Detekcia modifikovaných báz v dátach platformy MinION

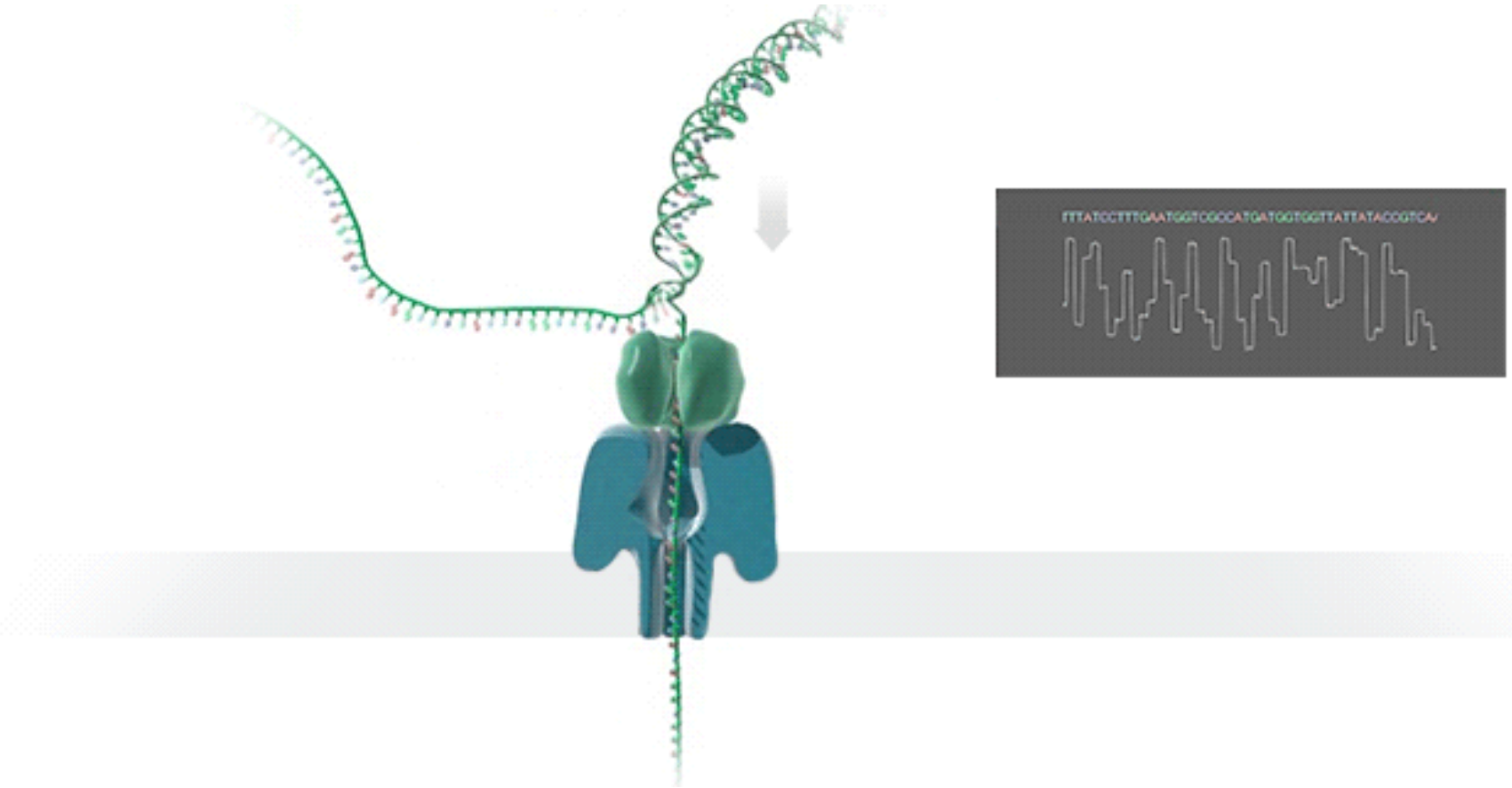
Rastislav Rabatin

Školiteľ: doc. Mgr. Tomáš Vinař, PhD.

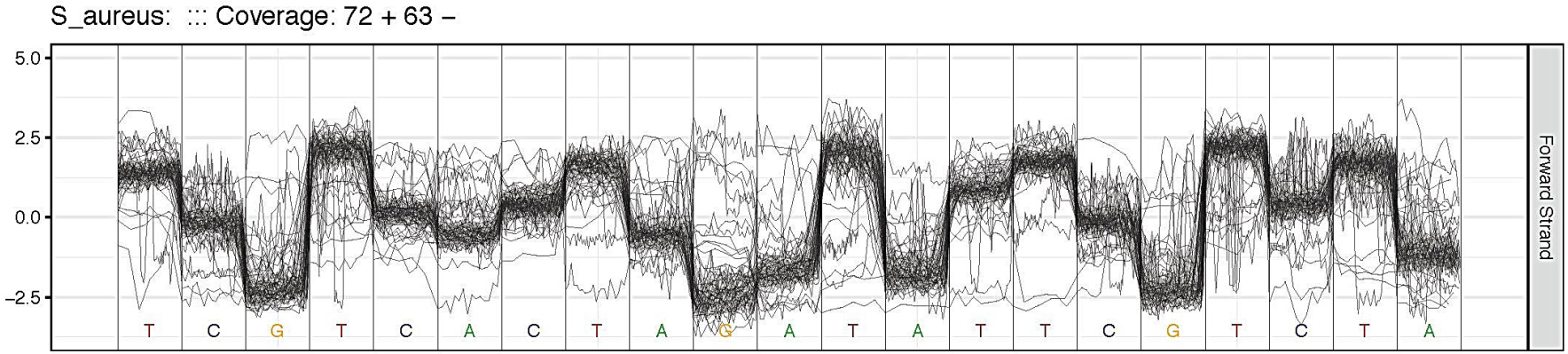
13. júna 2018

Čo je to MinION?

Nanopórové sekvenovanie



Surový signál

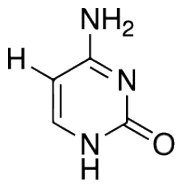


- Zašumený schodíkový signál.
- Úroveň signálu závisí na k-tici báz (kontext báz).
- Čas aký sa báza zrdží v nanopóre nie je konštantný.

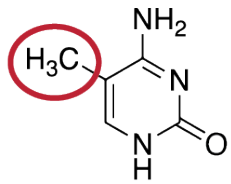
DNA modifikácie

- Genetické modifikácie – zmena bázy (C $\rightarrow$ G).
- Epigenetické modifikácie – neovplyvňujú genetický kód.
- Nás zaujímajú epigenetické modifikácie.
- Ovplyvňujú proces čítania DNA v živej bunke (napr. expresia génov).

DNA metylácia

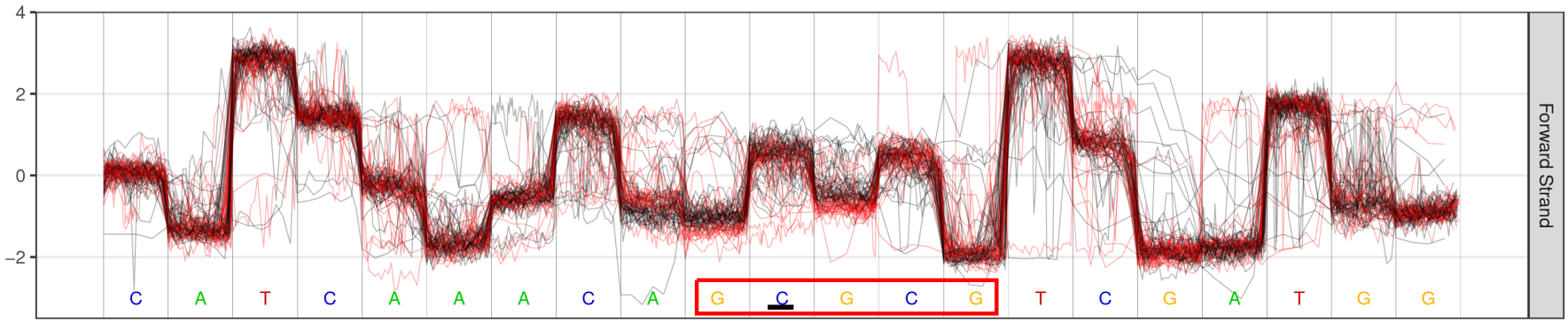


Cytosine



methylated Cytosine

Efekt metylácie na signál



Rôzne pohľady na problém – Klasifikácia

- Požiadavky na trénovacie dáta:
  - Obsahujú modifikované a nemodifikované bázy.
  - Vieme, ktoré bázy sú modifikované, a ktoré nie.
  - Každá modifikácia sa nachádza v každom možno kontexte báz.
- V prírode existuje veľa rôznych DNA modifikácií.
- Signál závisí na kontexte báz.
- Nie je jednoduché experimentálne vytvoriť takéto dáta.
- Predchádzajúce projekty: nanopolish (HMM), mCaller (MLP)

Rôzne pohľady na problém – Štatistické testy

- Žiaden model, žiadna trénovacia fáza.
- Porovnávame signál s modifikáciami a bez modifikácií.
- Potrebujeme vyššie pokrytie.
- Predchádzajúce projekty: nanoraw, tombo

Ciele našej práce

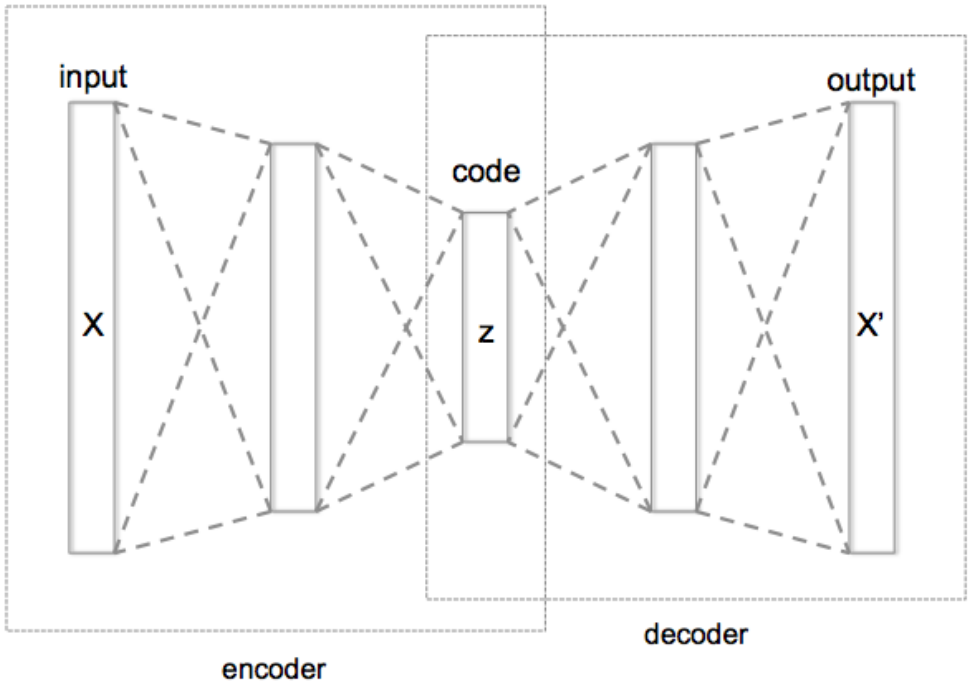
- Požiadavky na trénovacie dáta:
  - Dáta bez modifikácií.
  - Máme dostatočné pokrytie všetkých kontextov.
- Testovacia fáza: potrebujeme iba natrénovaný model.

Náš prístup

Detekcia anomálií

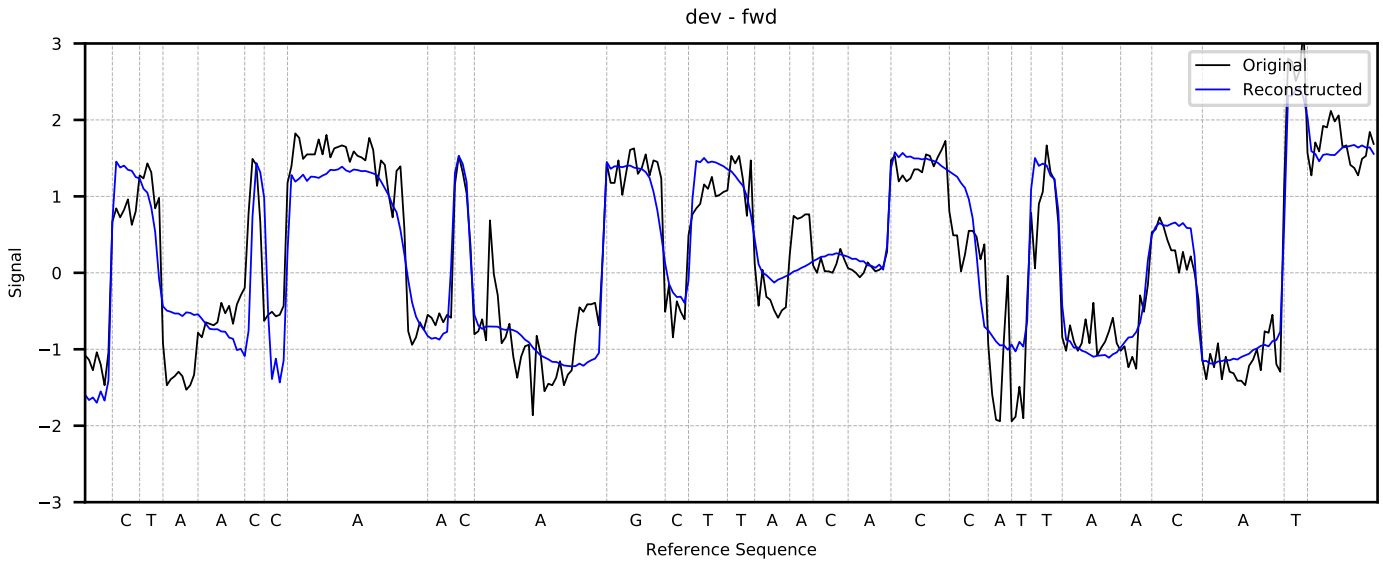
- Trénujeme iba na dátach bez modifikácií.
- Cieľ: model sa naučí charakteristiky dát bez modifikácií.
- Testovacia fáza: model vráti skóre anomálnosti.
- Výhoda: vieme objavovať nové DNA modifikácie.

Autoenkóder

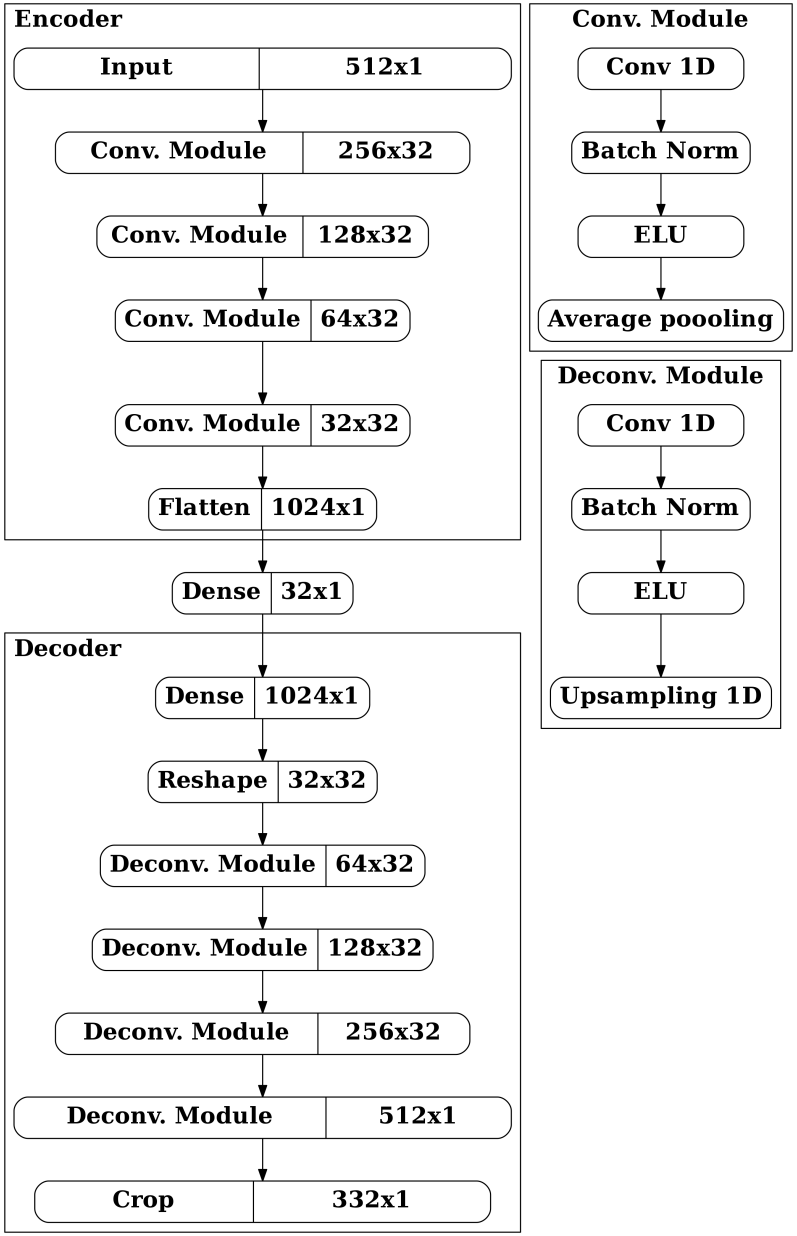


- Princíp: veľká rekonštrukčná chyba na dátach s modifikáciami
- Môžeme pridať aj kontext báz tomuto modelu.

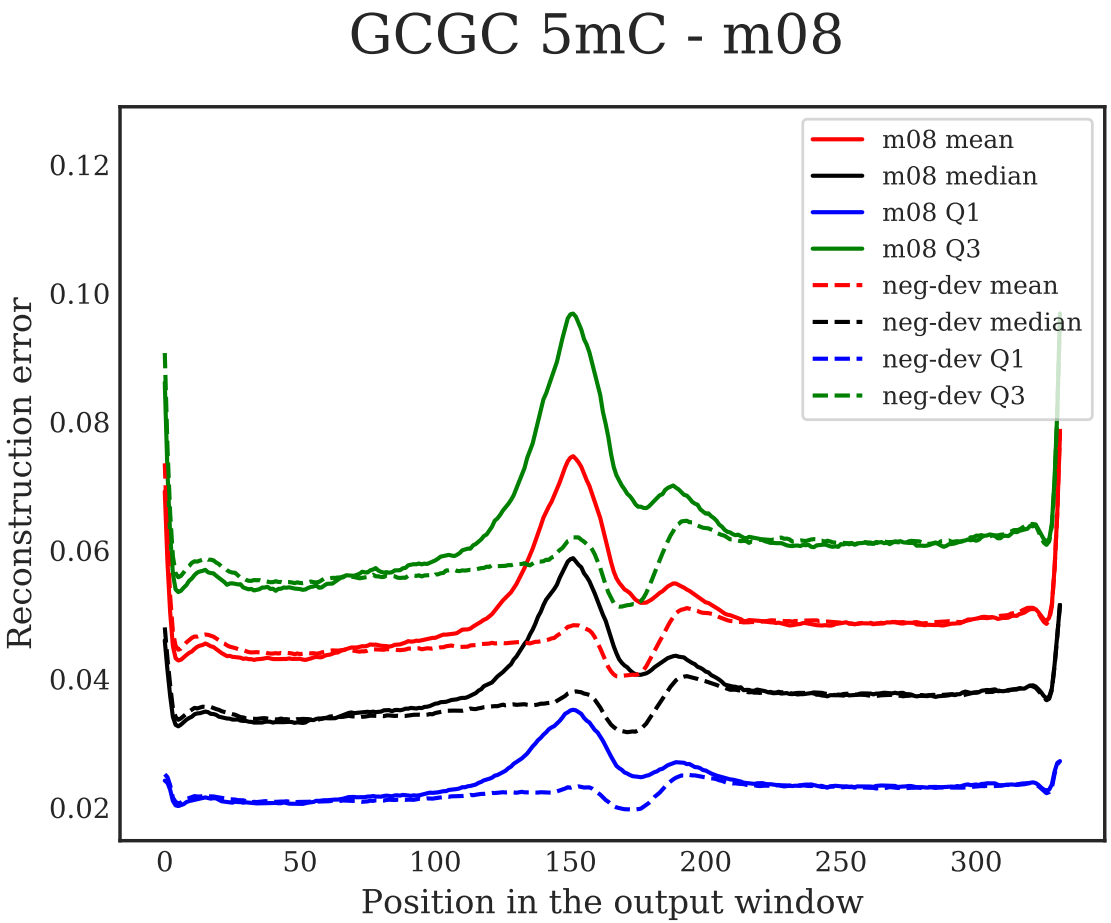
Zrekonštruovaný signál



Architektúry autoenkóderov



Chybové profily

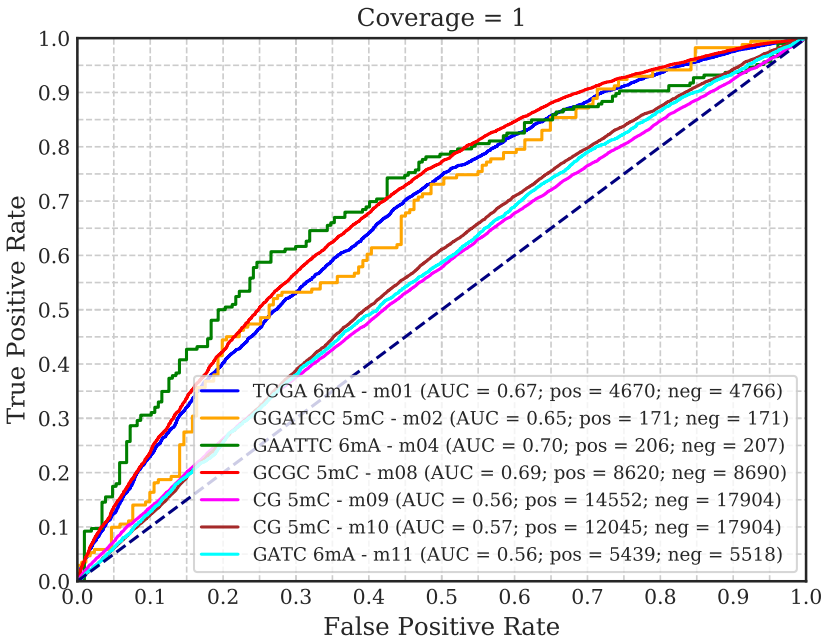


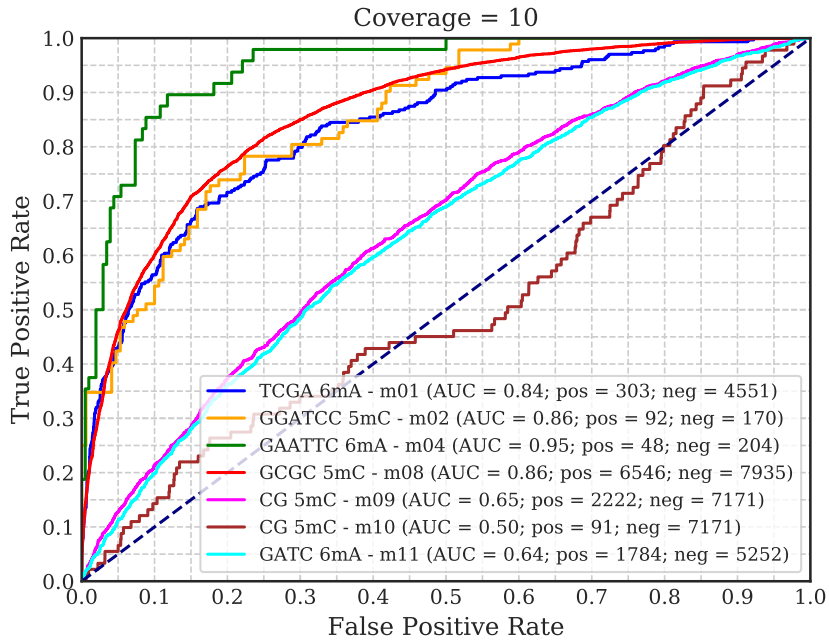
- Testovacia fáza: poznáme iba profil nemodifikovaných báz

Testovanie našej stratégie

- 7 modifikácií vyskytujúcich sa rôznych kontextoch.
- Metrika: ROC AUC, PR AUC

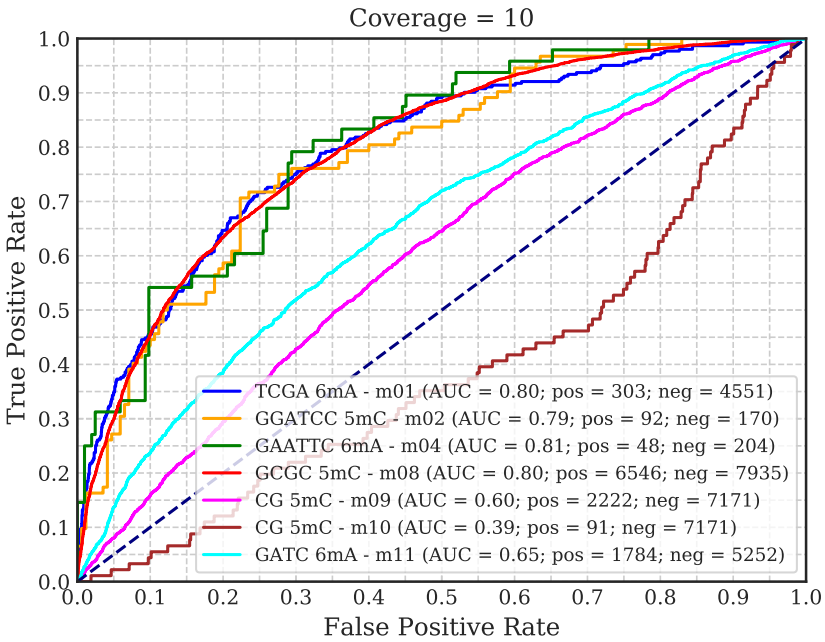
Jedno vs viacero čítaní



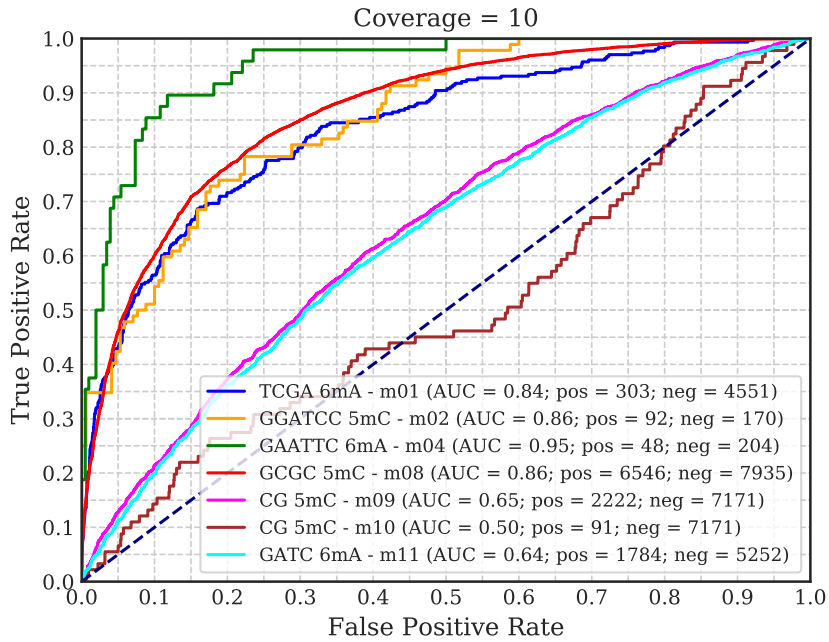


- S kontextom na vstupe dekódera

Pridanie kontextu na vstup dekódera

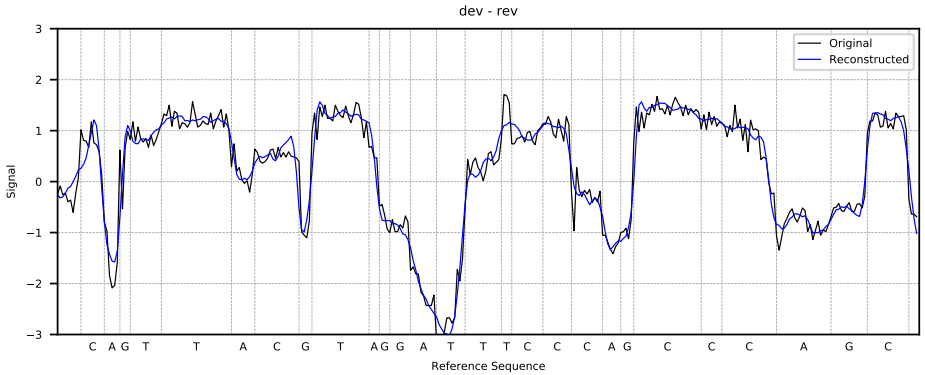


(a) Bez kontextu

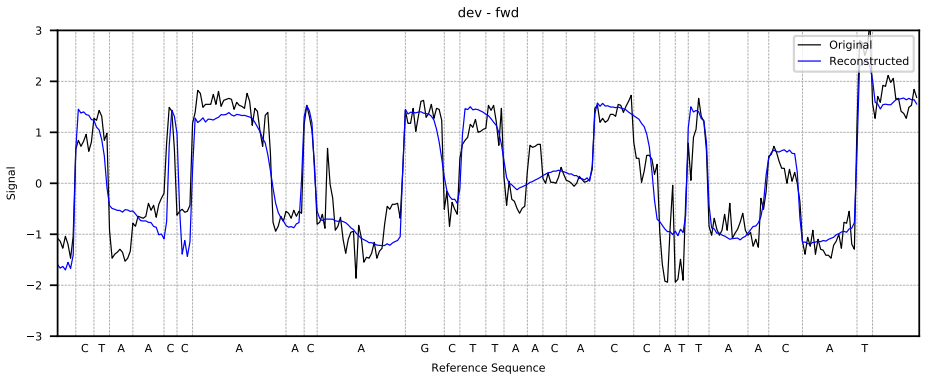


(b) S kontextom

CAE vs DAE

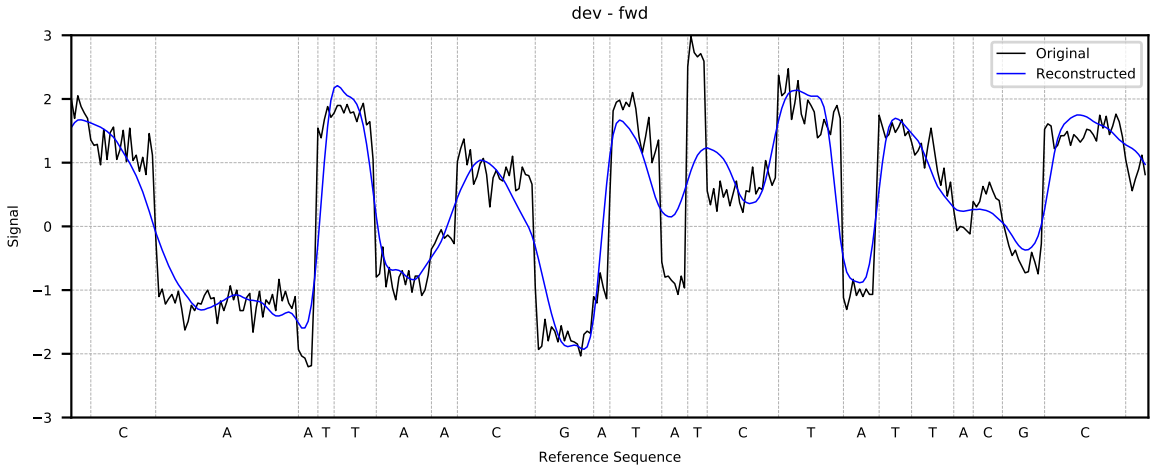


(c) DAE64



(d) CAE32

DAE32

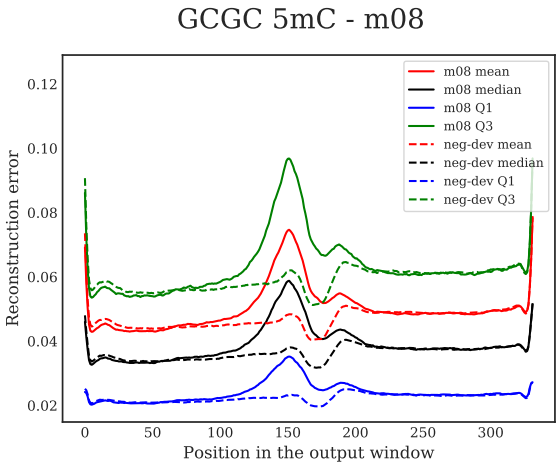

dev - fwd
Original
Reconstructed
Signal
3
2
1
0
−1
−2
−3
C A A T T A A C G A T A T C T A T T A C G C
Reference Sequence

Zhrnutie

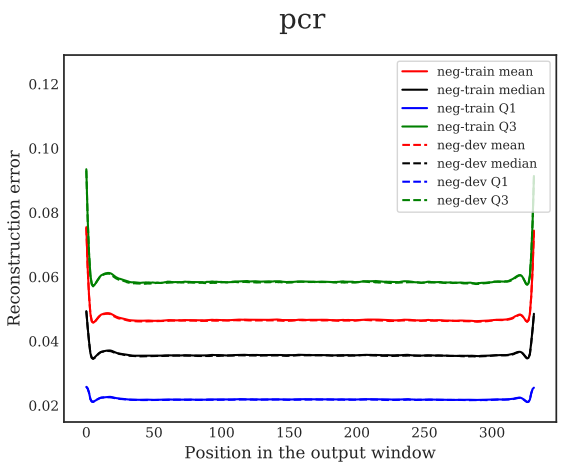
- Vyskúšali sme viacero architektúr autoenkóderov.
- Konvolučné neurónové siete fungovali lepšie.
- Potrebujeme viacero čítaní, aby sme vedeli detegovať modifikácie.
- Aj predchádzajúce prístupy potrebujú vyššie pokrytie.
- Pridanie kontextu báz pomáha.

Ďakujem za pozornosť.

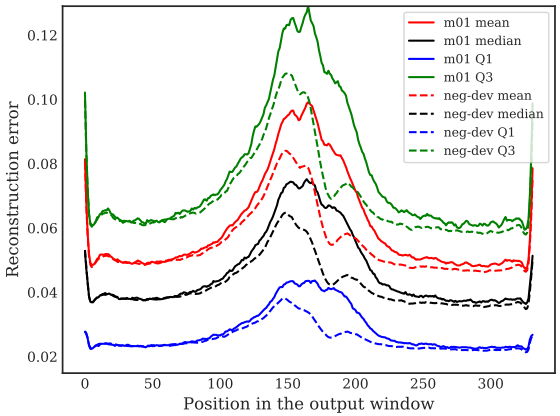
Chybové profily



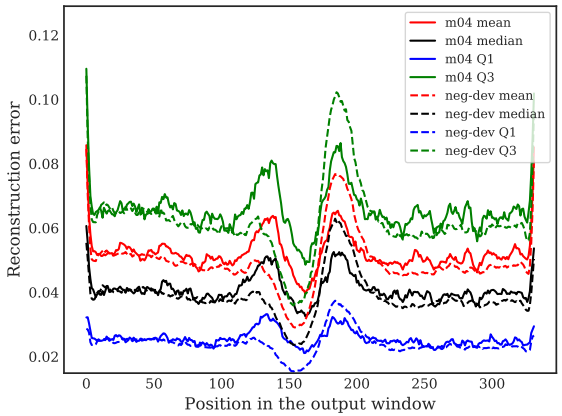
(a) GCGC



(b) Žiaden motív



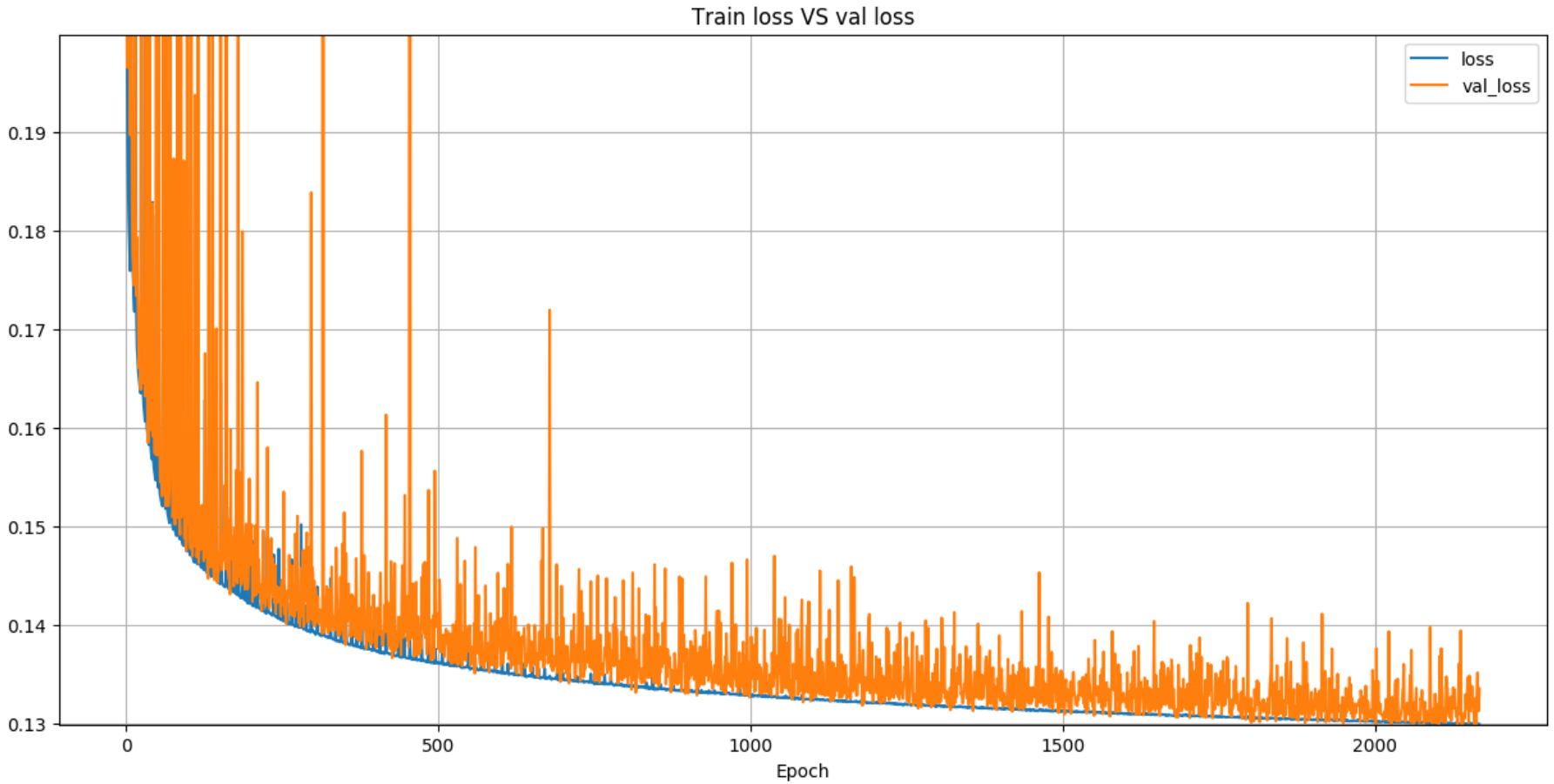
(c) TCGA



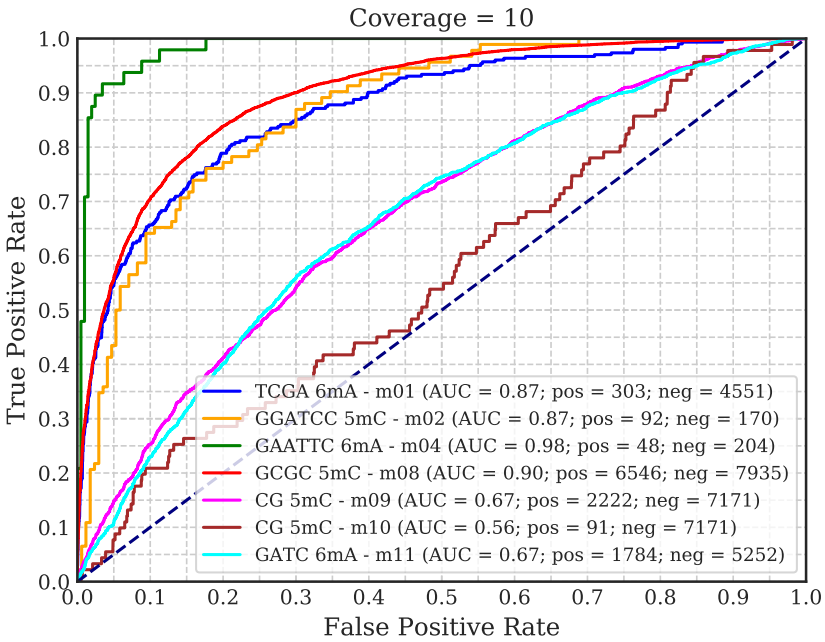
(d) GAATTC

Most of the networks were trained for 2000 epochs but our analysis showed that we could train these networks for only approximately 400 – 500 epochs. The validation error was still decreasing even after the 500th epoch but during the offline analysis we have not observed any systematic improvements for all of the methylations. Some methylations improved a little bit but then some got a little bit worse.

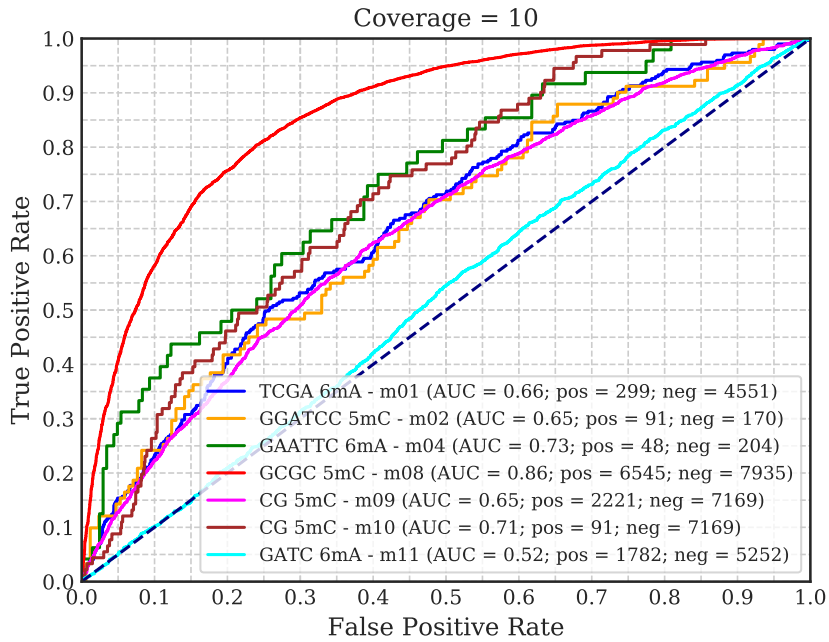
Chyba počas trénovania



Porovnanie s Tombom



(a) Náš najlepší model



(b) Tombo