalu se vyvíjejícího učitele; žádné páry ani labely. Pro cíl práce — popis **jednotlivých segmentů** křivky — je self-distillation nejvhodnější.

Architektura (Obr. 6). Student i učitel sdílejí stejnou strukturu: Conv1d patch embedding (kernel 8, stride 4 ⟶ **50% překryv** sousedních patchů) převede křivku na sekvenci tokenů; přidá se učitelný CLS token a učitelné poziční embeddingy; Transformer backbone s Multi-Head Self-Attention zachytí závislosti (např. korelaci délky plata se sklonem snap-backu); DINOHead promítá CLS reprezentaci na distribuci přes **1 024 prototypů**, na níž se počítá DINO loss (cross-entropy student vs. učitel). Učitel se neaktualizuje backpropagací, ale exponenciálním klouzavým průměrem (EMA) vah studenta. **Role CLS tokenu je omezena na trénink** — downstream analýzy (BoVW, similarity search, segmentace) pracují výhradně s per-patch embeddingy z výstupu enkodéru; CLS se zahazuje. Self-attention ale zajišťuje, že každý patch embedding nese týž kontext, který formoval CLS reprezentaci.



Obr. 6 — Vlevo: schéma DINO — student se učí z lokálních výřezů, učitel (EMA studenta) dává cíle z globálních; loss porovnává jejich distribuce nad prototypy (DP, obr. 4.1). Vpravo: tokenizace 1D křivky — Conv1d patch embedding, CLS token, poziční embeddingy (DP, obr. 4.2).

Augmentace (Obr. 7) — každá simuluje konkrétní instrumentální varianci: **horizontální posun** (jitter startu měření a piezo aktuátoru); **resampling** (variace rychlosti piezo a vzorkovací frekvence); **vyhlazení** Gaussem (potlačení vysokofrekvenčního šumu, důraz na makrostrukturu); **vertikální gain** (invariance vůči absolutní úrovni vodivosti — vazba molekuly, zesílení); **multi-crop** (globální pohledy 80–100 % křivky, lokální 20–40 %; student předpovídá globální reprezentaci z lokálních výřezů).



Obr. 7 — Augmentační pipeline pro 1D vodivostní křivky: originál, posun, resampling, vyhlazení, vertikální gain, multi-crop (DP, obr. 3.2).

Prevence kolapsu reprezentace. Degenerovaný stav „vše na stejný výstup„ triviálně splňuje DINO loss. Řešení: **centering** (odečtení průměru výstupů učitele — žádný prototyp nesmí saturovat) + **sharpening** (nižší teplota učitele ⟶ ostřejší cíle). Navíc pojistka **„Stability First“**: klesne-li loss pod 10^{-6} po 5 po sobě jdoucích batchích, teplotní schedule učitele se přitlumí faktorem 0,9.

Trénink a výběr modelu

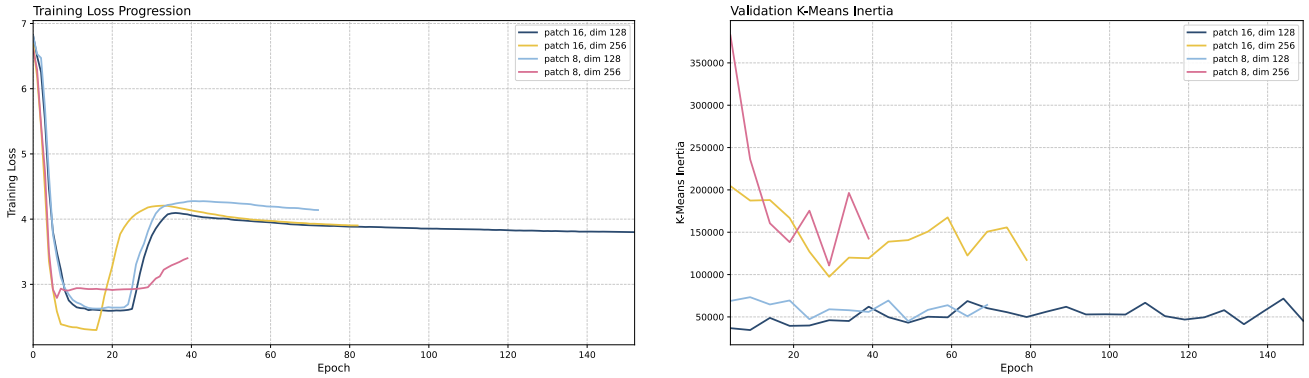
Na základě krátkých pretrénovacích běhů (15 kombinací) byly plně trénovány 4 konfigurace: patch {8, 16} × dim {128, 256} (Tab. 1). Velikost patche počet parametrů téměř neovlivňuje; dominuje embedding dimenze (backbone 3,2–3,3 M při dim 256 vs. 0,84–0,87 M při dim 128; projekční hlava ≈ 0,7–0,8 M). Trénink: HPCC, nodes s NVIDIA L40S (48 GB) GPU, 2× AMD EPYC 9654; každý běh ≈ 5 dní; batch 32.

Konfigurace	Parametry	Finální loss	Min. loss	Epocha min.	Trvání (dny)
Patch 16, dim 128	≈ 1,56 M	3,8004	2,5942	19	4,850
Patch 16, dim 256	≈ 4,03 M	3,9032	2,2992	16	4,847
Patch 8, dim 128	≈ 1,60 M	4,1394	2,6277	16	4,849
Patch 8, dim 256	≈ 4,11 M	3,4008	2,7912	6	4,848

Tab. 1 — Počty parametrů a souhrn tréninkových běhů čtyř konfigurací (zkráceno; DP, tab. 4.1 a 4.2).

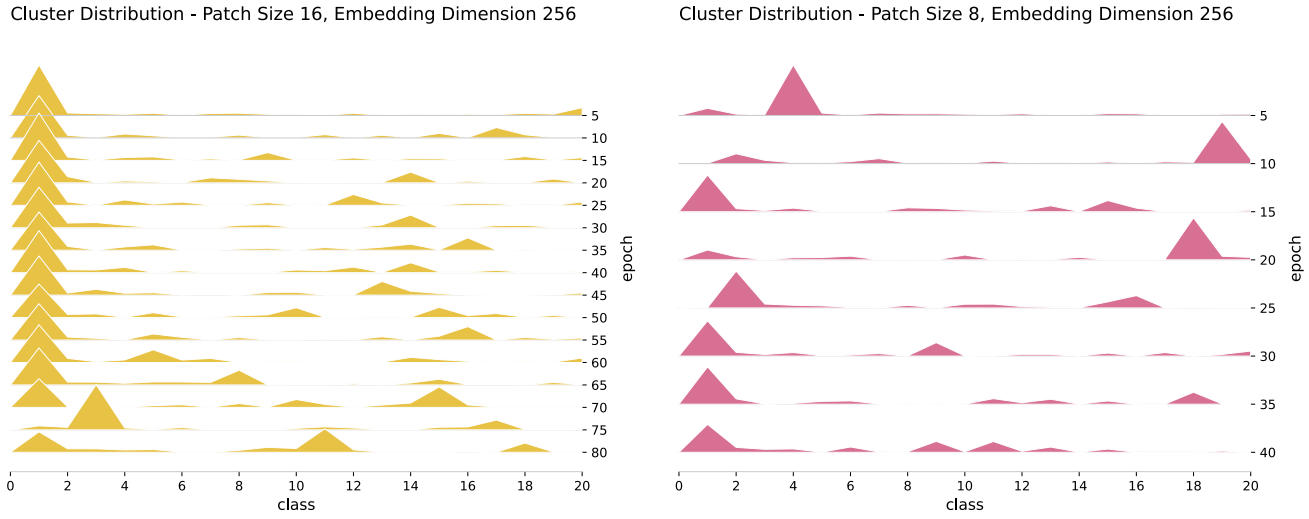
Proč DINO loss nevypovídá o kvalitě modelu (Obr. 8 vlevo). Loss u všech konfigurací prudce klesne během prvních epoch (minimum v ep. 6–19) a poté **roste** — primárně kvůli teplotnímu schedule učitele (lineárně 0,04 ⟶ 0,07 přes prvních 30 epoch), který změkčuje cílové distribuce a zvyšuje cross-entropy i při zlepšujících se reprezentacích; současně kosinový decay learning rate zpomaluje korekce studenta. Loss je tedy ovlivněn schedulem a **nelze ho použít k výběru modelu**.

Výběr přes validační K-means inertia (K=20) na patch embeddings učitele (Obr. 8 vpravo): velký pokles inertia = embeddingy se organizují do kompaktních skupin. 256-dim modely klesají výrazně (patch 8/dim 256: ≈ 375 000 ⟶ 110 000 za prvních 25 epoch), 128-dim zůstávají ploché — nižší dimenze nemá dostatečnou kapacitu na strukturovaný latentní prostor.



Obr. 8 — Vlevo: DINO loss — rychlý pokles, poté růst řízený teplotním schedulem (DP, obr. 4.5). Vpravo: validační K-means inertia (K=20) — výběrové kritérium (DP, obr. 4.6).

Waterfall grafy odhalují kolaps (Obr. 9). Model patch 16/dim 256 vykazuje **fixní dominantní prototyp** napříč epochami (kompaktní, ale degenerované řešení — navzdory klesající inertii). U patch 8/dim 256 dominantní třída mezi epochami **rotuje** — slovník se aktivně reorganizuje. Zvolen byl proto **patch 8 / dim 256, epocha 30**.



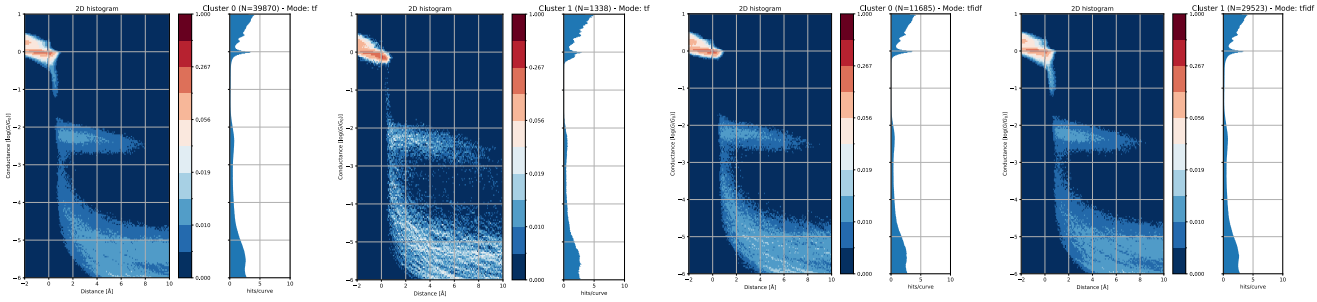
Obr. 9 — Vývoj distribuce patch tříd přes epochy. Vlevo patch 16/dim 256: známky kolapsu (perzistentní dominantní třída). Vpravo patch 8/dim 256: dominantní třída rotuje ⟶ zvolený model (DP, obr. 4.7).

Bag of Visual Words

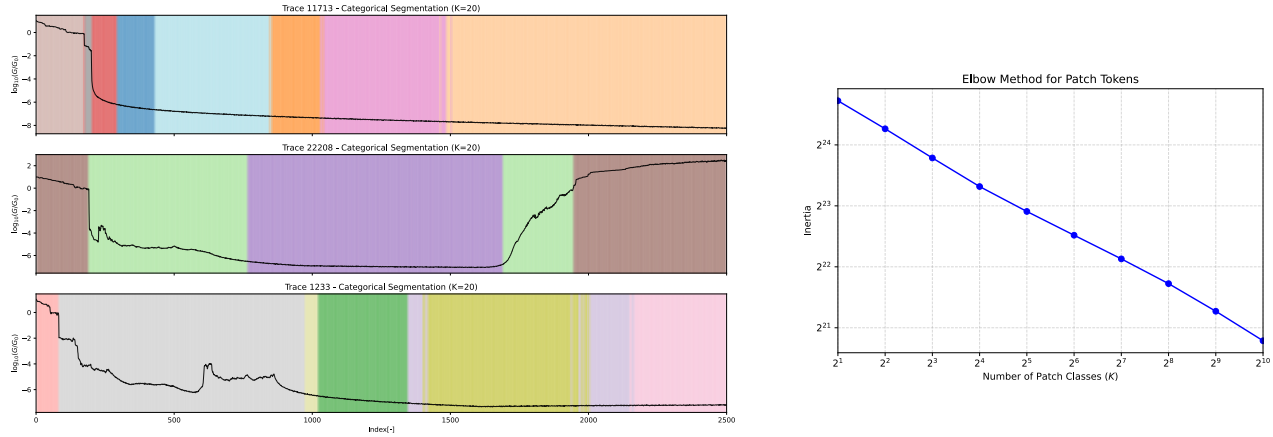
Zadání práce předpokládá klastrování s velkým K; BoVW byl přímou adaptací osvědčeného postupu z computer vision (SIFT → slovník → histogram „vizuálních slov„). Výsledek je negativní, ale diagnosticky cenný — **selhala agregace, nikoli embeddingy**.

Pipeline. Slovník 1 024 vizuálních slov K-means klastrováním náhodného vzorku patchů (celý trénink = 244 M patchů ≈ 250 GB RAM — nutný subsampling; i samotná validace ≈ 26 GB). Každá křivka → histogram četností slov; vážení TF vs. TF-IDF (IDF $= \log(N/(1 + \mathrm{DF}))$ potlačuje všudypřítomná slova, zesiluje vzácná), L_2 normalizace, globální K-means.

Výsledky. (1) TF a TF-IDF dávají **téměř identické** klastry (Obr. 10) — očekávané zvýhodnění vzácných slov se nekoná. (2) WCSS křivka **nemá loket** — klesá téměř lineárně v celém testovaném rozsahu (Obr. 11 vpravo): patch embeddingy jsou rozloženy hustě a spojitě, bez kompaktní klastrové struktury, takže žádná velikost slovníku není „přirozená„. (3) Per-trace segmentace je souvislá (koherentní) — bulk, plató, snap-back i tunelování vycházejí jako souvislé bloky (Obr. 11 vlevo) — ale slovník **nedrží identitu tříd napříč křivkami**: bulk každé křivky dostane jiné vizuální slovo, takže dvě křivky se stejnou fyzikální událostí sdílejí jen málo slov.



Obr. 10 — TF (dvojice vlevo) vs. TF-IDF (dvojice vpravo), $k = 2$: téměř identická přiřazení klastrů (DP, obr. 5.1).

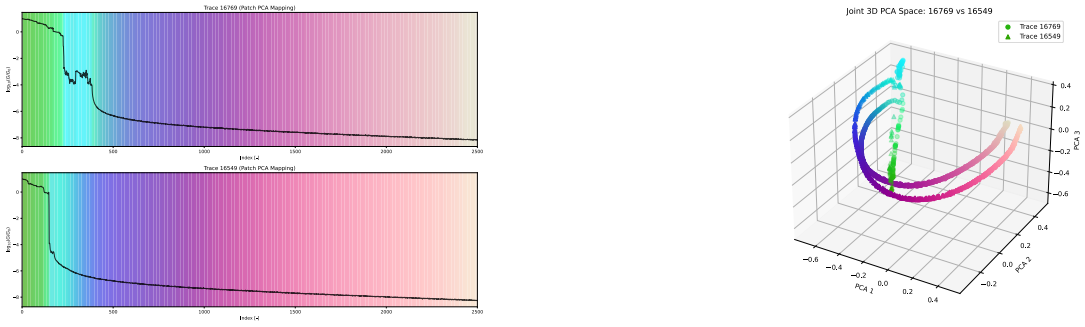


Obr. 11 — Vlevo: K-means segmentace patchů uvnitř křivek — koherentní bloky, ale třídy nesdílené napříč křivkami (DP, obr. 5.2). Vpravo: WCSS bez lokte — žádná přirozená velikost slovníku (DP, obr. 5.3).

Závěr: zjištění o agregaci, ne o embeddings — PCA projekce, self-similarity matice i podobnostní profily (str. 7–8) potvrzují, že embedding prostor fyzikální strukturu nese.

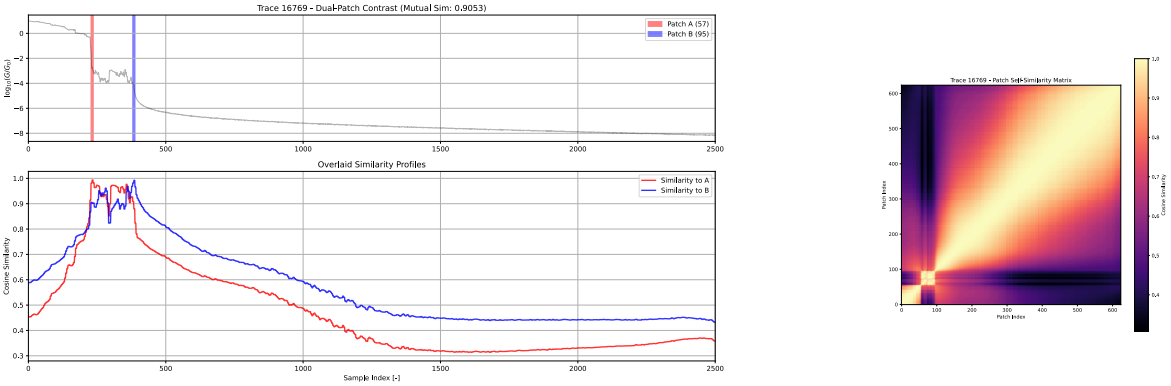
Embedding prostor a similarity search (hlavní výsledek)

Struktura embedding prostoru (Obr. 12). Projekce všech patch embeddingů na první 3 hlavní komponenty (mapované na RGB kanály) ukazuje, že fyzikálně odlišné segmenty — bulk, molekulární plató, tunelovací proud — obsazují **geometricky oddělené oblasti**. Embedding prostor kóduje fyziku měření, i když jej reprezentace pomocí histogramů nedokáže využít.



Obr. 12 — PCA/RGB visualisace patch embeddingů (vlevo) a společný 3D PCA prostor (vpravo): segmenty tvoří oddělené oblasti (DP, obr. 5.4).

Vnitřní kontrast a „similarity square„ (Obr. 13). Kosinová podobnost mezi patchi téže křivky ukazuje ostrý kontrast mezi segmenty; self-similarity matice zviditelňuje molekulární spoj jako koherentní čtverec na diagonále (před patchem 100).

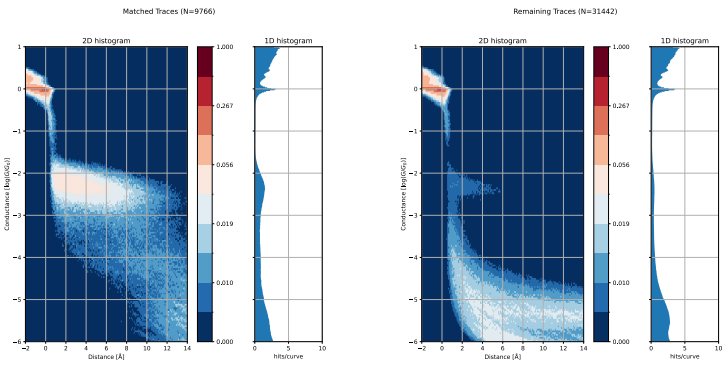


Obr. 13 — Vlevo: kontrast dvou referenčních patchů (57 vs. 95) v křivce 16769 (DP, obr. 5.5). Vpravo: self-similarity matice téže křivky — „similarity square„ = molekulární spoj (DP, obr. 5.6).

Vyhledávací pravidlo. Referenční patch = **patch 85 křivky 16769** (uvnitř molekulárního plata). Patch je shoda, pokud kosinová podobnost ≥ práh; křivka je „Matched„, obsahuje-li ≥ **10** shodných patchů. Průchod je streamovaný/batched — 25 M patchů validace by v RAM zabralo ≈ 26 GB, plná matice se nikdy nedrží v paměti. Metoda je **machine-agnostic**: pracuje s morfologií příznaků, ne s absolutním rozsahem vodivosti.

Práh S	Shodných	Podíl
≥ 0,70	18 043	43,8 %
≥ 0,75	13 233	32,1 %
≥ 0,80	**9 935**	**24,1 %**
≥ 0,85	7 074	17,2 %
≥ 0,90	4 152	10,1 %
≥ 0,95	1 529	3,7 %

Tab. 2 — Citlivost na práh, plné znění (křivka 16769, patch 85, MIN_PATCHES = 10, N = 41 208; DP, tab. 5.1).

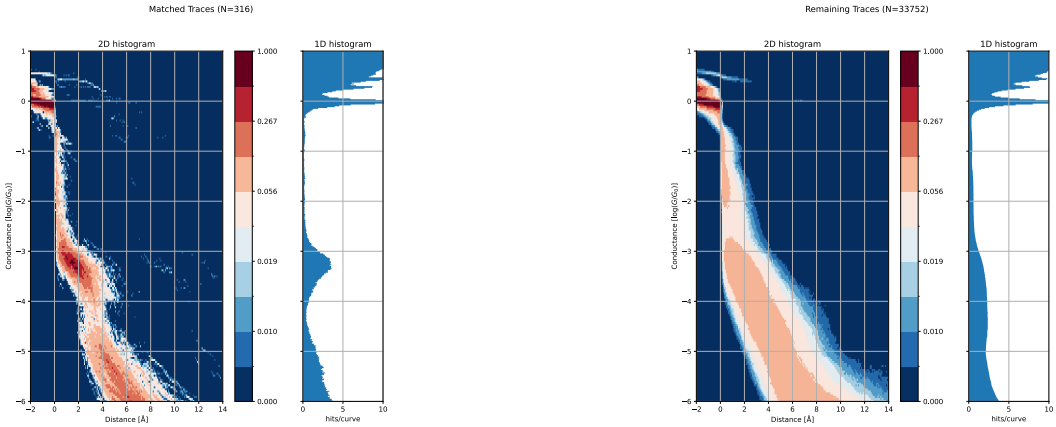


Obr. 14 — 2D histogramy: shodné křivky (vlevo) vs. zbytek validačního datasetu (vpravo) (DP, obr. 5.7).

Monotónní pokles 43,8 % ⟶ 3,7 % (Tab. 2) potvrzuje precision–recall trade-off; celý průchod trval 652,8 s, tj. **15,8 ms/křivka** na notebooku (Apple Silicon). Shodné křivky (Obr. 14 vlevo) jsou prosté čistě tunelovacích křivek — vysoká precision; ve zbytku (vpravo) zůstávají některé molekulární signatury — recall jediné kotvy je neúplný: jeden referenční patch nepokryje širokou oblast molekulárních rysů a přísné minimum 10 patchů odfiltruje i část platných událostí (řešení: multi-anchor search, str. 10).

Specificita na jednom měření a segmentace křivek

„Obtížný dataset„ 1954 (Obr. 15). Stejné pravidlo aplikované na jediné měření izoluje molekulární signaturu od tunelovacího pozadí (**316 z 34 068** křivek). Zbytek může obsahovat další signatury či artefakty — postup lze aplikovat iterativně.

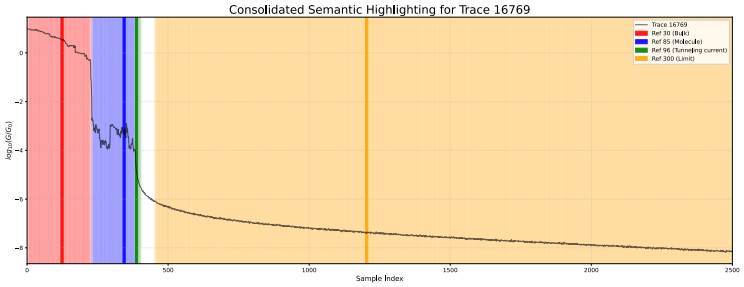


Obr. 15 — Dataset 1954: shodné křivky (vlevo) vs. zbytek (vpravo) — specificita similarity search na úrovni jednotlivého měření (DP, obr. 5.8).

Segmentace přes referenční patche (Tab. 3, Obr. 16). Porovnáním embeddingů proti referenčním patchům s prahem specifickým pro třídu lze křivku sémanticky segmentovat — bez jediné anotace. Bulk a limit lze ověřit triviálně (log-vodivost nad 0, resp. pod −6) — model je rozlišuje správně; snap-back patche se v prostoru shlukují s minimem odlehlých bodů (DP, obr. 5.10).

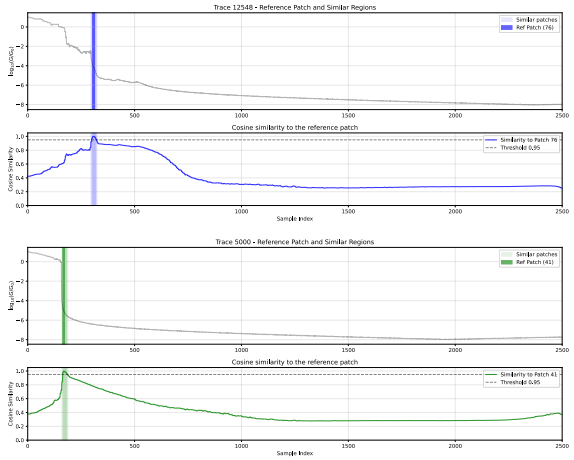
Segment	Práh S	Ref. patch
Bulk	≥ 0,80	30
Molekulární signatura	≥ 0,90	85
Tunelovací proud	≥ 0,95	96
Limit	≥ 0,65	300

Tab. 3 — Prahy a referenční patche segmentace (DP, tab. 5.2).



Obr. 16 — Konsolidovaná sémantická segmentace: bulk, molekulární signatura, tunelování a limit identifikovány přes podobnost embeddingů (DP, obr. 5.9).

Identifikace tunelovacího proudu (Obr. 17). Klíčový segment pro výpočet posuvu elektrod (indifaktor, str. 10). Similarity search jej spolehlivě nachází **navzdory rozsahu pouhých ≈ 24 datových bodů** (několik překrývajících se patchů) — a to jak v křivce s molekulou, tak v blank křivce. Nabízí se tak annotation-free alternativa k supervizovanému U-Netu používanému dosud; kvantitativní srovnání proti anotované referenci zatím chybí.



Obr. 17 — Identifikace tunelovacího proudu podobnostním profilem: křivka 12548 s molekulou (nahoře) a blank křivka 5000 (dole); panely (a) a (c) z DP, obr. 5.11.

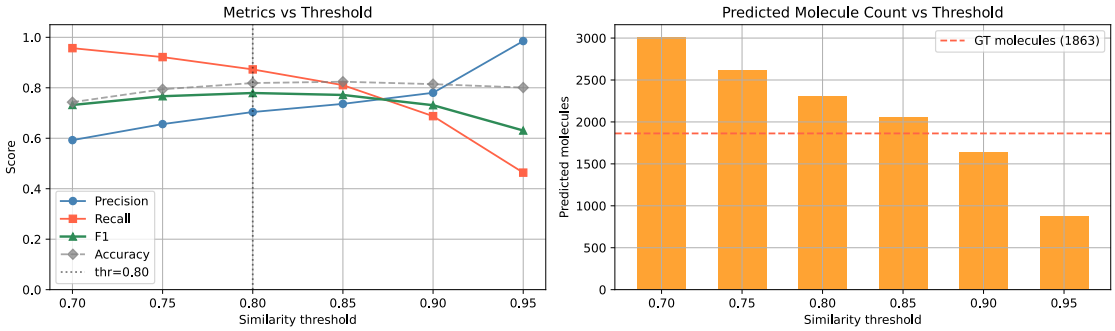
Cross-instrument validace (bp4k)

bp4k. Veřejný dataset bp4k (Bro-Jørgensen et al., Univ. Kodaň): **5 475 křivek** × 4 229 bodů, ground-truth labely **1 863** Molecule / **3 219** Background / 393 Noise (Noise z metrik vyloučen, N = 5 082). Zásadní odlišnosti od ÚOCHB dat: měření při **4 K** (vs. pokojová teplota), **neznámá vzorkovací frekvence**, jiné okno měření — model nikdy neviděl ani přístroj, ani molekulu.

Adaptace: pouze resampling, žádný finetuning. Křivky sjednoceny za limitem vodivosti, zarovnány na nástup bulk kontaktu, molekulární můstek přeškálován na **180 patchů (724 vzorků)** odpovídajících efektivnímu rozlišení modelu, doplněno na 2 500 bodů. Klasifikační pravidlo identické se str. 7: referenční patch (křivka 2, patch 145, label Molecule), ≥ 10 shodných patchů.

Práh	#Pred	Precision	Recall	F1	Accuracy
0,70	3 009	0,59	0,96	0,73	0,74
0,75	2 617	0,66	0,92	0,77	0,79
0,80	2 310	0,70	0,87	**0,78**	**0,82**
0,85	2 051	0,74	0,81	0,77	0,82
0,90	1 642	0,78	0,69	0,73	0,81
0,95	877	**0,99**	0,46	0,63	0,80

Tab. 4 — Citlivost na práh na bp4k, plné znění (referenční patch: křivka 2, patch 145; MIN_PATCHES = 10; N = 5 082 labelovaných Molecule/Background křivek; DP, tab. 5.3).



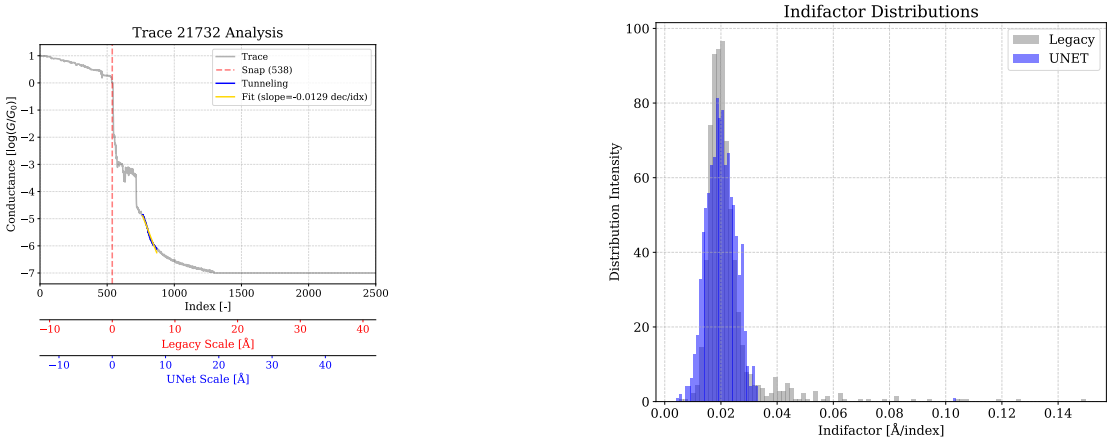
Obr. 18 — Metriky vs. práh a predikované počty na bp4k — bez finetuningu (DP, obr. 5.13).

Tři provozní body (Tab. 4): práh 0,70 najde **96 %** molekul — vhodný jako hrubý první filtr; práh 0,80 je vyvážený kompromis (**F1 0,78**, precision 0,70, recall 0,87); práh 0,95 nedává téměř žádné false positives (precision **0,99**, recall 0,46) — vhodný pro výběr čistých ukázkových křivek.

Interpretace. Jediný patch embedding z jedné molekulární křivky funguje jako **přenositelný prototyp napříč přístroji** — reprezentace kóduje fyziku molekulárního signálu, ne artefakty konkrétní aparatury. Spolu s PCA projekcemi a podobnostními profily to potvrzuje, že limitem BoVW pipeline nebyla kvalita embeddingů, ale histogramová agregace.

Indifaktor (Appendix C; POSTER 2026)

Posuv elektrod není měřen přímo — indifaktor je škálovací faktor převádějící indexy vzorků na ångströmy; je nezbytný pro korektní 2D histogramy. **Legacy vzorec** (DP, rov. C.1) odhaduje sklon ze dvou bodů křivky s empirickou konstantou $s = 1,28$. **Nový postup:** 1D U-Net segmentuje tunelovací proud, RANSAC robustně fituje jeho sklon (Obr. 19 vlevo); fyzikální základ — 1 dekáda poklesu ≈ 1 Å ve vakuu, v MesH empiricky **0,531 Å/dekádu** (kalibrováno proti legacy indifaktoru). **Výsledky:** směrodatná odchylka klesá z $\sigma \approx 0,0118$ (legacy) na $\sigma \approx 0,0057$ (U-Net) — **≈ 2× přesnější** (Obr. 19 vpravo); sklon je stabilní přes 30 000 křivek; U-Net potlačuje rostoucí trend chyby legacy metody; distribuce blank vs. 4,4′-bipyridin vykazují měřitelný posun — robustní kalibrace odliší jemné molekulární chování od baseline variací.



Obr. 19 — Vlevo: U-Net segmentace + RANSAC fit sklonu tunelovacího proudu (DP, obr. C.1). Vpravo: distribuce indifaktorů legacy vs. U-Net — ≈ 2× užší rozptyl (DP, obr. C.3).

Limity, budoucí práce, reprodukovatelnost

Limity.

- **Recall jediné reference:** molekulární rysy obsazují širokou oblast embedding prostoru; jediný referenční patch + minimum 10 shod část platných událostí vynechá.
- **Chybí segment-level ground truth:** evaluace na ÚOCHB datech je kvalitativní (2D histogramy, PCA); kvantitativní metriky jen pro bp4k (trace-level labely).
- **Fixní velikost patche** vs. disparita délek segmentů: bulk/limit = stovky bodů, plató/snap-back/tunelování = desítky ⟶ kompromis mezi redundancí a rozlišením hranic.

Budoucí práce (DP, §6.3).

- **Multi-anchor search** — sada komplementárních kotev ⟶ vyšší recall bez ztráty label-free charakteru.
- **Anotovaný benchmark:** ≥ 10 000 křivek se segment-level hranicemi — „ImageNet pro BJ„; dnes neexistuje.
- **Adaptivní velikost patche** podle naučeného detektoru.
- **DINO jako foundational model v iCluto** — fixní extraktor příznaků pro klasifikaci, klastrování, detekci anomálií.

Reprodukovatelnost (DP, Appendix B). Kompletní artefakty na thesis.icluto.oklimt.com: `icluto-0.1.10` wheel, váhy `dino_model_epoch30.pth` + metadata JSON, dataset `traces.npy`, notebooky `00–05` (model card, augmentace, segmentace, similarity search, BoVW, externí validace bp4k). Vyžaduje Python ≥ 3.11; instalace:

```
python -m venv .venv && source .venv/bin/activate && pip install icluto-0.1.10-py3-none-any.whl jupyter ipywidgets &&
jupyter lab notebooks/
```

Mini-glosář

G_0 kvantum vodivosti $2e^2/h$; jednotka, v níž se udává vodivost spoje.
snap-back prudký pokles vodivosti při prasknutí posledního atomového můstku.
molekulární plató nízkovodivostní plató — molekula přemosťuje mezeru mezi elektrodami.
tunelovací proud / indifaktor exponenciální pokles vodivosti s mezerou; jeho sklon kalibruje posuv elektrod (indifaktor).
patch / embedding krátký úsek křivky (8 vzorků, překryv 50 %) / jeho 256-dim vektorová reprezentace z modelu.
self-distillation student se učí predikovat výstup učitele, který je EMA kopií studenta; bez labelů (DINO).
CLS token agregační token — globální shrnutí křivky; používá se jen při tréninku.
BoVW Bag of Visual Words — histogram četností „vizuálních slov„ (prototypů patchů) na křivku.

Vybrané reference

Caron et al. (2021): *Emerging Properties in Self-Supervised Vision Transformers* (DINO). ICCV.
Oquab et al. (2023): *DINOv2: Learning Robust Visual Features without Supervision.*
Siméoni et al. (2025): *DINOv3.*
Bro-Jørgensen et al. (2022): *Trusting our machines* (dataset bp4k). Univ. Kodaň.
Klimt (2024): *iCluto v1* — bakalářská práce, PCA + K-means pipeline. ČVUT FEL.
Schubert (2023): *Stop using the elbow criterion for k-means.* SIGKDD Expl.
Xu, Tao (2003): *Measurement of Single-Molecule Resistance by Repeated Formation of Molecular Junctions* (STM-BJ). Science.