Univerzita Karlova

Přírodovědecká fakulta

Studijní program: Bioinformatika

Studijní obor: B-BINF

Tomáš Preisler

Analýza single cell dat za pomoci transformerů

Analysis of single cell data using transformers

Typ závěrečné práce:

Bakalářská práce

Vedoucí práce/Školitel:

Mgr. Jan Stuchlý, Ph.D.

Praha, 2024

Prohlašuji, že jsem závěrečnou práci zpracoval samostatně a že jsem uvedl všechny použité informační zdroje a literaturu. Tato práce ani její podstatná část nebyla předložena k získání jiného nebo stejného akademického titulu.

V Praze dne 29.4.2024 Tomáš Preisler

.....

Název práce: Analýza single cell dat za pomoci transformerů

Autor: Tomáš Preisler

Ústav: Katedra buněčné biologie

Vedoucí bakalářské práce: Mgr. Jan Stuchlý, Ph.D.

Abstrakt:

Tato bakalářská práce je zaměřena na využití metod strojového učení a deep learning, při snaze automatizovat anotace buněk na základě jejich genové exprese (single-cell data). Především se zabývá využitím transformerů a architektur od nich odvozených, tedy performery. Hlavní část práce se věnuje popisu a charakteristice, jak tyto architektury fungují v jejich základní formě, popis modelů vytvořených pro anotaci buněk a jejich následné porovnání.

Klíčová slova: single-cell RNA sekvenování, deep learning, transformery, anotace buněk

Title: Analysis of single cell data using transformers

Autor: Tomáš Preisler

Department: Department of cell biology

Supervisor: Mgr. Jan Stuchlý, Ph.D.

Abstract:

This bachelor's thesis is focused on the use of machine learning and deep learning methods in an attempt to automate the annotation of cells based on their gene expression (single-cell data). In particular, it is focused on the use of transformers and architectures derived from them, such as performers. The main part of the work is devoted to the description and explanation of how these architectures work in their basic form, the description of the models created for cell annotation and their comparison.

Keywords: single-cell RNA sequencing, deep learning, transformers, annotating cell types

Poděkování

Chtěl bych vyjádřit své upřímné poděkování mému školiteli panu, Mgr. Janu Stuchlému, Ph.D. za odborné vedení, ochotu, trpělivost a cenné rady, které mi pomohly při zpracování této práce. Také bych rád poděkoval rodičům za podporu a povzbuzení, která pro mě byla během mého tříletého studia klíčová. Děkuji i spolužákům Anet, Aničce, Márovi a Tomovi za jejich pomoc a ochotu při našem společném studiu.

Obsah

Úvod

Moderní Single-Cell technologie, jako je například single-cell RNA-sequencing, nám v současné době umožňují detailní deskripci heterogenity a identifikaci funkční rozdílnosti na úrovni jednotlivých buněk (Adil et al., 2021). Přesná anotace buněk se tak stává klíčovou problematikou tohoto odvětví a je zcela klíčová pro pokročilý biologický výzkum, případně pro medicinální využití (Haque et al., 2017; Heumos et al., 2023; Su et al., 2022). Ve spojitosti s rychlým rozvojem těchto technologií vzrůstá i komplexita získaných dat (Yang et al., 2022).

Jednou z klasických metod anotace buněk, je anotace, kde jsou clustery buněk manuálně anotovány na základě marker genů. Samotné clusterování buněk je výsledkem unsupervised learning algoritmů, jako je například k-means (Yang et al., 2022). Marker geny jsou takové geny, jejichž úrovně exprese se používají k identifikaci a charakterizaci specifických typů buněk v rámci heterogenní populace (Huang et al., 2021; Su et al., 2022). Problémem této metody je její závislost na sadách marker genů, které jsou tvořeny na základě výzkumů, a proto může docházet k jejich nepřesné identifikaci kvůli chybám a nepřesnostem v experimentech, které následně vedou k chybám při rozpoznávání buňky (Huang et al., 2021). Další podstatnou překážkou jsou nové typy buněk, které zatím nejsou dostatečně prostudovány a neexistují pro ně kvalitní sady marker genů (Huang et al., 2021; Su et al., 2022; Yang et al., 2022).

Spearmanova, nebo Pearsonova korelace jsou standardní typy korelací využívány tzv. correlation based metodami (Yang et al., 2022). Tyto metody jsou založeny na postupech, které měří korelaci exprese genů mezi datasetem anotovaných buněk a buňkou, jenž se snažíme rozpoznat. Tyto metody jsou potencionálně náchylné k tzv. batch efektu experimentů. I když existují metody pro jeho korekci, je velmi složité rozlišovat biologickou relevanci od čistě technických rozdílů, kdy využívané metody pro výpočet korelace nemusí být dostatečně robustní pro měření korelace dvou vysoce dimenzionálních scRNA-seq datových matic (Haghverdi et al., 2018; Korsunsky et al., 2019; Yang et al., 2022).

Ke zpracování dat toho typu byly již v minulosti vyvinuty a zároveň i využity deep learning techniky, jako je právě anotace pomocí klasifikátorů nebo anotace buněk pomocí náhodných rozhodovacích stromů (Alquicira-Hernandez et al., 2019; de Kanter et al., 2019). Tyto metody přináší mnoho výhod, umožňují hledání vzorů v datech, nezávisí na sadách marker genů a jejich výhodou je celková robustnost modelů vůči šumu v datech (Alquicira-Hernandez et al., 2019; Yang et al., 2022). Zároveň nevýhodou je, že velikost a složitost dat, je pro tyto modely příliš

vysoká, a proto je nezbytné často přistupovat k metodám dimenzionální redukce, jako je například principal component analysis (PCA) (Alquicira-Hernandez et al., 2019; Yang et al., 2022). Při aplikaci dimenzionální redukce může docházet ke ztrátě podstatných informací z dat, a právě proto v nedávně době vzrostla snaha o vývoj modelů, které by byly schopné analyzovat scRNA-seq data, aniž by muselo docházet k velikému zásahu do samotných dat (Chen et al., 2023; Cui et al., 2023; Yang et al., 2022).

V současné době též dochází k významnému posunu analýzy přirozeného jazyka za pomoci tranformerů a architektur od nich odvozených – např. performerů (Choromanski et al., 2020; Vaswani et al., 2017), které jsou díky své kapacitě a možnosti paralelního zpracování vstupních dat ideální architekturou pro řešení tohoto úkolu (Yang et al., 2022). Transformery jsou od roku 2017 považovány za state of the art v oblasti analýzy a zpracování textu, kde nahradily rekurentní neuronové sítě (Schmidt, 2019; Turner, 2023; Vaswani et al., 2017).

Architektura transformerů je aktuálně také s jistým úspěchem aplikována pro analýzu scRNA-seq dat, které díky velké kapacitě těchto modelů nepotřebují projít dimenzionální redukcí ani žádnou další úpravou, která by mohla potenciálně vést ke ztrátě informací, jako je například mezi-genová interakce (Cui et al., 2023; Yang et al., 2022). Myšlenkou využití transformerů je paralela mezi buněčnou expresí a větou přirozeného textu, kde slovo je reprezentováno jako exprese genu a vztahy mezi slovy jsou koexprese[1] jednotlivých genů. Tento pokrok může vést k výraznému zrychlení a zpřesnění automatické anotace buněk na základě scRNA-seq dat a vnést podrobnější vhled do mezigenových vztahů a interakcí, za pomoci analýzy vnitřních parametrů těchto modelů (Yang et al., 2022).

[1] Koexprese genů je jev, kdy dva nebo více genů jsou společně aktivovány a exprimovány v buňce. Tento proces naznačuje, že dotyčné geny mohou být regulovány stejnými transkripčními faktory, sdílet podobné regulační mechanismy, nebo být součástí stejných biologických cest nebo funkčních skupin.

Single-cell

Single-cell RNA sequencing technologie způsobily výrazný posun v pochopení kompozice buněk v organismech a interakcí buněk mezi sebou (Jovic et al., 2022). „*Tyto technologie umožnily měření transkriptomových profilů v bezprecedentním měřítku a rozlišení, čímž způsobili naprostou revoluci v molekulární biologii“* (Heumos et al., 2023).

Single-cell RNA sequencing (scRNA-seq) umožňuje analýzu genové exprese na úrovni jednotlivých buněk v heterogenních populacích. Na rozdíl od standardních RNA sequencing metod, které měří průměrnou genovou expresi celé populace buněk a získávají tak informaci o průměrném chování dané buněčné populace, scRNA-seq nabízí mnohem detailnější vhled do heterogenních populací díky tomu, že analyzuje transkriptom každé buňky zvlášť (Haque et al., 2017; Heumos et al., 2023).

Příprava scRNA-seq dat

Prvním krokem je izolace buněčné populace z tkáně, která je předmětem zkoumání. Před lyzí je třeba buňky separovat., abychom mohli analyzovat genovou expresi každé buňky zvlášť. Micromanipulace, fuorescence activated cellsorting (FACS), laser capture microdissection (LCM) a microfluidic technology jsou metody, které jsou v současné době využívány pro tento úkol (Zhou et al., 2021). Přehled metod a jejich bližší popis je dostupný v apendixu práce (Obrázek 29).

Jakmile jsou buňky separovány, následuje lyze, díky které je možno zachytit mRNA jednotlivých buněk (Haque et al., 2017). Pokud chceme zajistit, aby docházelo k zachycení pouze mRNA s Poly(A) koncem, standardně je využit Poly(T) primer (Haque et al., 2017; Passmore & Coller, 2022; Slomovic et al., 2006; Zlatanova, 2023). Tento krok zajistí, že například ribosomální RNA (rRNA) nebude zachycena Poly(T) primerem, protože nedisponuje Poly(A) koncem, a tudíž nebude dále vstupovat do analýzy (Haque et al., 2017; Slomovic et al., 2006; Zlatanova, 2023).

Zachycená Poly(A)-mRNA je následně konvertována do komplementárního DNA řetězce (cDNA) za pomoci reverzní transkripce (Haque et al., 2017). Primery reverzní transkripce mohou mít na sobě sekvence navíc, které mohou následně sloužit například jako identifikátory původu mRNA (templátu pro cDNA) (Haque et al., 2017).

Polymerázová řetězová reakce (PCR) je ve většině případů využívána pro amplifikaci vzniklé cDNA. U většiny postupů dochází také k nukleotidovému barcode-taggingu, který má v sobě zakódovanou informaci o buňce, ze které cDNA pochází. Pokud zde neprobíhá k barcode-tagging znamená to, že k němu již došlo v předchozích krocích postupu. Amplifikované a označené cDNA ze všech buněk jsou následně společně sekvenovány za pomoci next-generation sequencing (NGS) metod, čímž získáme knihovnu sekvencí, kde u každé sekvence víme, z jaké buňky pochází (barcode-tagging nese informaci o buňce, ze které cDNA pochází) (Haque et al., 2017; Townes et al., 2019).

Knihovna je následně za pomoci bioinformatických nástrojů analyzována. V první fázi se ověřuje kvalita a variabilita knihovny, ve fázi druhé dochází k samotné analýze (Haque et al., 2017; Heumos et al., 2023; Luecken & Theis, 2019). Diagram procesu zisku scRNA-seq dat dostupný v apendixu práce (Obrázek 25).

Díky tomuto postupu získáváme databázi profilů genových expresí jednotlivých buněk (Haque et al., 2017).

Standardní metody analýzy scRNA-seq dat

Ze všech naměřených genů jsou standardně vybrány tzv. highly variable genes (HVGs) (Su et al., 2022). HVGs jsou takové geny, jejichž úrovně exprese se významně liší mezi různými buňkami nebo vzorky v populaci. Tyto geny často hrají zásadní roli při určování buněčné identity, funkce a reakce na podněty z prostředí. V rámci analýzy obvykle bývá zpracováváno mezi 1000 až 5000 genů (Cui et al., 2023; Townes et al., 2019; Yang et al., 2022).

Tento počet HVGs je ale příliš vysoký pro vizuální kontrolu, a proto musí dojít k dimenzionální redukci dat. Metody principal component analysis (PCA), multidimensional scaling (MDS), t-distributed stochastic neighbor embedding (t-SNE), on-negative matrix factorization (NMF) a uniform manifold approximation and projection (UMAP) jsou standardně využívány pro tento účel (Su et al., 2022; Szubert et al., 2019; Townes et al., 2019).

Jednou z aplikací scRNA-seq je identifikace a následná anotace buněčných subpopulací v populaci buněk. Clusterování je klasickým krokem při standardní analýze. ScRNA-seq data však mohou obsahovat vysoké množství šumu, který by mohl bránit ve správné identifikaci buněčných populací. Toto úskalí napomáhá vyřešit již dříve zmíněná dimenzionální redukce Nicméně, nežádoucím jevem dimenzionální redukce je, že můžeme přijít o podstatné informace o expresi genů (Su et al., 2022; Townes et al., 2019; Yang et al., 2022).

Clusterování scRNA-seq dat je obecně založeno na algoritmech, jako je klasický k-means algoritmus (Kiselev et al., 2017; Su et al., 2022). Mezi tyto algoritmy se řadí například single-cell consensus clustering (SC3) (Kiselev et al., 2017; Su et al., 2022). Ke clusterování je také využíván R balíček Seurat, kde je clusterování inspirováno metodou Nearest Neighbor Networks (Heumos et al., 2023; Su et al., 2022). Autoři článku *„Data analysis guidelines for single-cell RNA-seq in biomedical studies and clinical applications"* tvrdí, že studie naznačují, že metody SC3 a metody z knihovny Seurat jsou lepší nebo podobně dobré jako všechny ostatní konkurenční metody při clusterování buněk do subpopulací, které jsou založeny na podobných technikách (Su et al., 2022).

Samotná anotace buněčných subpopulací (vytvořených clusterů) může probíhat manuálně. Nicméně tento postup je velmi časově náročný, a navíc může vést k chybám, způsobených subjektivním pohledem anotátora. Z tohoto důvodu byly vyvinuty metody pro automatickou anotaci buněk, které lze rozdělit do tří hlavních skupin (Su et al., 2022; Yang et al., 2022).

- marker gene-based metody
- reference transcriptome-based metody
- supervised machine learning-based metody

Marker gene-based metody

Marker gene-based metody závisí na dostupnosti markerů specifických pro buněčný typ, které však nemusí existovat pro všechny buňky v populaci, jenž se snažíme anotovat (Su et al., 2022; Yang et al., 2022).

CellMarker nebo PangleDB jsou standardními zdroji těchto marker-genů pro velké spektrum buněk lidské a myší populace. Celkově obsahují okolo 14 000 setů marker-genů, které jsou specifické pro různé typy buněk (Franzén et al., 2019; Hu et al., 2023; Su et al., 2022).

Reference transcriptome-based metody

Reference transcriptom-based metody využívají labeled cell type-specific scRNA-seq datasety. Následně počítají korelaci mezi vstupními daty a datasetem. Vstupní scRNA-seq data jsou dále anotována na základě dat z datasetu, se kterým má nejvyšší korelaci (de Kanter et al., 2019; Su et al., 2022; Yang et al., 2022).

Nástroje, které se využívají pro tento přístup jsou například scmap, scMatch, SingleR a CHETAH (de Kanter et al., 2019; Kiselev & Hemberg, 2017).

Supervised machine learning-based metody

Supervised machine learning-based metody jsou založeny na předtrénování klasifikátorů na referenčních genových expresí buněk. Tyto klasifikátory jsou následně využity pro klasifikaci buněčných typů na základě vstupních scRNA-seq dat (Su et al., 2022).

SingleCellNet je jedním z programů, který je využíván pro klasifikaci buněk. Tento program je založený na multi-class random forest klasifikátorech (Tan & Cahan, 2019). Dalším programem je například ACTINN, který je založen na identifikaci buněčných typů za pomoci neuronových sítí (Ma & Pellegrini, 2020).

Strojové učení

Strojové učení představuje obor v rámci umělé inteligence a informatiky, který se soustředí na využívání dat a algoritmů k imitaci procesu učení člověka. Učením se rozumí zlepšování se v úkolu, který má daný model řešit a řešit ho zároveň i efektivně (Goodfellow et al., 2016).

„Podle Arthura Samuela je strojové učení definováno jako studijní obor, který počítačům poskytuje schopnost učit se bez explicitního programování“ (Mahesh, 2018).

Strojové učení není jediný algoritmus, ale naopak třída mnoha různých algoritmů, které jsou zaměřeny na různé typy dat. Je důležité tedy zdůraznit, že neexistuje žádný univerzální algoritmus, který by byl nejvhodnější pro řešení všech problémů, ve kterých se využívá strojového učení (Alzubi et al., 2018; Goodfellow et al., 2016; Mahesh, 2018).

Algoritmy strojového učení se dají dle jejich stylu učení rozdělit do tří kategorií:

- supervised learning – učení s učitelem
- unsupervised learning – učení bez učitele
- reinforcement learning – zpětnovazebné učení

Supervised learning (učení se s učitelem)

„Supervised Learning je paradigma strojového učení pro získávání informací o vztahu vstup-výstup systému založeného na sadě tréninkových dat, u kterých známe vstup i správný výstup“ (Liu & Wu, 2012). Z této definice plyne, že model při tréninku dostává na vstupu data, ke kterým předem známe správný výsledek a na základě porovnání výstupu modelu a správného

výsledku se model učí efektivně řešit daný problém (Goodfellow et al., 2016). Tento přístup vyžaduje dataset, který je předem zpracován, a tudíž potřebuje externí asistenci (Mahesh, 2018).

Unsupervised learning (učení bez učitele)

Na rozdíl od supervised learningu, kde trénovací data obsahují zároveň i správné výstupy (labels), data u unsupervised learningu správné výstupy nemají. Unsupervised Learning je tedy přístup strojového učení, kdy se model učí detekovat vzory v datasetech, které nemají přímo označený správný výstup (Goodfellow et al., 2016; Naeem et al., 2023). Tyto modely se tedy musí samy naučit hledat struktury a vztahy v datech (Goodfellow et al., 2016).

Mezi algoritmy unsupervised learningu, které se využívají v single-cell analýze, se například řadí PCA Principal Component Analysis), K-Means Clustering a autoencodery.

Semi Supervise Learning

Jedná se o kombinaci supervised a unsupervised learningu. Tréninkový dataset obsahuje jak data se správným výstupem (labeled data) tak i data, která neobsahují informaci o správném výstupu (unlabeled data). Tento přístup je využíván při tréninku jazykových modelů (Mahesh, 2018).

Reinforcement learning (zpětnovazebné učení)

Reinforcement learning je technika strojového učení, která trénuje model tak, aby se rozhodoval k dosažení nejlepších výsledků (Mahesh, 2018). Tato metoda napodobuje učící se proces „pokus-omyl“. Model se pokouší vyřešit daný problém a prostředí ho odměňuje nebo trestá za kroky, které dělá (Goodfellow et al., 2016). Kroky modelu, jež vedou k vyřešení problému jsou v průběhu tréninku posilovány a naopak kroky, které model odvádějí od cíle, jsou penalizovány.

Deep learning

Deep learning je pododvětví machine learningu. Deep learning je zaměřen na učení hlubokých neuronových sítí, díky kterému je schopen naučit se reprezentaci dat s několika úrovněmi abstrakce. Deep learning modely si tvoří vlastní datovou reprezentaci. Díky tomuto přístupu dochází v posledních letech k významným pokrokům v oblastech rozpoznávání řeči, rozpoznávání objektů, detekce objektů, aj. Deep learning je také využíván při vývoji nových léků a v genomice (Lecun et al., 2015; Schmidhuber, 2014).

Mezi hlavní zástupce deep learningu se řadí hluboké feed-forward neuronové sítě, rekurentní neuronové sítě, konvoluční neuronové sítě a v neposlední řadě transformery (Schmidt, 2019; Vaswani et al., 2017).

Word Embedding

Embedding je technika reprezentace dat, kde k jednotlivým slovům ze slovníku přiřadíme vektor reálných čísel, který efektivně zaznamenává jejich sémantický význam ve spojitém vektorovém prostoru[2] (Almeida & Xexéo, 2019). Slova, která mají v daném kontextu podobný význam, by v tomto prostoru měla být namapována blízko sebe. Tyto embedding vektory následně slouží jako vstupní vrstva pro neuronové sítě. Každé slovo je reprezentovano jako vektor (Almeida & Xexéo, 2019; Goodfellow et al., 2016).

Během tréninku tyto embedding vektory zdokonalujeme na základě kontextu, ve kterém se slova vyskytují (Goodfellow et al., 2016).

Word embedding je standardně základem moderních modelů zpracování přirozeného jazyka a jsou integrovány do pokročilejších neuronových architektur, jako jsou rekurentní neuronové sítě (RNN), sítě dlouhodobé krátkodobé paměti (LSTM) a transformery (Almeida & Xexéo, 2019; Goodfellow et al., 2016; Vaswani et al., 2017).

K embeddingu se dá poté přistupovat pomocí one-hot vektoru (vektor obsahující pouze jednu jedničku na pozici odpovídající příslušnému slovu a jinak samé nuly).

Větu jsme schopni zakódovat pomocí one-hot matice. Každý řádek poté odpovídá jednomu one-hot vektoru.

Word2vec

Word2vec je technika, která umožňuje vektorovou reprezentaci slov. Vektory zachycují informace o významu slova i o jeho použití v kontextu (Mikolov et al., 2013). Modely jsou neuronové sítě s dvěma vrstvami, kde na vstupu obdrží část textu, následně každé slovo z tohoto textu umístí do vektorového prostoru. Aby byl embedding efektivní, vektorový prostor musí být dostatečně velký. Standardně se jedná o několikaset dimenzionální prostory. Word2vec využívá k vytvoření těchto jedinečných vektorů dvě techniky Bag-Of-Words (CBOW) a continuously sliding skip-gram (Mikolov et al., 2013).

[2] Prostor, obsahující vektory se spojitými souřadnicemi.

Gene2vec

Metoda Gene2vec je technika, která se zabývá vytvářením distribuovaných reprezentací genů na základě jejich koexprese (Du et al., 2019). Tato technika se inspiruje koncepty z oblasti distribuovaného zpracování přirozeného jazyka (NLP), kde jsou slova reprezentována vektorovými prostorovými modely. Gene2vec využívá tyto vzory koexprese k vytvoření distribuovaných vektorových reprezentací genů. (Du et al., 2019)

Gene2vec se stal naprosto klíčový v tréninku modelů založených na architektuře transformerů, které se zaměřují na scRNA-seq data a následnou anotaci buněk.(Yang et al., 2022)

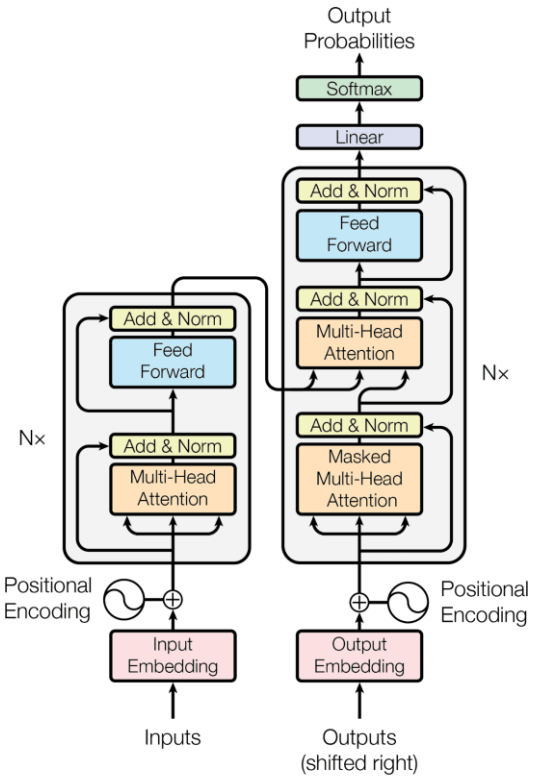
Transformery

Vaswani et al. (2017) a jeho tým byli první, kteří představili myšlenku transformeru. Ty se staly prakticky okamžitě modely, které jsou dodnes s menšími úpravami využívány ke zpracování přirozeného jazyka a jsou na nich založeny jazykové modely, jako je například BERT nebo GPT-4 (Devlin et al., 2018; Khurana et al., 2023; OpenAI et al., 2023). *„Díky transformerům dochází k pokrokům ve zpracování“ přirozeného jazyka, počítačového vidění, a časoprostorového modelování“* (Turner, 2023).

„Transformery umožňují mnohem více paralelizace a mohou se stát novým state of the art v překladu textu poté, co byly trénovány pouhých dvanáct hodin na osmi GPU P100“ (Vaswani et al., 2017). Paralelizace u rekurentních sítí nebyla možná, kvůli rekurentní části modelu. Tuto část ovšem transformery neobsahují a spoléhají se pouze na attention mechanismus (Schmidt, 2019; Vaswani et al., 2017).

Architektura základního transformeru

Transformery mají dvě hlavní podjednotky, encoder a decoder. (Vaswani et al., 2017) Hlavní funkcí encoderu je mapovat vstupní sekvenci na sekvenci spojitých reprezentací, která je následně přiváděna do decoderu. Decoder přijímá výstup encoderu společně s výstupem decoderu z předchozího kroku, následně vygeneruje výstupní sekvenci. *„V každém kroku je model auto-regresivní, využívá dříve vygenerované symboly jako dodatečný vstup při generování dalšího symbolu“* (Vaswani et al., 2017). Vaswani v původním implemetaci modelu využil 6 encoderů a 6 decoderů (N = 6).



Obrázek 1 Transformer – architektura modelu encoder-decoder. Obrázek převzat z článku „Attention Is All You Need“ (Vaswani et al., 2017).

Úprava vstupních dat

Transformer není schopen analyzovat text v jeho přirozené reprezentaci. Proto jsou všechna data vstupující do modelu nejdříve upravena. Model využívá standardního embeddingu (Vaswani et al. (2017) využívali embedding s 512 dimenzemi). Ten ovšem neobsahuje informaci o pozici slova v sekvenci, proto je k tomuto embeddingu přičten tzv. positional embedding (Turner, 2023; Vaswani et al., 2017). Díky positional embeddingu model získává informaci o přesné pozici slova v sekvenci, bez nutnosti využití rekurentní nebo konvoluční sítě (Vaswani et al., 2017). *„Positional embedding má stejnou dimenzi jako standardní embedding, a to z toho důvodu, aby je šlo bez problémů sečíst dohromady“* (Vaswani et al., 2017).

Vaswani et al. (2017) v původní implementaci využili k výpočtu positional embeddingu sinus a cosinus v tomto tvaru:

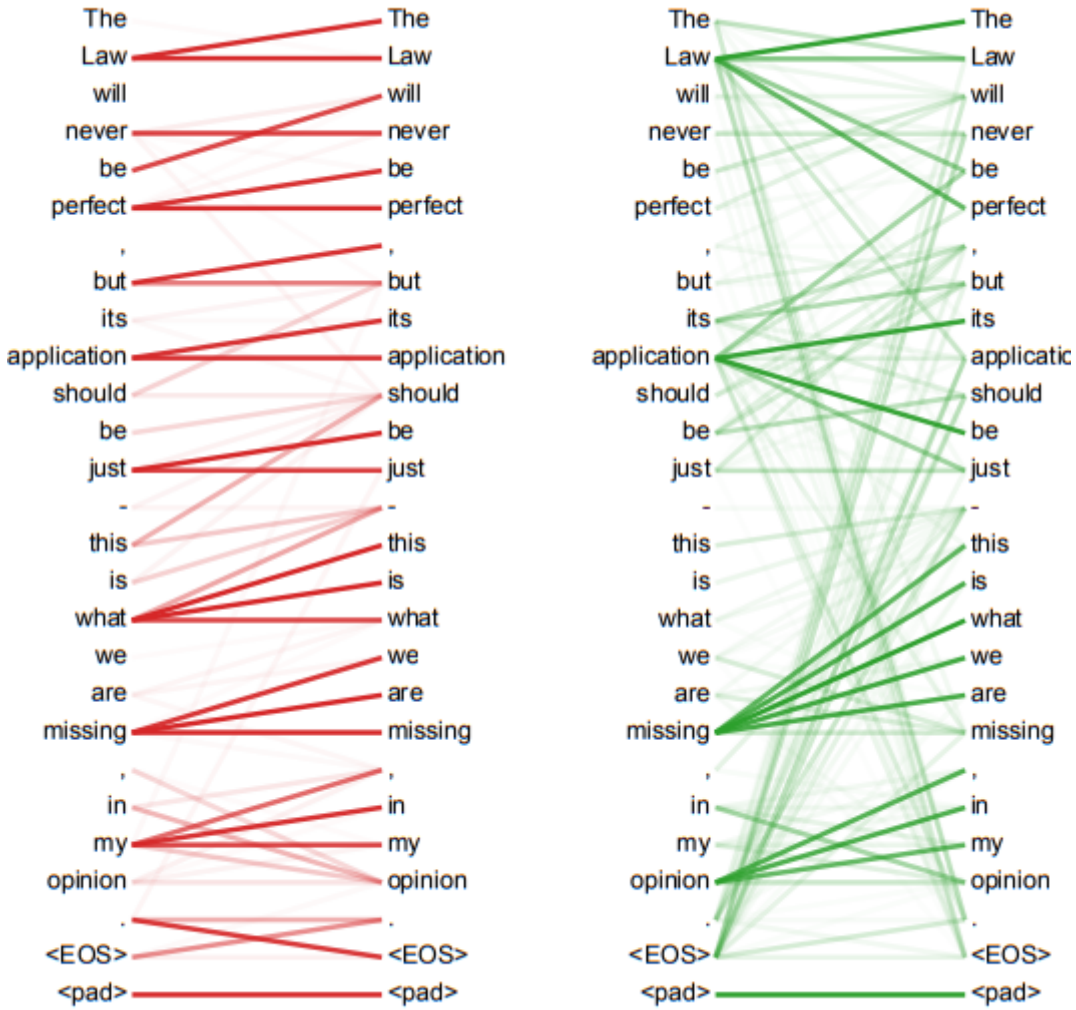
$$PE_{(pos,2i)} = sin\,(pos/10000^{2i/d_{model}})$$

$$PE_{(pos,2i+1)} = cos\,(pos/10000^{2i/d_{model}})$$

Kde *pos* označuje pozici slova, *i* dimenzi a *d* celkový počet dimenzí, se kterými model pracuje. „*Z toho vyplývá, že každá dimenze positional embeddingu odpovída sinusoidě“* (Vaswani et al., 2017).

Attention

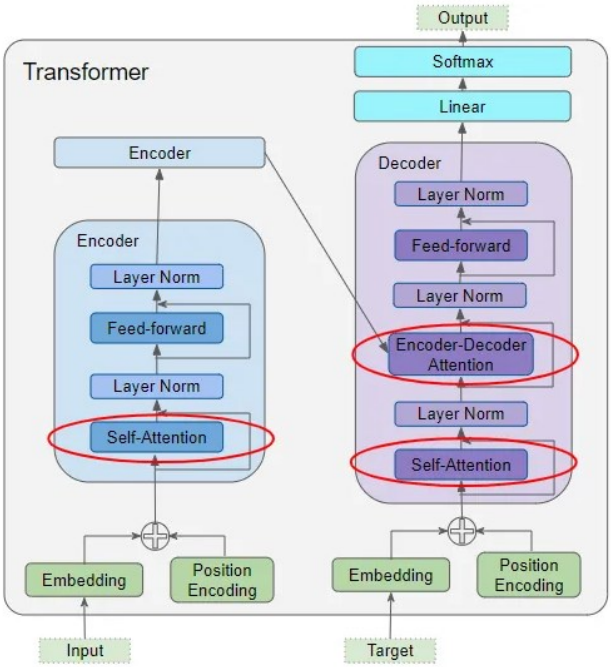
Attention byla navržena k identifikaci korelací mezi slovy ve větě za předpokladu, že se během tréninku naučila vzorce vět z tréninkových dat. Tato korelace je zachycena v maticích vah (W), buď pomocí unsupervised předtréninku, nebo za pomoci supervised fine-tuningu (Vaswani et al., 2017).



Obrázek 2 Příklad attention ze dvou attention heads. Každá attention head zachycuje rozdílné vztahy slov. Obrázek převzat z článku „Attention Is All You Need“ (Vaswani et al., 2017).

Attention využíváme v transformeru na třech místech (Vaswani et al., 2017).

- Self-Attention v encoderu – Attention input sekvence na sebe samou
- Self-attention v decoderu – Attention target sekvence na sebe samou
- Encoder-Decoder-attention v decoderu – Attention target sekvence na input sekvenci



Obrázek 3 Model transformeru se zvýrazněnými místy, kde se využívají různé druhy attention. Obrázek převzat z článku „Transformers Explained Visually (Part 3): Multi-head Attention, deep dive“ (Doshi, 2021).

Vaswani ve svém článku napsal, že attention lze popsat jako mapování query a množiny key-value na výstup, kde query, keys, values a výstup jsou vektory.

Z této definice vyplývá, že pro výpočet attention je nutno znát tři hodnoty Q (query), K (keys), V (values). Tyto hodnoty jsou vypočítávány pomocí lineární projekce vstupního embeddingu pro každé slovo. Každé slovo má tedy přiřazeny tři hodnoty, ze kterých se následně počítá attention (Vaswani et al., 2017). Mějme vstupní sekvenci X a předpokládejme, že má dimenzi d_{model}. Lineární projekce je parametrizována maticemi vah W_Q, W_K, W_V. Tyto váhy jsou upravovány při tréninku modelu. Pro jednu attention head (Obrázek 26) Q, K a V vypočítáme následovně:

$$Q = XW_Q$$

$$K = XW_k$$

$$V = XW_V$$

V transformerech je využíván tzv. multi-head attention mechanismus[3]. Lineární projekce jsou rozděleny hned do několika attention heads. Mějme *h* attention heads, dále pro každou attention head *i, kde* $i \in (1, \dots, h)$ máme samostatné matice vah $W_{Q,i}$, $W_{K,i}$, $W_{V,i}$. Výpočet následně probíhá stejně jako u single head attention:

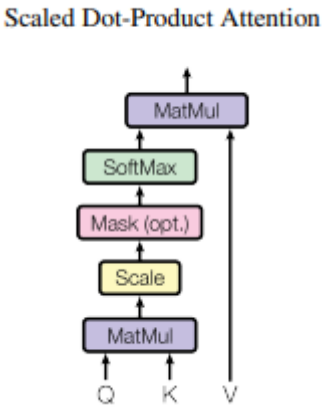
$$Q_i = XW_{Q,i}$$

$$K_i = XW_{K,i}$$

$$V_i = XW_{V,i}$$

Po získání Q, K, V je samotná Attention spočítána (Vaswani et al., 2017):

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$



Obrázek 4 Diagram výpočtu scaled dot-product attention z matic Q, K, V. Obrázek obsahuje i nepovinný krok maskování, který je vysvětlen níže. Obrázek převzat z článku „Attention Is All You Need" (Vaswani et al., 2017).

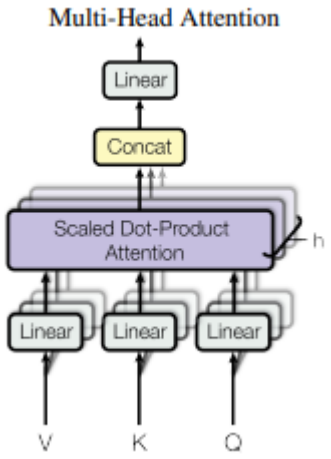
„Pro velké d_k *(definováno níže) skalární součiny Q a K rostou enormně do velikosti a softmax je následně tlačen do oblastí, kde má extrémně malý gradient. Proto jsou výsledné hodnoty přeškálovány* $1/\sqrt{d_k}$*"* (Vaswani et al., 2017). Tento vzorec poskytuje výsledný výstup self-attention bloku v transformerech.

[3] Vstupní data jsou rozdělena do více částí (hlav). Každá attention head aplikuje attention mechanismus nezávisle na ostatních, což umožňuje modelu zachytit různé pohledy na vstupní data.

Při používání multi-head attention je využit tento vzorec:

$$MultiHead(Q, K, V) = Concat(head_1, \dots, head_h)W_O$$

$$kde\ head_i = Attention(QW_{Q,i}, KW_{K,i}, VW_{V,i})$$



Obrázek 5 Diagram průběhu výpočtu multi-head attention z matic Q, K, V. Obrázek převzat z článku „Attention Is All You Need“ (Vaswani et al., 2017).

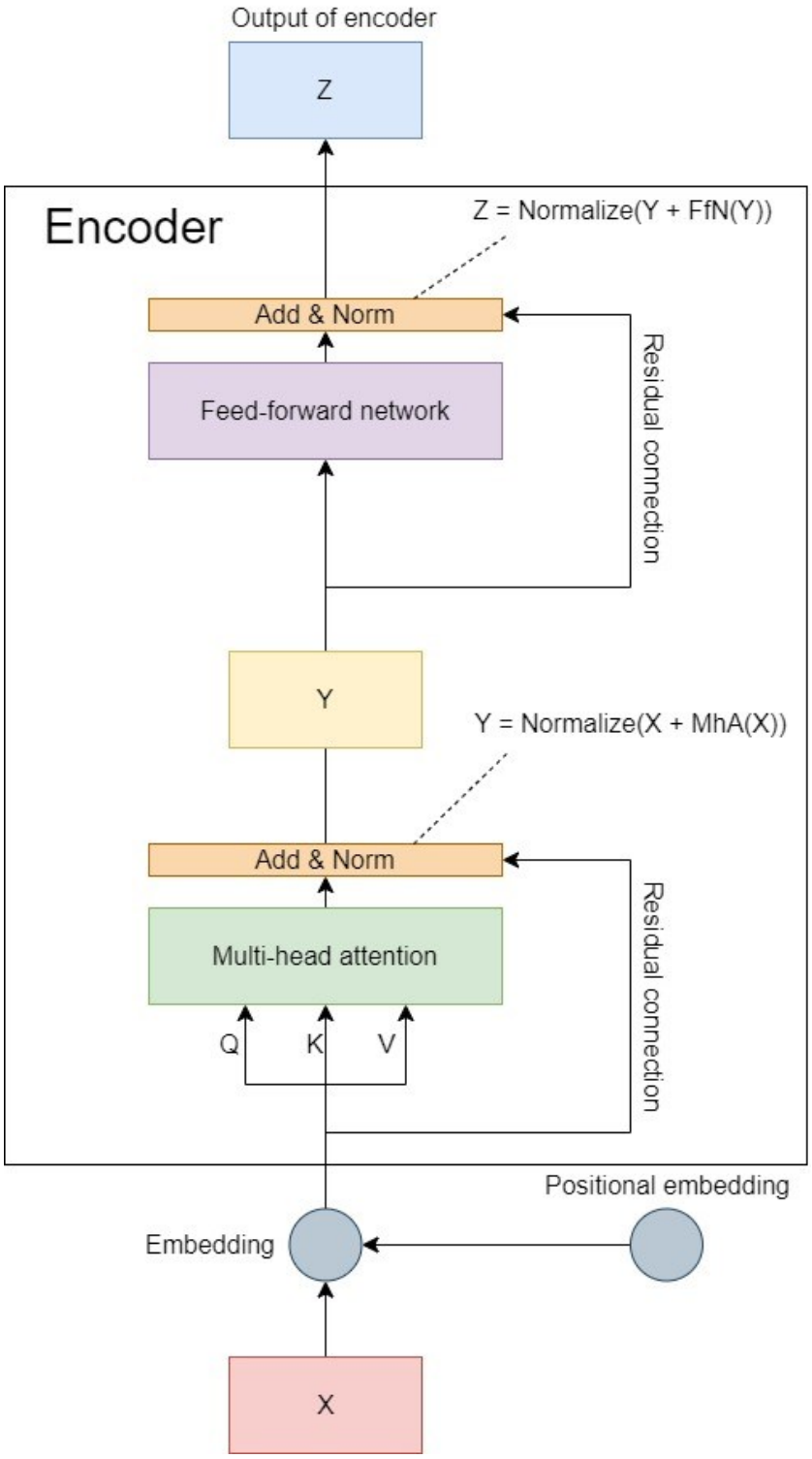
Kde W_O slouží jako lineární projekce[4] aplikovaná na konkatenovaná data[5] (Vaswani et al., 2017). Jednotlivé matice projekcí mají tyto rozměry: $W_{Q,i} \in \mathbb{R}^{d_{model} \times d_k}$, $W_{K,i} \in \mathbb{R}^{d_{model} \times d_k}$, $W_{V,i} \in \mathbb{R}^{d_{model} \times d_k}$, $W_O \in \mathbb{R}^{hd_v \times d_{model}}$, kde h je počet attention heads a $i \in (1, \dots, h)$. Vaswani et al. (2017) použil $d_v = d_k = d_{model} / h = 64$.

Multi-head attention přináší mnoho výhod. Hlavní výhodou je možnost se na vztahy mezi slovy dívat různými pohledy. Každá hlava poté následně zachycuje jeden z těchto pohledů (Vaswani et al., 2017). Příklad výpočtu attention lze najít v apendixu práce (Obrázek 26).

[4] Konkatenovaný vektor je transformován lineární projekcí do požadované výstupní dimenze
[5] Konkatenovaná data – Výstupy z každé hlavy jsou spojeny do jednoho vektoru.

Architektura encoderu

Encoder je první ze dvou podjednotek transformeru (Vaswani et al., 2017). Architektura encoderu je detailně popsána na obrázku níže (Obrázek 6).



Obrázek 6 diagram encoder bloku obsahující dvě podvrstvy: Multi-head attention podvrstvu a feed-forward neuronovou síť. Mezi každou podvrstvou vede residuální spoj, který přenáší hodnotu získanou v předešlé podvrstvě. Po sečtení výstupu každé podvrstvy a residuálního spoje je tento součet normalizován. Původní transformer obsahoval 6 encoder bloků.

Výpočet encoderu je možno popsat rovnicemi:

$$Y = LayerNorm(X + MhA(X))$$

$$Z = LayerNorm(Y + FfN(Y))$$

Kde X je vstupní embedding, Y je mezivýsledek encoderu po aplikování Multi-head attention (MhA), residuálního spoje a normalizace a Z je výsledný output encoder bloku po aplikování feed-forward neuronové sítě (FfN) na Y, residuálního spoje a normalizace. LayerNorm značí normalizaci dat (Vaswani et al., 2017).

Architektura decoderu

Decodery jsou založeny na stejné myšlence jako encodery. Jeden blok decoderu, který byl popsán v původní implementaci transfomeru, je ale složen ze tří podvrstev, nikoliv ze dvou, jak je tomu u encoderu. První podvrstvou je multi-head attention podvrstva, která je upravena o maskování. Druhou podvrstvou je také multi-head attention vrstva, ta je již standardní, jako u encoderu (Vaswani et al., 2017). Této vrstvě se také říká Encoder-Decoder-attention, jelikož matice Q je vypočítána z hodnot zpracovaných decodérem a matice K a V je počítána z výstupu encoderu (Obrázek 7) (Turner, 2023).

Masked multi-head attention blok obsahuje navíc od standardního multi-head attention bloku maskovací podčást, která zajišťuje auto-regresi modelu (Vaswani et al., 2017). Auto-regresivní znamená, že query počítá self attention pouze pro slova (tokeny), které jsou před ním. Toho lze ve výpočtu docílit tak, že přidáme maskovací matici (M) a celou sekvenci posuneme o jednu pozici. Tímto zajistíme, že attention pro slovo na pozici i je počítána pouze ze slov na pozici od 0 do i-1 (Obrázek 28) (Choromanski et al., 2020; Turner, 2023; Vaswani et al., 2017).

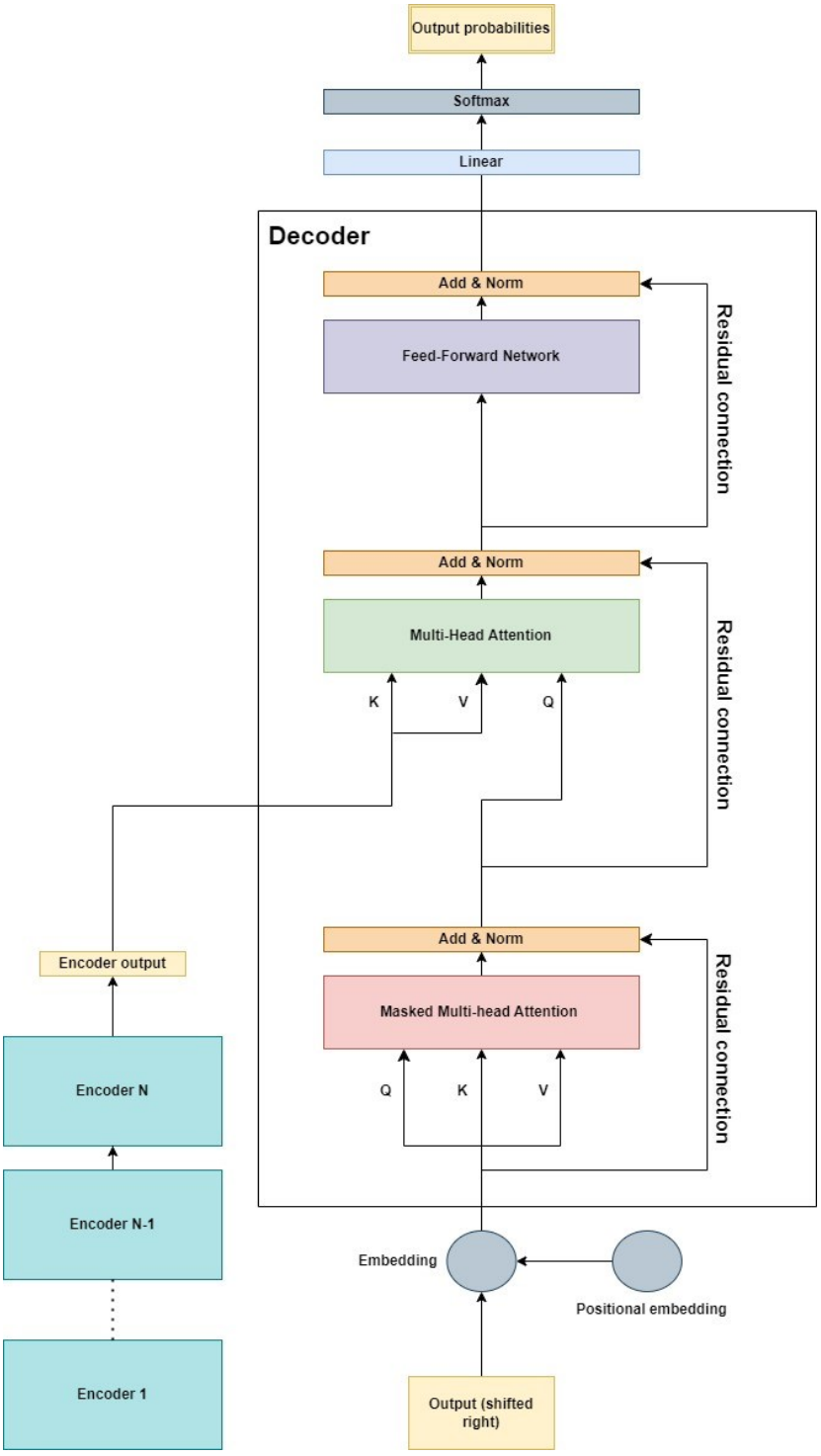
$$Attention(Q, K, V) = softmax\left(\frac{QK^T + M}{\sqrt{d_k}}\right)V$$

$$kde\ M = \begin{pmatrix} 0 & \cdots & -\infty \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{pmatrix}$$

Matice M obsahuje pod a na hlavní diagonále hodnotu 0. Nad hlavní diagonálou je poté hodnota $-\infty$. Rozměry matice M odpovídají rozměrům matice QK^T. Touto úpravou softmax funkce

přiřadí 0 každému slovu (tokenu), který následuje po query tokenu (počítáme attention pouze pro slova, která se nachází před query).

Druhá multi-head attention podvrstva funguje již bez maskování. Tato podvrstva je zodpovědná za mapování sekvence z decoderu na encoder sekvenci (Obrázek 7) (Vaswani et al., 2017).



Obrázek 7 Diagram decoder bloku obsahující tři podvrstvy. Masked multi-head attention podvrstvu, Multi-head attention podvrstvu a feed-forward neuronovou síť.

Poslední podvrstva je opět fully-connected feed-forward neuronová síť. Na výstup decoderu je následně použita lineární projekce a softmax funkce. Finálním výsledkem jsou tzv. output probabilities, které zachycují pravděpodobnosti, že dané slovo ze vstupní sekvence má být namapováno na slovo z výstupní sekvence (Obrázek 27) (Vaswani et al., 2017).

Vaswani et al. (2017) ve svém článku využíval výsledné výstupní pravděpodobnosti k překladu vět. Optimalizace maximalizuje pravděpodobnosti správného překladu slov.

Performery

Ačkoliv transformery jsou naprosto revoluční nástroj pro analýzu textu, tak jsou ve své základní formě neaplikovatelné na úplná scRNA-seq data, která se skládají z tisíců sloupců. Pokud bychom chtěli použít genovou expresi buňky jako větu, tato věta by se skládala z několika tisíců slov, a jelikož časová i paměťová složitost transformerů je kvadratická vůči vstupu, tak je tento proces takřka nemožný (Vaswani et al., 2017; Yang et al., 2022). Tento problém řeší úprava ve výpočtu attention. Této architektuře se následně říká performery (Choromanski et al., 2020).

Performery jsou modely založené na architektuře transformerů, které ale nahrazují standardní attention mechanismus rychlejší metodou, kde je attention matice A dekomponována do matic Q´ a K´, pro které platí, že $Q'(K')^T \approx \exp((QK^T)/\sqrt{d})$ (Obrázek 8). Výhoda v tomto přístupu je, že performery využívají postupy pouze s lineární (na rozdíl od kvadratické u standardních transformerů) složitostí vůči vstupu (Choromanski et al., 2020).

O aproximaci attention se již pokoušely metody, jako je například Locality-Sensitive Hashing (LSH) (Choromanski et al., 2020; Kitaev et al., 2020). Tyto přístupy však nepřináší tak přesnou aproximaci, jako je aproximace pomocí tzv. Fast Attention Via positive Orthogonal Random features (FAVOR+) algoritmu, který autoři popsali v článku „Rethinkin Atteniton With Performers“. FAVOR+ dokáže aproximovat softmax funkci v lineárním čase vůči délce vstupu (Choromanski et al., 2020).

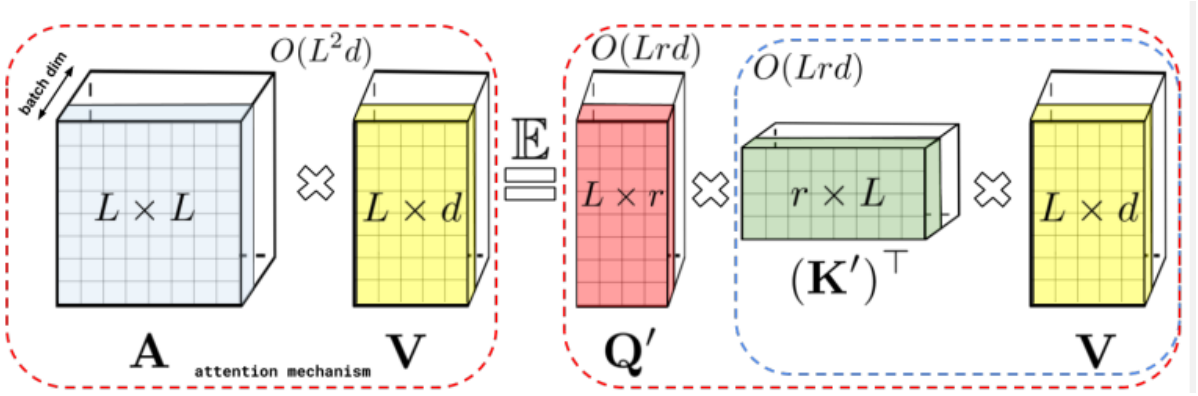
FAVOR+

Mějme attention matici $\boldsymbol{A} \in \mathbb{R}^{L\times L}$ kde $A(i,j) = K(q_i^T, k_j^T)$, kde q_i a k_j jsou sloupce z matice **Q** (Query) a **K** (Keys) a mějme kernel $K: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$definovaný pro randomizované mapování $\phi: \mathbb{R}^d \to \mathbb{R}_+^r$ jako:

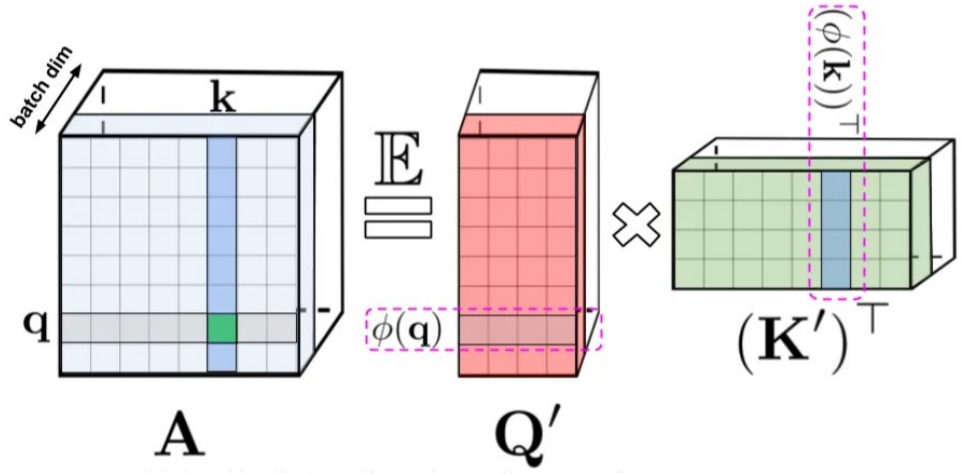
$$K(x,y) \stackrel{\text{def}}{=} \mathbb{E}(\phi(\mathrm{x})^T, \phi(\mathrm{y})^T)$$

Kernely nám umožňují reprezentovat skalární součin v jiném prostoru. Výstup kernel funkce bude stejný jako skalární součin dvou vstupů, které jsou nejprve upraveny pomocí funkce φ (Choromanski et al., 2020). To nám umožňuje učinit velikost matic Q' a K' pouze lineárně závislé na délce vstupu (nikoliv kvadraticky). Následně můžeme nejprve vynásobit matici K´ s maticí V a poté až s maticí Q´, čímž se vyhneme nutnosti pracovat explicitně s celou attention maticí A (Obrázek 8) (Choromanski et al., 2020).

$\phi(\mathbf{u})$ se nazývá *random feature map* pro $\mathbf{u} \in \mathbb{R}^d$. Pro $Q', K' \in \mathbb{R}^{L \times r}$, které mají řádky definovány jako $\phi(q_i^T)^T$ a $\phi(k_i^T)^T$ následně platí tvrzení, které popisuje Obrázek 8 a Obrázek 9:



Obrázek 8 Aproximace standardního attention mechanismu AV za pomocí random feature maps. Z obrázku vyplývá, že standardní attention mechanismus má časovou složitost O(L^2d) a prostorovou složitost O(L^2 + Ld). Attention mechanismus pomocí FAVOR+ má časovou složitost O(Lrd) a prostorovou složitost O(Lr + Ld + rd). Obrázek převzat z článku „Rethinking Attention with Performers“ (Choromanski et al., 2020).



Obrázek 9 Detailní ukázka rozložení attention matice na dvě menší matice Q´ a K´. Pro zisk řádků/sloupců matic Q´/K´ je využit random feature map funkce. Obrázek převzat z článku „Brief Review — Rethinking Attention with Performers“ (Tsang, 2022).

Definujme funkci ϕ s funkcemi $f_1, \ldots, f_l: \mathbb{R} \to \mathbb{R}$, funkcí $h: \mathbb{R}^d \to \mathbb{R}$ a náhodné vektory $w_1, \ldots, w_m \sim D$ pro nějakou distribuci $D \in \mathcal{P}(\mathbb{R}^d)$ jako:

$$\phi(\mathbf{x}) = \frac{h(\boldsymbol{x})}{\sqrt{m}}(f_1(w_1^T\boldsymbol{x}), \ldots, f_1(w_m^T\boldsymbol{x}), \ldots, f_l(w_1^T\boldsymbol{x}), \ldots, f_l(w_m^T\boldsymbol{x}))$$

Softmax kernel, který definuje standardní attention matici A je:

$$SM(x, y) \stackrel{\text{def}}{=} \exp(x^T y)$$

Pro $x, y \in \mathbb{R}^d$ a $z = x + y$ lze ukázat[6], že:

$$SM(x, y) = \mathbb{E}_{w \sim \mathcal{N}(0, I_d)}\left[\exp\left(w^T x - \frac{\|x\|^2}{2}\right)\exp\left(w^T y - \frac{\|y\|^2}{2}\right)\right]$$

w jsou náhodné vektory generovány z normální distribuce. Z rovnice jde vidět, že softmax kernel lze dekomponovat na dvě podčásti a to umožňuje *positive random map unbiased* aproximaci[7] (PRFs) takovou, že:

$$h(x) = \exp\left(-\frac{\|x\|^2}{2}\right), l = 1, f_1(x) = \exp(x)\ a\ \mathcal{D} = \mathcal{N}(0, I_d)$$

nebo:

$$h(x) = \frac{1}{\sqrt{2}}\exp\left(-\frac{\|x\|^2}{2}\right), l = 2, f_1(x) = \exp(x), f_2(x) = \exp(-x)\ a\ \mathcal{D} = \mathcal{N}(0, I_d)$$

Tyto estimátory nazveme $\widehat{SM_m^+}$ a $\widehat{SM_m^{hyp+}}$ (Tabulka estimátorů dostupná v apendixu - Tabulka 1). Autoři článku dokázali, že positive hyperbolic random features estimator ($\widehat{SM_m^{hyp+}}$) je lehce přesnější než positive random features estimator ($\widehat{SM_m^+}$) (Choromanski et al., 2020). Oba estimátory efektivně aproximují softmax funkci, ale jejich výpočetní složitost je pouze lineárně závislá na délce vstupu (nikoliv kvadraticky jako tomu je u transformerů), což z performerů dělá výborný nástroj pro analýzu extrémně dlouhých sekvencí, jako je například genový profil buňky (Choromanski et al., 2020; Yang et al., 2022).

[6] Důkaz č. 1

[7] $\mathbb{E}_{w \sim \mathcal{N}(0, I_d)}$ je možné aproximovat jako $\frac{1}{m} * (součet\ hodnot, které\ jsou\ závislé\ na\ w)$. Z tohoto důvodu lze následně softmax kernel aproximovat pomocí funkce ϕ.

Přesný popis FAVOR+ algoritmu a další jeho specifikace jsou dostupné v článku „Rethinking Attention With Performers“ (Choromanski et al., 2020).

Fine tuning

Fine tuning je proces, při kterém je již natrénovaná neuronová síť dotrénována na nové sadě dat. Toto dotrénování může probíhat na celé neuronové sítí, případně jsou váhy určitých vrstev zafixovány (nemění se v průběhu tréninku) a dotrénovává se pouze určitá část vrstev.

Fine tuning standardně probíhá tak, že se na natrénovanou neuronovou síť přidají nové vrstvy, které jsou následně trénovány na specifický úkol. Při tomto procesu jsou všechny původní váhy neuronové sítě zafixovány a mění se váhy pouze v dodatečně přidaných vrstvách, nebo je určitý počet vrchních vrstev původního modelu trénován společně s přidanými vrstvami.

Analýza single cell dat s využitím transformerů

Transformery a architektury od nich odvozené (performery) se ukazují jako velice efektivní architektura pro analýzu scRNA-seq dat, jelikož jsou schopny zpracovat multidimenzionální data, aniž by na ně byla využita dimenzionální redukce (Yang et al., 2022) a díky tomu jsou schopny uchovat všechny informace obsažené v datasetu, jako je například mezigenová interakce. Dalším velkým přínosem je, že transformery jsou schopny zachytit závislosti genů na sobě, a to díky jejich attention mechanismu (Cui et al., 2023; Yang et al., 2022).

Zkoumané modely v této práci jsou scBERT, scGPT a TOSICA. Všechny tyto modely jsou založeny na architektuře transformerů, ale každý sílu této architektury používá jiným způsobem. Odkazy na zdrojové kódy modelů jsou dostupné v apendixu práce.

Dodatečné informace ke všem datasetům, které budou v textu zmíněny jsou dostupné v apendixu práce (Tabulka 2, Tabulka 3, Tabulka 4).

scBERT

ScBERT je jazykový model určený pro anotaci buněk na základě scRNA-seq dat. Model využívá architekturu performerů. Díky této architektuře dokáže analyzovat data v rozsahu několika tisíců sloupců, v této souvislosti pak nemusí docházet k selekci HVGs a dimenzionální redukci (Yang et al., 2022).

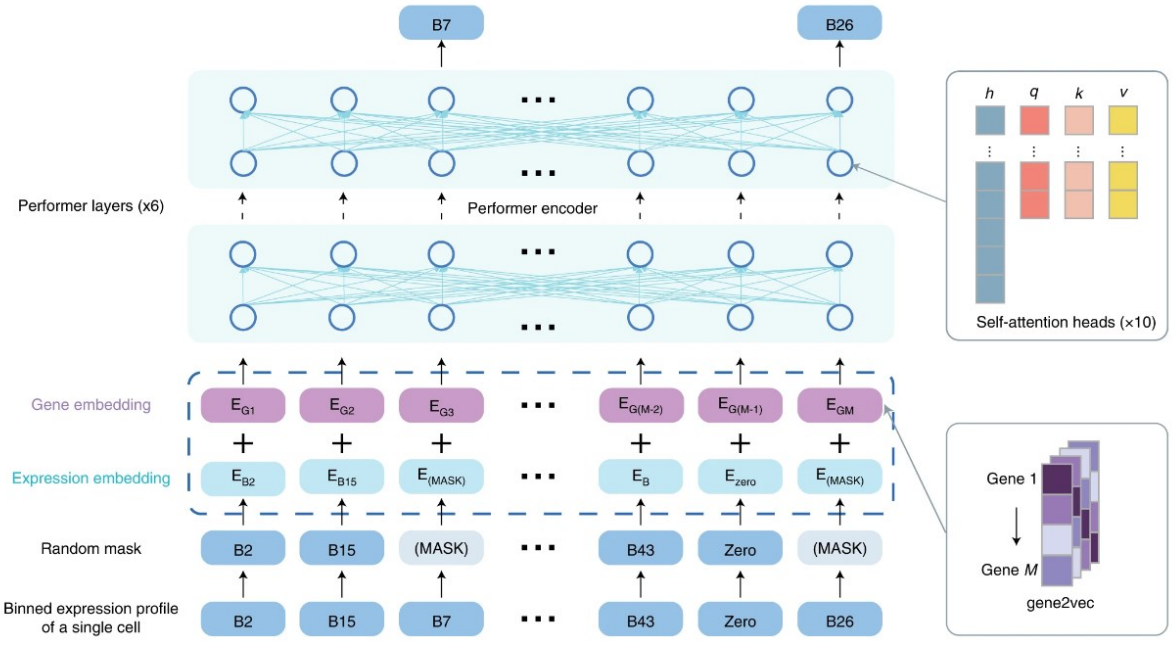
ScBERT je založený na stejné myšlence, jako dříve vyvinutý jazykový model BERT (Devlin et al., 2018; Yang et al., 2022). To znamená, že model je předem předtrénován za pomoci maskování unlabeled dat (unsupervised learning) a následně je využit supervised fine tuning ke specializaci modelu. Předtrénovaný model i kompletní kód je dostupný na GitHubu autorů.

Předtrénování modelu

Model je primárně předtrénován na několika milionech unlabeled scRNA-seq dat, kde mají zastoupení různé typy buněk. Zásadní je rozdíl dat, která analyzuje BERT a která analyzuje scBERT. BERT dostává na vstup diskrétní proměnou (slovo), zatímco scBERT dostává na vstupu spojitou veličinu (genovou expresi – frekvenci). Řešením je tyto spojité veličiny rozdělit do intervalů. Tato metoda se nazývá binning. Jednotlivé exprese genů jsou umístěny do jednoho intervalu, ve kterém leží velikost exprese daného genu. Tímto krokem je genová exprese převedena do diskrétní veličiny (Yang et al., 2022).

U genové exprese nezáleží na pořadí genů (na rozdíl od věty), a proto jsou na začátku předtrénování modelu jednotlivé geny náhodně přeskupeny (Cui et al., 2023; Yang et al., 2022). Určitá část expresí je následně zamaskována (od modelu se očekává doplnění maskované exprese). K takto připraveným datům jsou následně přičteny dva typy embeddingů (gene embedding a expression embedding). Gene embedding je získán za pomoci techniky gene2vec (každý gen má unikátní embedding). Tento embedding využívá model k zachycení sémantické podobnosti z hlediska koexprese genů (Du et al., 2019; Yang et al., 2022).

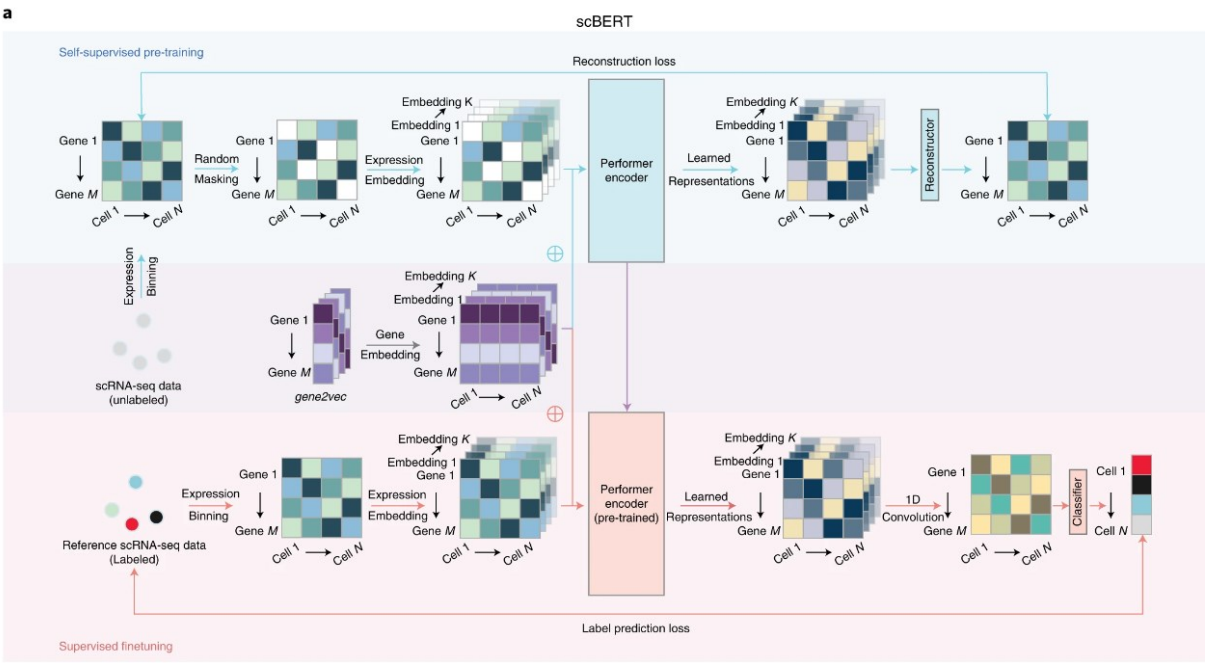
Druhý embedding zachycuje informaci o genové expresi a je generován modelem v závislosti na intervalu, ve kterém genová exprese daného genu leží. Takto předzpracovaná data vstupují do vrstev performerů (v základní verzi autoři využili 6 vrstev, tento parametr je ale ve výsledném kódu nastavitelný). Každý performer obsahuje 10 self-attention heads, které zachycují vztahy mezi jednotlivými geny. Výstupem těchto vrstev je genová exprese buňky, kde jsou doplněny genové exprese, které byly zamaskovány (Obrázek 10). Autoři tvrdí, že tento styl předtrénování performerů vede k umožnění modelu naučit se a pochopit syntax mezigenových interakcí, což výrazně přispívá k eliminaci batch efektu při následné klasifikaci (Yang et al., 2022).



Obrázek 10 Diagram předtrénování modelu scBERT. Genová exprese je pomocí binningu převedena do diskrétní veličiny. Následně jsou některé intervaly zamaskovány a je přičten gene a expression embedding. Takto předzpracovaná data vstupují do performerů. Výstupem je odmaskovaná exprese buňky. Obrázek převzat z článku „scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data“ (Yang et al., 2022).

Fine tuning pro anotaci buněk

Fine tuning předtrénovaného modelu probíhá za pomocí labeled datasetu, kde ke každé genové expresi je znám typ buňky. Na vrstvu performerů je přidána jedna jednodimenzionální konvoluční vrstva a klasifikátor (Obrázek 11), které jsou následně dotrénovány ke generování predikce typu buňky, a to na základě její genové exprese. Předzpracování labeled scRNA-seq dat probíhá identicky, pouze je vynechán krok maskování. Fine tuning probíhá pouze u konvoluční vrstvy na klasifikátoru (parametry performerů jsou zmraženy).

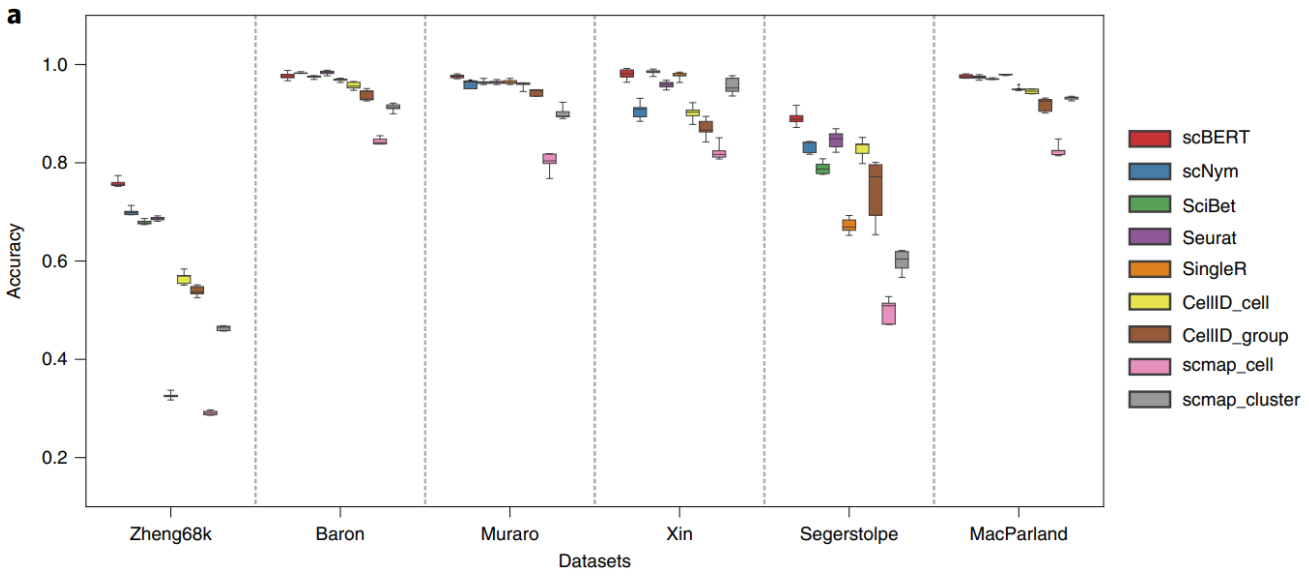


Obrázek 11 Diagram průběhu tréninku (předtrénink + fine tuning) modelu scBERT. Na model je po předtrénování přidána konvoluční vrstva a klasifikátor, na kterých následně probíhá fine tuning, aby model byl schopen anotovat buňky. Obrázek převzat z článku „scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data“ (Yang et al., 2022).

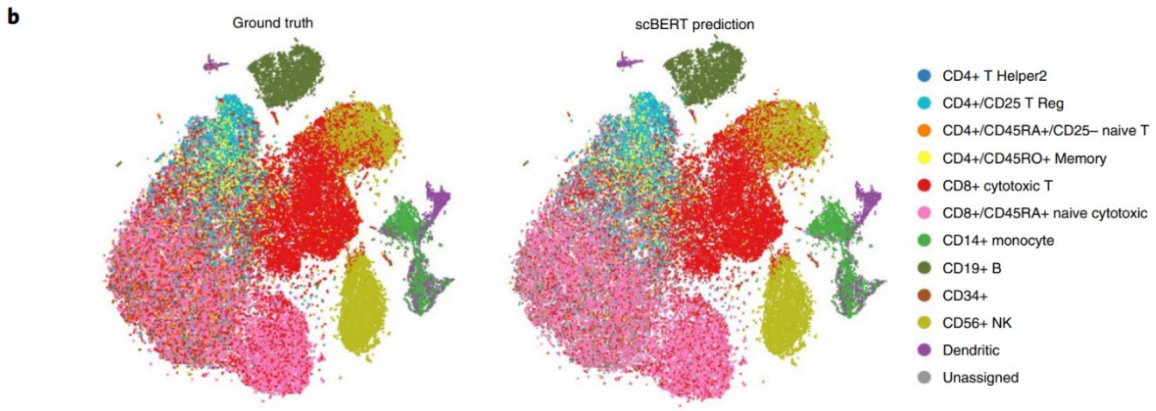
Přesnost modelu

Model byl otestován na řadě datasetů (Zheng68k, Baron, Muraro, Xin, Segerstolpe a MacParland – bližší informace k datasetům obsahuje Tabulka 2) a následně porovnán se standardními nástroji pro anotaci buněk, jako je například Seurat, SciBet a CellID_cell (Obrázek 14). ScBERT se řadí mezi nástroje s největší přesností (Obrázek 12), čímž jeho autoři potvrdili, že jsou transformery skutečně schopny přizpůsobit se scRNA-seq datům a úspěšně je analyzovat. Významnou předností využití transformerů je to, že model je schopný předjímat, s jakou pravděpodobností je anotace buňky správná. Pokud je pravděpodobnost nižší než 50 %, tak model buňku označí za „novel type“. Model je tedy schopen také najít a identifikovat dosud jím nepozorované buňky, na které nebyl trénován.

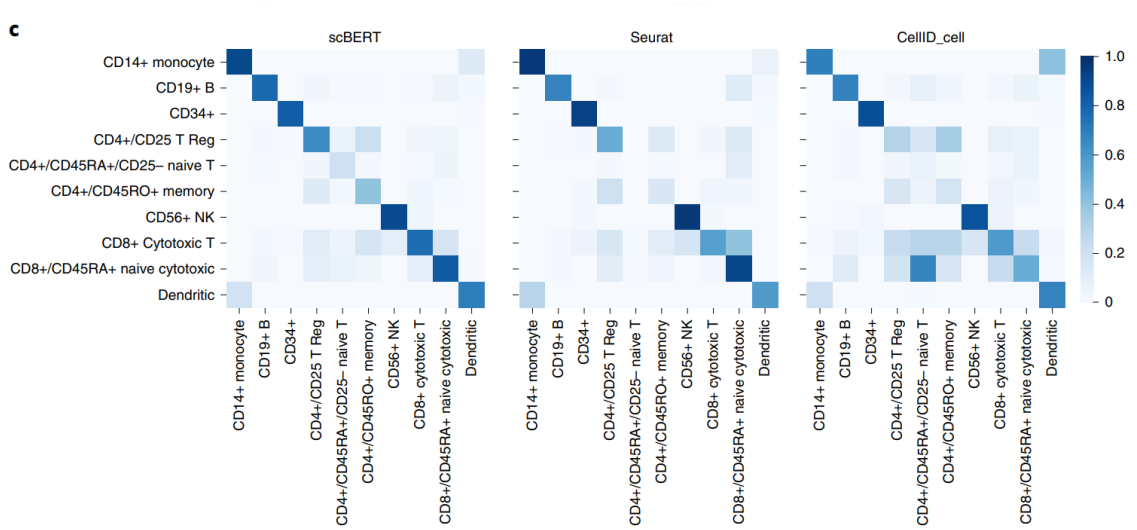
Autoři také nastínili možnost interpretace vnitřních parametrů, kde ukázali, že geny, na kterých má model vysokou attention, jsou z velké části již identifikované marker geny pro dané typy buněk, aniž by tato informace byla modelu jakkoliv uměle dodána (Yang et al., 2022).



Obrázek 12 Graf přesnosti všech testovaných nástrojů pro anotaci buněk na různých datasetech. scBERT se vždy řadil mezi nejlepší nástroje. Obrázek převzat z článku „scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data“ (Yang et al., 2022).



Obrázek 13 Příklad predikce modelu scBERT na datasetu Zheng68k. Obrázek převzat z článku „scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data“ (Yang et al., 2022).



Obrázek 14 Porovnání accuracy nástrojů scBERT, Saurat a CellID_cell na datasetu Zheng68k. Obrázek převzat z článku „scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data“ (Yang et al., 2022).

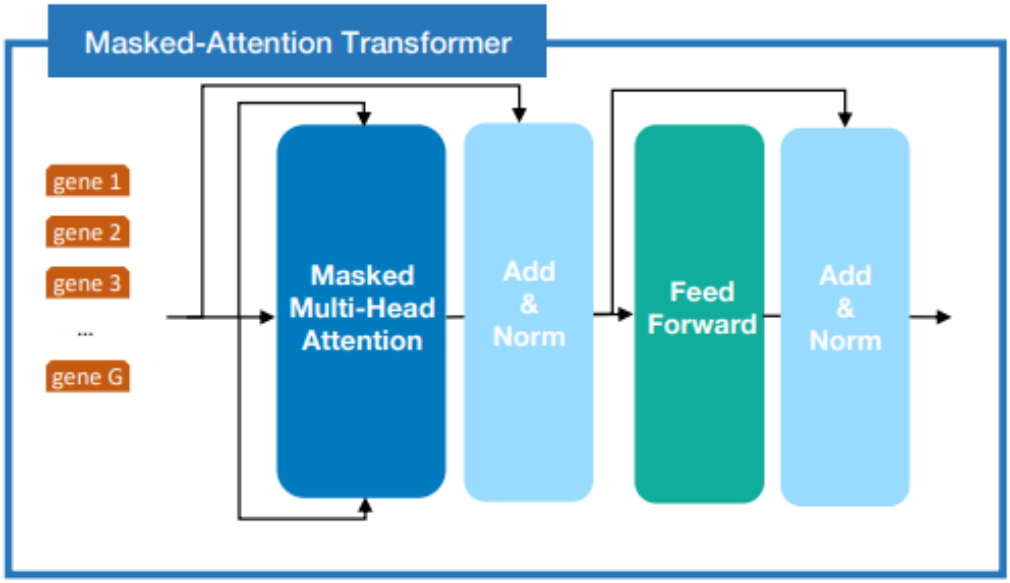
scGPT

Model scGPT (Single-cell Generative Pre-trained Transformer) je model založený na konceptu generativního předtrénování, které je velmi úspěšně využíváno například při trénování modelů na zpracování přirozeného jazyka (NLP), na single-cell data (Cui et al., 2023).

Model scGPT, popsaný Haotianem Cui et al., je model založený na architektuře transformerů, který je předtrénovaný na více než 10 milionech buněk. Model využívá podobného přístupu jako scBERT, kde genová exprese buňky je brána jako věta. U tohoto modelu ovšem dochází k selekci HVGs, tudíž nemusí být využity performery, ale transformery jsou dostačující architekturou.

Popis modelu

Model využívá embeddingu, který má 512 dimenzí. Model je složený z 12 transformerů, uspořádaných za sebou, kde každá obsahuje 8 attention heads a FFN obsahuje jednu vrstvu o velikosti 512 neuronů. Transformer využívaný v scGPT je založen na decoderu popsaném v původní implementaci transformerů (Obrázek 15) v článku „Attention is all you need“ (Vaswani et al., 2017).



Obrázek 15 Diagram Masked-attention transformeru využitém v modelu scGPT. Obrázek převzat z článku „scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI 2“ (Cui et al., 2023).

Předtrénování modelu

Předtrénování scGPT je zásadním krokem k tomu, aby byl tento model schopen zachytit komplexní biologické vzory a interakce, které jsou v scRNA-seq datech. Model využívá tzv. generativního přístupu, což je hlavní rozdíl od modelu scBERT, který k předtréninku využíval maskování částí genové exprese buňky a následné predikce této exprese (Cui et al., 2023; Yang et al., 2022). Tento způsob předtréninku se dá najít u standardních modelů pro zpracování přirozeného jazyka jako je BERT nebo například Roberta (Cui et al., 2023; Devlin et al., 2018; Yang et al., 2022).

Generativní předtrénování bylo využito s velkým úspěchem u modelů jako OpenAI GPT3 a GPT4 (Cui et al., 2023; OpenAI et al., 2023). Toto předtrénování spočívá v myšlence, že model se snaží předpovědět další nejpravděpodobnější token ze vstupu za předpokladu, že zná všechny vstupní tokeny před ním, což je přímé využití masked attention popsané v původní implementaci transformerů (Obrázek 15) (Cui et al., 2023; Vaswani et al., 2017).

Modely GPT (Generative pre-trained transformer) se řadí mezi tzv. decoder-only modely. To znamená, že vstupní data jsou okamžitě vkládána do decoderu, aniž by byl jakýmkoliv způsobem využíván encoder pro zpracování více abstraktní reprezentace textu.

Autoři modelu scGPT využili generativní předtrénink na dvě úlohy. První úlohou je generování neznámé genové exprese genu na základě známých expresí, což nazvali generování na základě

„gene prompt“. Druhou úlohou bylo generování kompletní genové exprese na základě popisu buňky – generování na základě „cell prompt“ (Cui et al., 2023).

Data jsou před tréninkem předzpracována. Předzpracování dat spočívá v normalizaci a následným binningem genových expresí (podobně jako u modelu scBERT). Ke genům je pak následně přidán expression embedding, který zachycuje míru exprese genu v buňce (Cui et al., 2023).

Autoři zveřejnili předtrénované modely scGPT jak pro celý lidský organismus, tak i specificky předtrénované modely pro různé lidské orgány (mozek, srdce, játra, plíce, krev). Tímto se výrazně usnadnila práce s modelem.

Cell prompt

Cell prompt je takový vstup, který popisuje charakteristiku nebo stav buňky, jako vstup do modelu scGPT, který následně generuje celý profil genové exprese této buňky. Tento přístup by mohl být zvláště užitečný v případech, kde je třeba předpovědět, jak by buňka určitého typu nebo v určitém stavu mohla exprimovat všechny své geny, i když přímá měření všech těchto genů nejsou k dispozici (Cui et al., 2023).

Gene prompt

Gene prompt se zaměřuje na predikci exprese specifických genů, a to na základě známé exprese jiných genů ve stejné buňce, nebo sadě buněk. Gene prompts mohou být například využity k odvození aktivity genů, které je obtížné nebo nákladné přímo měřit, případně k předpovědi, jak by změny v expresi určitých genů mohly ovlivnit expresi jiných (Cui et al., 2023).

Fine tuning

ScGPT není model určený čistě k buněčné anotaci. Jeho předtrénink umožňuje dle autorů model využívat i k úlohám, jako je například korekce batch efektu v datech, nebo predikce genetických poruch (Cui et al., 2023).

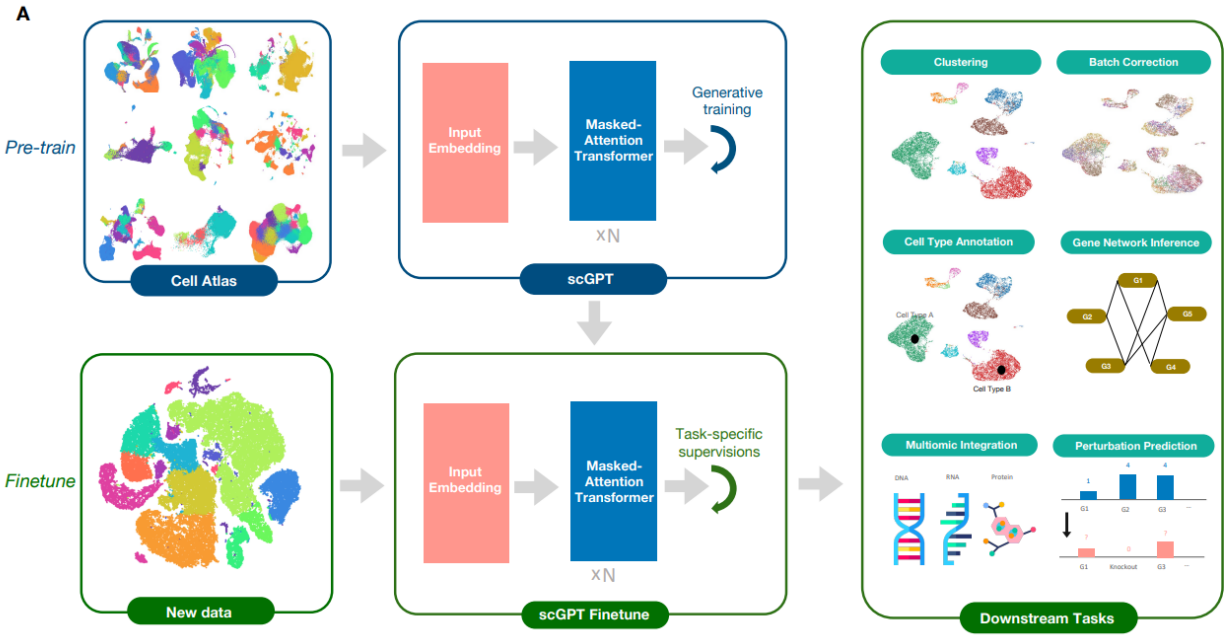
Autoři vytvořili pipelines pro uživatele, které výrazně usnadňují fine tuning modelu na specifický úkol na datech uživatele. Tyto pipelines jsou dostupné na GitHubu autorů.

Fine tuning pro anotaci buněk

Při fine tuningu modelu pro anotaci buněk je na scGPT model přidán Cell Type Classification (CLS) modul. CLS modul je autory navržen tak, aby využíval naučené reprezentace buněk

pomocí předtrénovaného scGPT, pro anotaci na základě profilů genové exprese (Cui et al., 2023). Jedná se o Multi-layer Perceptron (MLP) klasifikátor, který je optimalizován na základě cross entropy loss mezi predikovanými cell type pravděpodobnostmi a správnými anotacemi (ground-truth).

Samotný fine tuning probíhá vždy na labeled datasetu, kde má každá buňka k sobě přiřazený svůj typ. Během fine tuningu se model naučí spojovat určité specifické vzorce genové exprese s konkrétními typy buněk, přičemž využívá předtrénovaného těla scGPT (Cui et al., 2023).

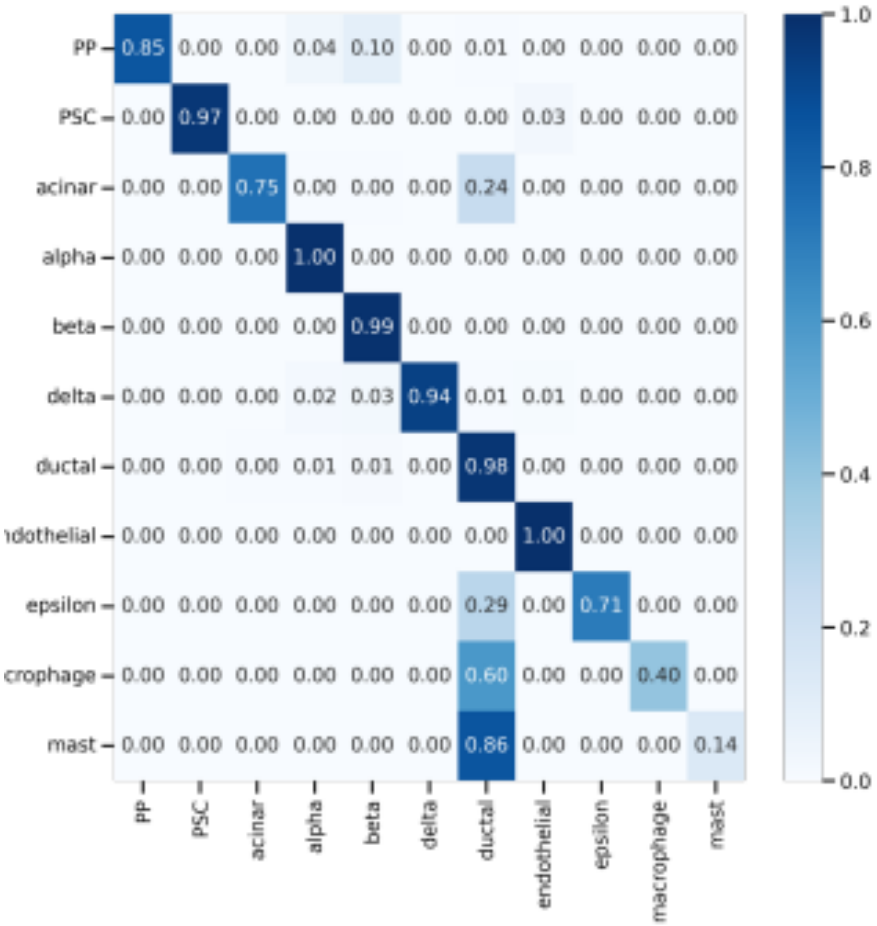


Obrázek 16 Diagram znázorňující předtrénink, fine tuning a následné využití modelu. Model je předtrénován na přibližně 10 milionech buněk. Fine tuning probíhá vždy specificky podle úlohy, na kterou chceme scGPT využít. Model je využitelný například pro anotaci buněk a clusterování. Obrázek převzat z článku „scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI 2“ (Cui et al., 2023).

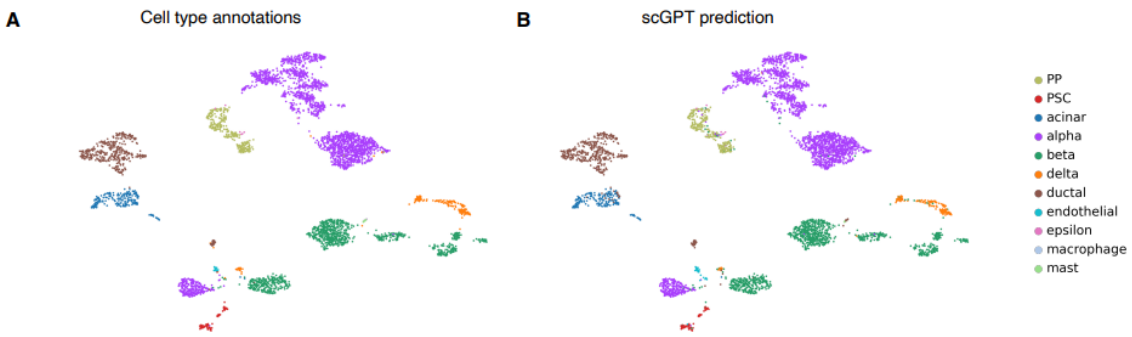
Přesnost modelu

Model scGPT byl otestován na hPancreas datasetu (Tabulka 3), který obsahuje expresi lidských buněk z pankreasu. Model predikoval anotace buněk s 96.7 % přesností, což je výrazně vyšší úspěšnost, než které dosáhly modely scBERT a TOSICA. Byl úspěšný i v predikci nad jinými

datasety, ve kterých opět překonal zmíněné modely v accuracy, precision, recall i v macroF1 (Obrázek 23, Obrázek 24) (Chen et al., 2023; Cui et al., 2023).



Obrázek 17 Normalizovaná confusion matice predikcí scGPT na datasetu hPancreas. Obrázek převzat z článku „scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI 2“ (Cui et al., 2023).



Obrázek 18 UMAP vizualizace buněk hPancreas datasetu Ground-truth vs scGPT prediction. scGPT predikuje s 96.7 % úspěšností. Obrázek převzat z článku „scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI 2“ (Cui et al., 2023).

scGPT je velmi silným nástrojem pro analýzu scRNA-seq dat, jelikož je nejen schopný anotovat buňky, ale je i využitelný v mnoha dalších úlohách, které jsou pro práci se single-cell daty nutností. scGPT potvrzuje, že generativní přístup je využitelný nejen pro zpracování přirozeného textu, ale i pro single cell a dává velmi dobrý základ pro budoucí modely založené na transformerech s generativním přístupem (decoder-only architecture) (Cui et al., 2023).

TOSICA

TOSICA je model založený na architektuře transformerů, který je přímo určen pro anotaci buněk na základě scRNA-seq dat (Chen et al., 2023). Podobně jako u scGPT dochází k selekci HVGs. Model je dostupný na GitHubu autorů.

Popis modelu

Model je složený z tří hlavních částí: Cell embedding vrstvy, Multi-head self-attention vrstvy a Cell-Type klasifikátoru (Chen et al., 2023).

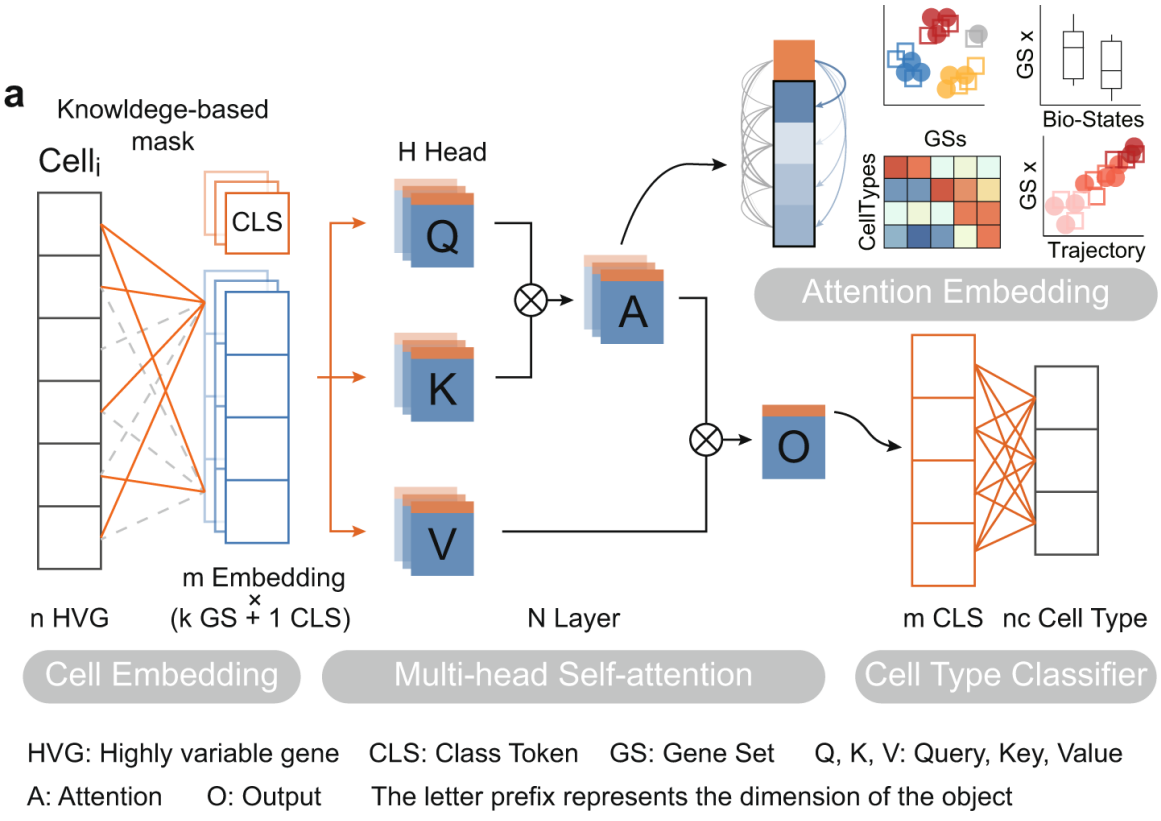
Cell embedding je proces, kde TOSICA transformuje genovou expresi buňky do sady tokenů, kde každý token reprezentuje skupinu genů organizovaných často podle biologických drah nebo regulonů[8], což umožňuje modelu zachytávat biologicky smysluplné vzorce, namísto soustředění se na individuální genové exprese. Tato transformace je standardní lineární projekce, kde jsou genové exprese násobeny maticí vah, čímž vysokodimenzionální data genové exprese přesouvá do prostoru s nižší dimenzí. Aby bylo zajištěno, že každý token zachycuje specifické biologické informace (dráhu nebo regulon), je během gene-token konverze aplikována tzv. knowledge-based matice[9], která vynuluje spoje v matici vah, jež neodpovídají předem definovaným gene-pathway asociacím (Chen et al., 2023). Tento proces zajistí, že každý token přesně reprezentuje odlišný biologický koncept, jelikož dostane informaci o genech pouze ze specifické dráhy. K takto vytvořenému embeddingu je následně přidán speciální token, který autoři nazývají Class token (CLS). CLS je trénovatelný parametr, který slouží jako shrnutí reprezentace celkového profilu genové exprese buňky. CLS následně hraje zásadní roli v mechanismu self-attention modelu (Obrázek 19), kde usnadňuje agregaci relevantních signálů pro cell-type klasifikaci (Chen et al., 2023).

[8] Regulon je označení pro skupinu genů, které jsou regulovány jako jednotka. Obecně je tato jednotka řízená stejným regulačním genem, který exprimuje protein působící jako represor nebo aktivátor.
[9] Knowledge-based matice, která se stará o zachycení regulonů a biologických drah do různých tokenů, není trénovatelný parametr. Tato matice je modelu předem dodána a model nemění hodnoty v ní uložené (expertní vstup).

Takto zpracovaný embedding genové exprese následně vchází do multi-head attention mechanismu (Obrázek 19), který se drží všech základních principů popsaných v sekci věnované původní architektuře transformerů, jenž byla představena v článku *„Attention is all you need“* (Chen et al., 2023; Turner, 2023; Vaswani et al., 2017).

Výsledná output matice O (matice získána vynásobením attention matice A value maticí V) není využita pro klasifikaci celá, ale je použita pouze její CLS část (Obrázek 19). Myšlenkou je, že CLS část v output matici po průchodu multi-head attention vrstvou, získala informaci o mnoha buněčných drahách, a proto pro klasifikaci stačí využít pouze CLS, která je převedena do vektoru obsahující pravděpodobnosti popisující o jaký typ buňky se jedná (Chen et al., 2023).

TOSICA je dále schopna analyzovat attention mezi CLS a ostatními tokeny (tokenky popisující různé dráhy nebo regulony). Tyto hodnoty autoři nazvali attention embedding a je možné ho využívat k další analýze scRNA-seq dat, jako je například analýza trajektorie buňky, a ne pouze ke klasifikaci buněk (Chen et al., 2023).



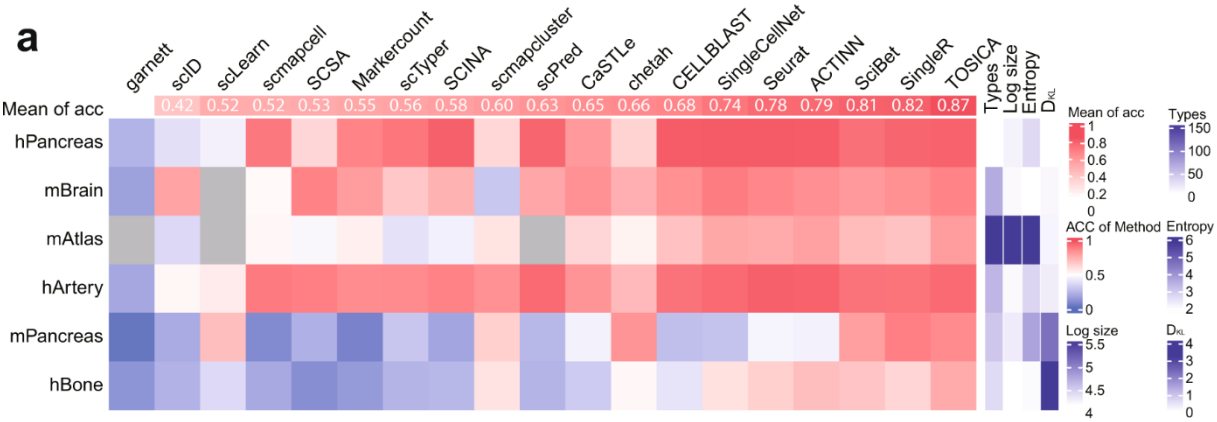
Obrázek 19 Diagram modelu TOSICA. Na vstupní data (HVGs buňky) je aplikována knowledge-based maska, která vytváří embedding, ke kterému je přidán CLS token. Takto předzpracovaná data následně vstupují do multi-head attention mechanismu. Pro anotaci buněk je využita pouze CLS část output matice (O). Attention matice (A) je dále využitelná pro specifickou analýzu genové exprese buňky. Obrázek je převzat z článku „Transformer for one stop interpretable cell type annotation“ (Chen et al., 2023).

Trénink modelu

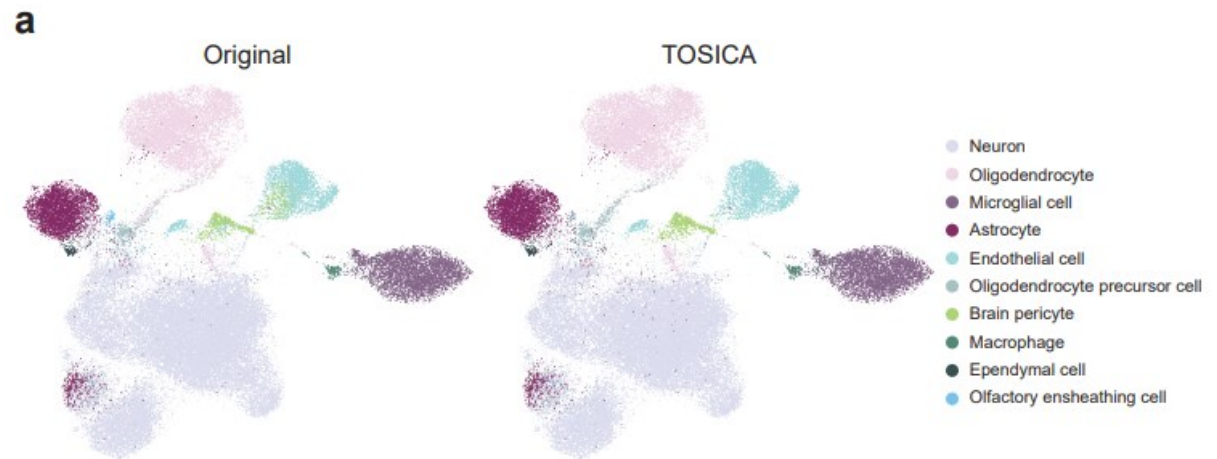
Model není na rozdíl od modelů scBERT nebo scGPT žádným způsobem předtrénováván, ale je trénován přímo s pomocí supervised learning na labeled datasetu (Chen et al., 2023; Cui et al., 2023; Yang et al., 2022). Jedná se tedy o výrazně jednodušší model, který anotuje buňky s velmi dobrou přesností, která je konkurenceschopná s modely scBERT a scGPT. Jedná se o uživatelsky jednodušší a výrazně rychlejší způsob tréninku, než tomu bylo u předchozích modelů (Chen et al., 2023; Cui et al., 2023).

Přesnost modelu

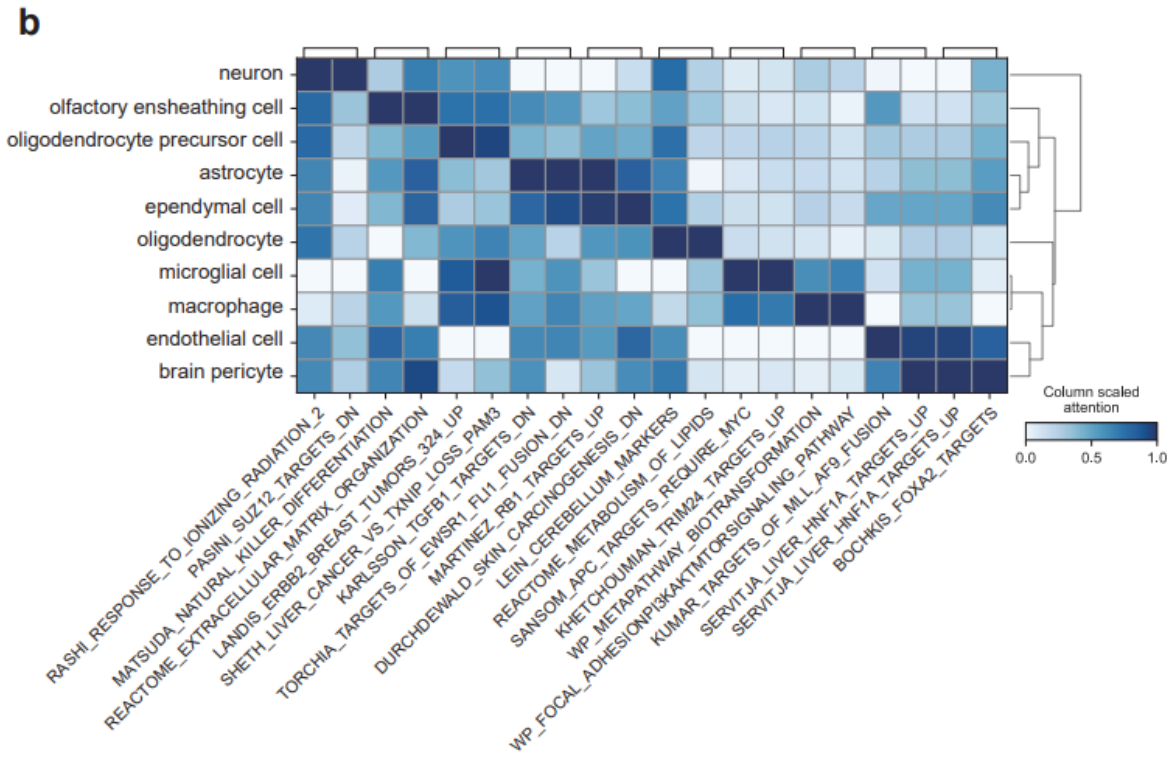
Model byl testován na šesti různých datasetech (hArtery, hBone, hPancreas, mBrain, mPancreas, mAtlas – bližší informace k datasetům obsahuje Tabulka 2) a byl porovnán s dalšími 18 nástroji pro anotaci buněk. TOSICA se vždy umístila mezi nejlepšími nástroji pro anotaci buněk. S průměrnou accuracy 86.69 % se stala nejlepším z 19 nástrojů (Obrázek 20).



Obrázek 20 Graf accuracy všech nástrojů, proti kterým byl model TOSICA porovnáván. Sloupce jsou seřazeny podle průměrné přesnosti každé metody na všech datech (nahoře). Počet typů buněk (Types), počet buněk (Log size), Shannon-Entropie (Entropy) v referenci a Kullback-Leiblerova divergence (DKL) mezi reference a query jsou označeny vpravo. Šedá znamená, že tento dataset je pro tuto metodu příliš velký, aby s ním mohla pracovat. Obrázek je převzat z článku „Transformer for one stop interpretable cell type annotation“ (Chen et al., 2023).



Obrázek 21 UMAP reprezentace mBrain datasetu a predikovaných buněčných typů pomocí TOSICA modelu. Obrázek je převzat z článku „Transformer for one stop interpretable cell type annotation“ (Chen et al., 2023).

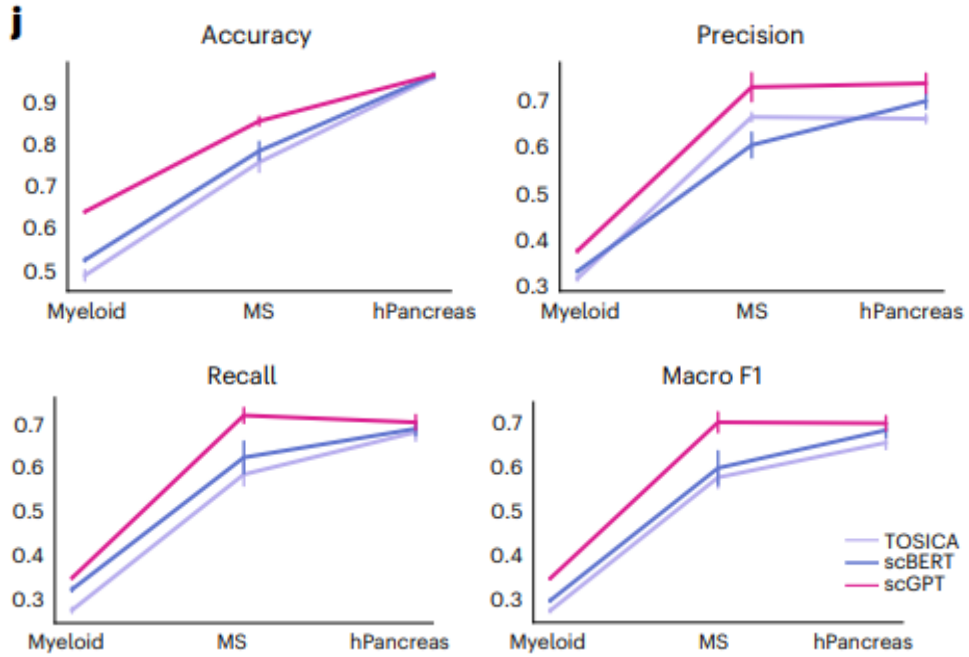


Obrázek 22 Graf attention mezi jednotlivými regulačními drahami a buněčnými typy. Graf ukazuje, jak jsou různé dráhy důležité pro klasifikaci buněk. Čím větší attention na dráze, tím důležitější dráha pro daný typ buňky je. Obrázek je převzat z článku „Transformer for one stop interpretable cell type annotation“ (Chen et al., 2023).

Model TOSICA je výrazně jednodušší než dříve zmíněné modely, ale i přes jeho jednoduchost je TOSICA schopný porazit většinu standardních metod dnes využívaných pro anotaci buněk. Toto ukazuje, že aplikace transformerů na scRNA-seq data je efektivní. TOSICA ale zaostává ve srovnání s modelem scGPT. Výsledky ovšem ukazují, že model TOSICA je na určitých datasetech lepší než model scBERT (Obrázek 23, Obrázek 24)

Srovnání modelů

Autoři modelu scGPT srovnali svůj vlastní model s modely scBERT, scGPT a TOSICA na datasetech myeloid, MS a hPancreas. Výsledky ukazují, že scGPT je nejlepší na všech testovaných datasetech (Obrázek 23, Obrázek 24) (Cui et al., 2023).



Obrázek 23 Výsledky srovnání modelů scGPT, TOSICA a scBERT, které bylo provedeno autory modelu scGPT. Obrázek převzat z článku „scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI 2“ (Cui et al., 2023).

Dataset	Model	Classification Metrics			
		Accuracy	*Precision*	*Recall*	*MacroF1*
Myeloid	scGPT (fine-tuned)	**0.642**	**0.366**	**0.347**	**0.346**
	scGPT (from-scratch)	0.606	0.304	0.339	0.309
	TOSICA	0.488	0.316	0.276	0.275
	scBert	0.525	0.331	0.323	0.298
Multiple Sclerosis	scGPT (fine-tuned)	**0.856**	**0.729**	**0.720**	**0.703**
	scGPT (from-scratch)	0.798	0.660	0.623	0.600
	TOSICA	0.758	0.664	0.585	0.578
	scBert	0.785	0.604	0.624	0.599
hPancreas	scGPT (fine-tuned)	**0.968**	**0.735**	**0.725**	**0.718**
	scGPT (from-scratch)	0.936	0.665	0.668	0.622
	TOSICA	0.960	0.661	0.681	0.656
	scBert	0.964	0.699	0.689	0.685

Obrázek 24 Tabulka výsledků srovnání modelů scGPT, TOSICA a scBERT na datasetech Myeloid, Multiple Sclerosis a hPancreas. Obrázek převzat z článku „scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI 2“ (Cui et al., 2023).

Závěr

Modely založené na architektuře transformerů se ukazují jako velice silným nástrojem pro anotaci buněk na základě scRNA-seq dat. Všechny tři popsané modely (scBERT, scGPT, TOSICA) ukázaly lepší či srovnatelnou úspěšnost než metody standardní, které nejsou založeny na této architektuře. Zároveň oproti standardním metodám nabízí větší pochopení mezigenových interakcí díky jejich attention mechanismu, který je schopen tyto interakce zachytit.

S rychlým vývojem nových architektur je ovšem možné, že se transformery brzy stanou zastaralou architekturou, která bude následně nahrazena. Architekturou, která transformery nahradí, by se v budoucnu mohla stát nová architektura MAMBA, která není založena na attention, ale je tzv. Selective-State-Spaces-based (SSM-based), což z ní dělá výrazně rychleji trénovatelný model s vysokou přesností (Gu & Dao, 2023).

Využitá literatura

Adil, A., Kumar, V., Jan, A. T., & Asger, M. (2021). Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis. In *Frontiers in Neuroscience* (Vol. 15). Frontiers Media S.A. https://doi.org/10.3389/fnins.2021.591122

Almeida, F., & Xexéo, G. (2019). *Word Embeddings: A Survey*. http://arxiv.org/abs/1901.09069

Alquicira-Hernandez, J., Sathe, A., Ji, H. P., Nguyen, Q., & Powell, J. E. (2019). ScPred: Accurate supervised method for cell-type classification from single-cell RNA-seq data. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1862-5

Alzubi, J., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. *Journal of Physics: Conference Series*, *1142*(1). https://doi.org/10.1088/1742-6596/1142/1/012012

Chen, J., Xu, H., Tao, W., Chen, Z., Zhao, Y., & Han, J. D. J. (2023). Transformer for one stop interpretable cell type annotation. *Nature Communications*, *14*(1). https://doi.org/10.1038/s41467-023-35923-4

Choromanski, K., Likhosherstov, V., Dohan, D., Song, X., Gane, A., Sarlos, T., Hawkins, P., Davis, J., Mohiuddin, A., Kaiser, L., Belanger, D., Colwell, L., & Weller, A. (2020). *Rethinking Attention with Performers*. http://arxiv.org/abs/2009.14794

Cui, H., Wang, C., Maan, H., & Wang, B. (2023). scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI 2. *BioRxiv*. https://doi.org/10.1101/2023.04.30.538439

de Kanter, J. K., Lijnzaad, P., Candelli, T., Margaritis, T., & Holstege, F. C. P. (2019). CHETAH: a selective, hierarchical cell type identification method for single-cell RNA sequencing. *Nucleic Acids Research*, *47*(16), E95. https://doi.org/10.1093/NAR/GKZ543

Devlin, J., Chang, M.-W., Lee, K., Google, K. T., & Language, A. I. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. https://github.com/tensorflow/tensor2tensor

Doshi, K. (2021, January 17). *Transformers Explained Visually (Part 3): Multi-head Attention, deep dive | by Ketan Doshi | Towards Data Science*.

https://towardsdatascience.com/transformers-explained-visually-part-3-multi-head-attention-deep-dive-1c1ff1024853 [citováno 15.4.2024]

Du, J., Jia, P., Dai, Y., Tao, C., Zhao, Z., & Zhi, D. (2019). Gene2vec: Distributed representation of genes based on co-expression. *BMC Genomics*, *20*. https://doi.org/10.1186/s12864-018-5370-x

Franzén, O., Gan, L. M., & Björkegren, J. L. M. (2019). PanglaoDB: A web server for exploration of mouse and human single-cell RNA sequencing data. *Database*, *2019*(1). https://doi.org/10.1093/database/baz046

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep Learning*. http://www.deeplearningbook.org

Gu, A., & Dao, T. (2023). *Mamba: Linear-Time Sequence Modeling with Selective State Spaces*. https://github.com/state-spaces/mamba.

Haghverdi, L., Lun, A. T. L., Morgan, M. D., & Marioni, J. C. (2018). Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors. *Nature Biotechnology*, *36*(5), 421–427. https://doi.org/10.1038/nbt.4091

Haque, A., Engel, J., Teichmann, S. A., & Lönnberg, T. (2017). A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications. In *Genome Medicine* (Vol. 9, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s13073-017-0467-4

Heumos, L., Schaar, A. C., Lance, C., Litinetskaya, A., Drost, F., Zappia, L., Lücken, M. D., Strobl, D. C., Henao, J., Curion, F., Aliee, H., Ansari, M., Badia-i-Mompel, P., Büttner, M., Dann, E., Dimitrov, D., Dony, L., Frishberg, A., He, D., … Theis, F. J. (2023). Best practices for single-cell analysis across modalities. *Nature Reviews Genetics*, *24*(8), 550–572. https://doi.org/10.1038/s41576-023-00586-w

Hu, C., Li, T., Xu, Y., Zhang, X., Li, F., Bai, J., Chen, J., Jiang, W., Yang, K., Ou, Q., Li, X., Wang, P., & Zhang, Y. (2023). CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. *Nucleic Acids Research*, *51*(D1), D870–D876. https://doi.org/10.1093/nar/gkac947

Huang, Q., Liu, Y., Du, Y., & Garmire, L. X. (2021). Evaluation of Cell Type Annotation R Packages on Single-cell RNA-seq Data. *Genomics, Proteomics and Bioinformatics*, *19*(2), 267–281. https://doi.org/10.1016/j.gpb.2020.07.004

Jovic, D., Liang, X., Zeng, H., Lin, L., Xu, F., & Luo, Y. (2022). Single-cell RNA sequencing technologies and applications: A brief overview. *Clinical and Translational Medicine*, *12*(3). https://doi.org/10.1002/ctm2.694

Khurana, D., Koli, A., Khatter, K., & Singh, S. (2023). Natural language processing: state of the art, current trends and challenges. *Multimedia Tools and Applications*, *82*(3), 3713–3744. https://doi.org/10.1007/s11042-022-13428-4

Kiselev, V. Y., & Hemberg, M. (2017). *scmap - A tool for unsupervised projection of single cell RNA-seq data*. https://doi.org/10.1101/150292

Kiselev, V. Y., Kirschner, K., Schaub, M. T., Andrews, T., Yiu, A., Chandra, T., Natarajan, K. N., Reik, W., Barahona, M., Green, A. R., & Hemberg, M. (2017). SC3: Consensus clustering of single-cell RNA-seq data. *Nature Methods*, *14*(5), 483–486. https://doi.org/10.1038/nmeth.4236

Kitaev, N., Kaiser, Ł., & Levskaya, A. (2020). *Reformer: The Efficient Transformer*. http://arxiv.org/abs/2001.04451

Korsunsky, I., Millard, N., Fan, J., Slowikowski, K., Zhang, F., Wei, K., Baglaenko, Y., Brenner, M., Loh, P. ru, & Raychaudhuri, S. (2019). Fast, sensitive and accurate integration of single-cell data with Harmony. *Nature Methods*, *16*(12), 1289–1296. https://doi.org/10.1038/s41592-019-0619-0

Lecun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539ï

Liu, Q., & Wu, Y. (2012). Supervised Learning. In *Encyclopedia of the Sciences of Learning* (pp. 3243–3245). Springer US. https://doi.org/10.1007/978-1-4419-1428-6_451

Luecken, M. D., & Theis, F. J. (2019). Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*, *15*(6). https://doi.org/10.15252/msb.20188746

Ma, F., & Pellegrini, M. (2020). ACTINN: Automated identification of cell types in single cell RNA sequencing. *Bioinformatics*, *36*(2), 533–538. https://doi.org/10.1093/bioinformatics/btz592

Mahesh, B. (2018). Machine Learning Algorithms-A Review. *International Journal of Science and Research*. https://doi.org/10.21275/ART20203995

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient Estimation of Word Representations in Vector Space*. http://arxiv.org/abs/1301.3781

Naeem, S., Ali, A., Anam, S., & Ahmed, M. M. (2023). An Unsupervised Machine Learning Algorithms: Comprehensive Review. *International Journal of Computing and Digital Systems*, *13*(1), 911–921. https://doi.org/10.12785/ijcds/130172

OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., … Zoph, B. (2023). *GPT-4 Technical Report*. http://arxiv.org/abs/2303.08774

Passmore, L. A., & Coller, J. (2022). Roles of mRNA poly(A) tails in regulation of eukaryotic gene expression. In *Nature Reviews Molecular Cell Biology* (Vol. 23, Issue 2, pp. 93–106). Nature Research. https://doi.org/10.1038/s41580-021-00417-y

Schmidhuber, J. (2014). *Deep Learning in Neural Networks: An Overview*. https://doi.org/10.1016/j.neunet.2014.09.003

Schmidt, R. M. (2019). *Recurrent Neural Networks (RNNs): A gentle Introduction and Overview*. http://arxiv.org/abs/1912.05911

Slomovic, S., Laufer, D., Geiger, D., & Schuster, G. (2006). Polyadenylation of ribosomal RNA in human cells. *Nucleic Acids Research*, *34*(10), 2966–2975. https://doi.org/10.1093/nar/gkl357

Su, M., Pan, T., Chen, Q. Z., Zhou, W. W., Gong, Y., Xu, G., Yan, H. Y., Li, S., Shi, Q. Z., Zhang, Y., He, X., Jiang, C. J., Fan, S. C., Li, X., Cairns, M. J., Wang, X., & Li, Y. S. (2022). Data analysis guidelines for single-cell RNA-seq in biomedical studies and clinical applications. In *Military Medical Research* (Vol. 9, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s40779-022-00434-8

Szubert, B., Cole, J. E., Monaco, C., & Drozdov, I. (2019). Structure-preserving visualisation of high dimensional single-cell datasets. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-45301-0

Tan, Y., & Cahan, P. (2019). SingleCellNet: A Computational Tool to Classify Single Cell RNA-Seq Data Across Platforms and Across Species. *Cell Systems*, *9*(2), 207-213.e2. https://doi.org/10.1016/j.cels.2019.06.004

Townes, F. W., Hicks, S. C., Aryee, M. J., & Irizarry, R. A. (2019). Feature selection and dimension reduction for single-cell RNA-Seq based on a multinomial model. *Genome Biology*, *20*(1). https://doi.org/10.1186/s13059-019-1861-6

Tsang, S.-H. (2022, November 26). *Brief Review — Rethinking Attention with Performers | by Sik-Ho Tsang | Medium*. https://sh-tsang.medium.com/brief-review-rethinking-attention-with-performers-e9fba834ab95 [citováno 26.4.2024]

Turner, R. E. (2023). *An Introduction to Transformers*. http://arxiv.org/abs/2304.10557

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need*. http://arxiv.org/abs/1706.03762

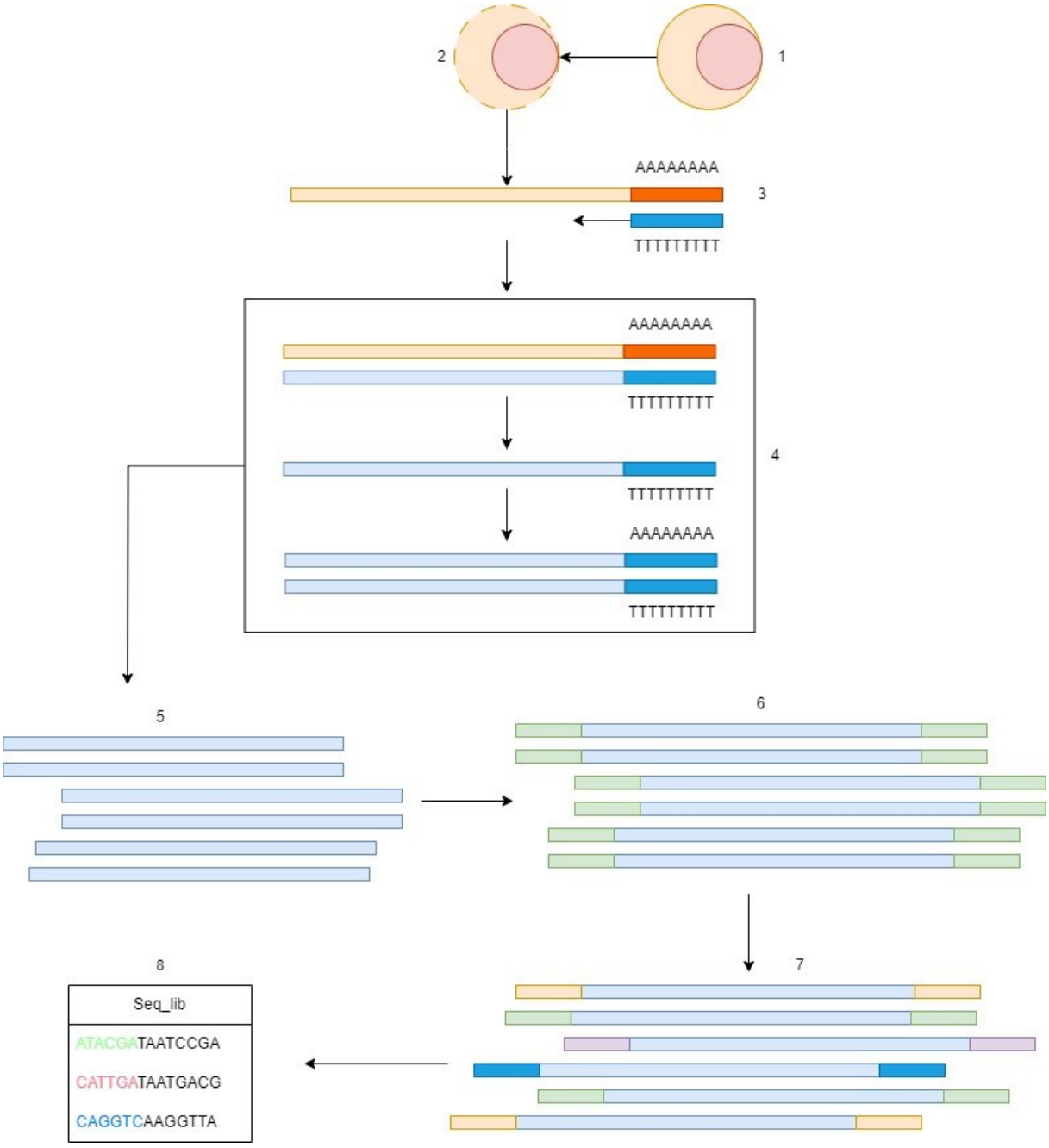
Yang, F., Wang, W., Wang, F., Fang, Y., Tang, D., Huang, J., Lu, H., & Yao, J. (2022). scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, *4*(10), 852–866. https://doi.org/10.1038/s42256-022-00534-z

Zhou, W. min, Yan, Y. yan, Guo, Q. ru, Ji, H., Wang, H., Xu, T. tian, Makabel, B., Pilarsky, C., He, G., Yu, X. yong, & Zhang, J. ye. (2021). Microfluidics applications for high-throughput single cell sequencing. In *Journal of Nanobiotechnology* (Vol. 19, Issue 1). BioMed Central Ltd. https://doi.org/10.1186/s12951-021-01045-6

Zlatanova, J. (2023). *Molecular Biology*. Garland Science. https://doi.org/10.1201/9781003132929

Apendix

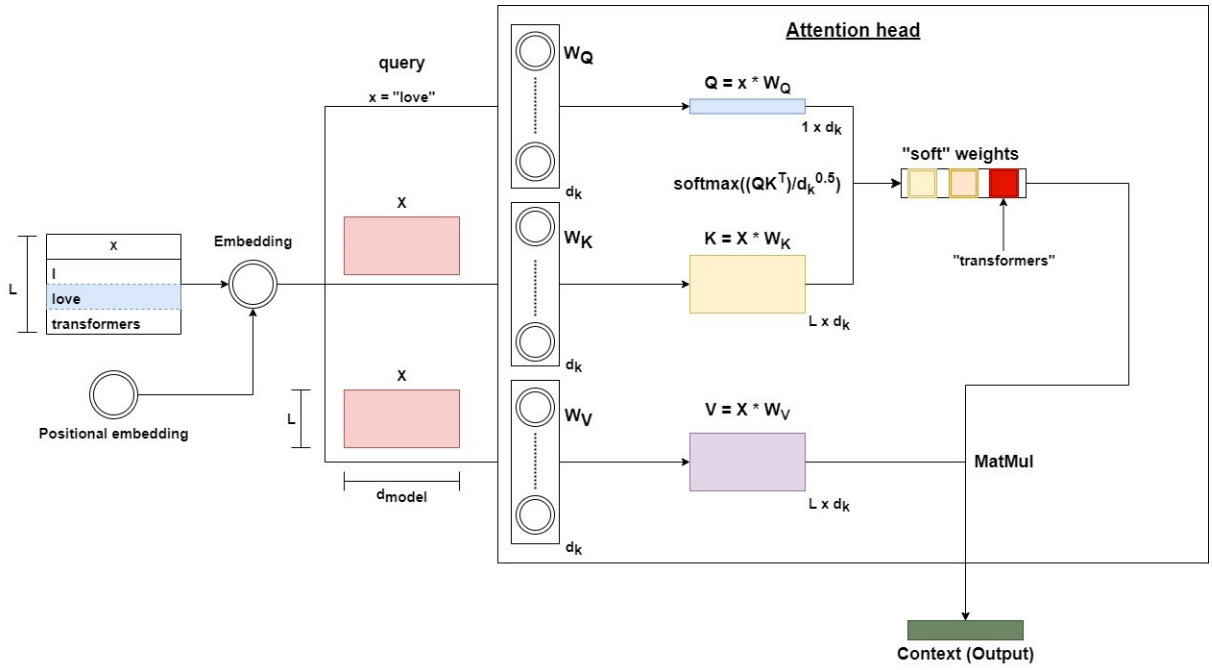
Diagram procesu zisku scRNA-seq dat



Obrázek 25 Diagram průběhu přípravy knihovny pro scRNA-seq. 1) Izolace buněk z tkáně. 2) Lyze buňky – lyze musí probíhat tak, aby nebyla porušena a znehodnocena mRNA. 3) mRNA zachycena poly(T) primerem. Poly(T) primer nasedá na mRNA na místě, kde se nachází poly(A) konec. 4) Konverze Poly(T)-primed mRNA na cDNA za pomocí reverzní transkripce 5) cDNA amplifikace (standardně za použití PCR) 6) barcode-tagging 7) spojení cDNA jednotlivých buněk do jednoho poolu. 8) Vytvoření knihovny sekvencí (NGS). Obrázek je inspirován obrázkem v článku „A practical guide to single-cell RNA-sequencing for biomedical research and clinical applications“ (Haque et al., 2017).

Příklad výpočtu attention

Uvažujme větu „I love transformers“. Výpočet self-attention pro slovo „love“ bude probíhat následovně (Obrázek 26):



Obrázek 26 Diagram výpočtu attention pro query = „love“. Vstupní embedding vstupuje do attention head, kde jsou spočítány hodnoty Q, K a V, na kterých jsou dále prováděny matematické operace, které vedou k získání výstupu (kontextu daného slova). Obrázek je inspirovaný popisem mechanismu attention v článku „Attention Is All You Need“ (Vaswani et al., 2017).

Věta je nejprve převedena do maticové reprezentace za pomoci embbeddingu a positional embeddingu. Slovo „love“ je v tomto příkladě použito jako query a počítáme pro něj jeho kontext. Slova (včetně slova „love“) ve větě jsou použita k výpočtu matice K (keys) a V (values).

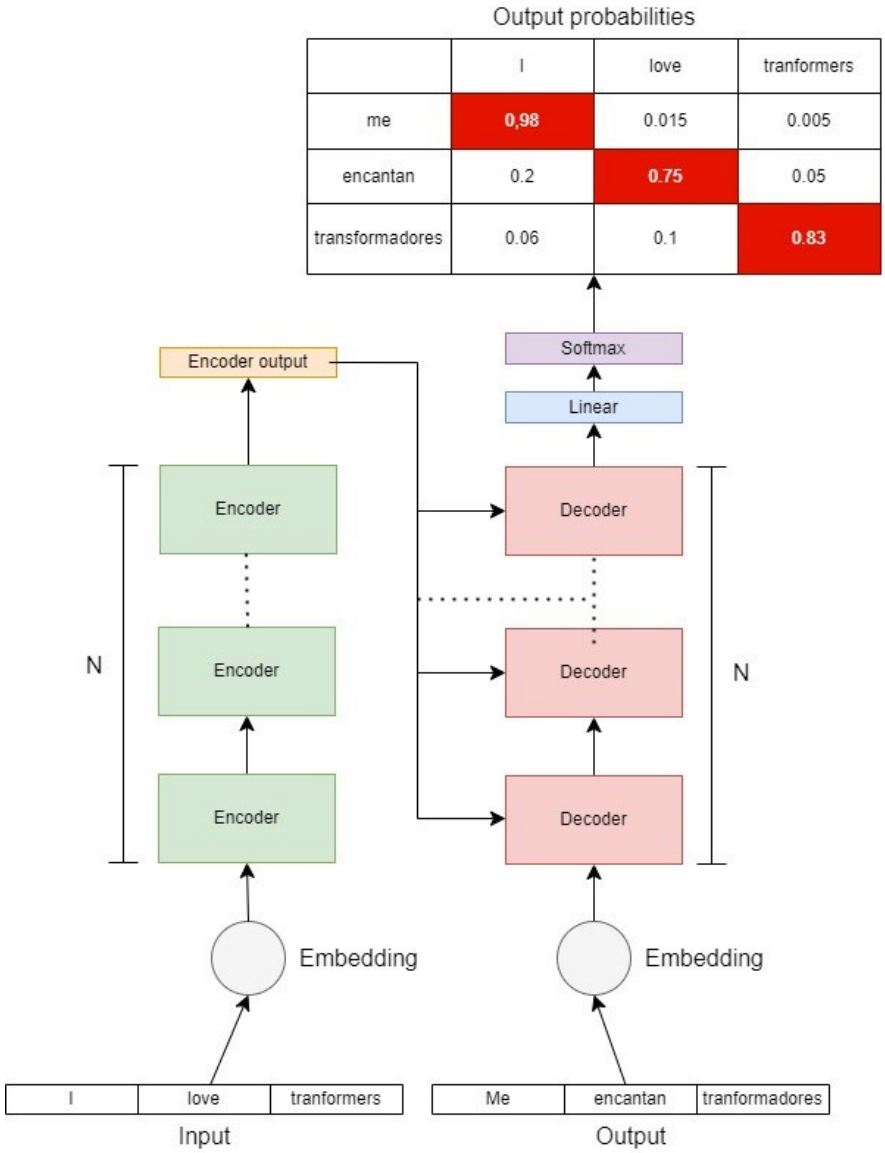
Query vektor pro slov „love“ je porovnán (pomocí skalárního součinu) s ostatními slovy v keys. To pomáhá modelu najít nejrelevantnější slovo pro naše query. V tomto případě vychází slovo „transformers“ jako nejrelevantnější. Na výsledné hodnoty aplikujeme softmax funkci, abychom získali vektor pravděpodobností (v našem případě velikosti 3), které se nasčítají na 1. Tento vektor poté vynásobíme maticí V (values), čímž výrazně zesílíme signál důležitých slov ve větě a snížíme signál u méně důležitých slov (Vaswani et al., 2017).

Po rozepsání celého výpočtu kontextu pro jedno slovo získáváme rovnici:

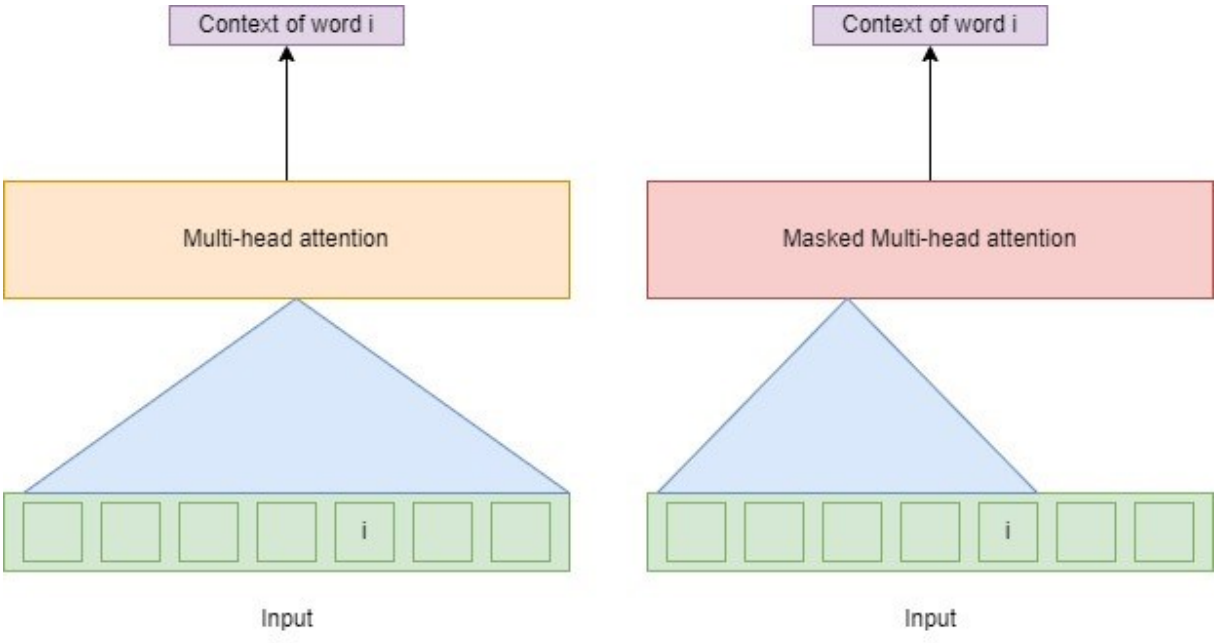
$$Context_x = softmax\left(\frac{xW_Q(XW_K)^T}{\sqrt{d_K}}\right)XW_V$$

Ze vzorce vyplývá, že není nutné počítat query postupně za sebou, ale jsme schopni počítat kontext více slov společně, jelikož kontext slova není ovlivněn kontextem jiných slov, ale pouze vstupem, který se v průběhu nemění.

Tato paralelizace výrazně zrychluje výpočet a díky ní jsou transformery podstatně výkonnější než rekurentní neuronové sítě (Turner, 2023).



Obrázek 27 Příklad výpočtu output pravděpodobností pro překlad věty "I love transformers". Výstupem je matice pravděpodobností (output probabilities), která říká, jak má být dané slovo přeloženo. Perfektní model by předpovídal hodnotu 1 pro správný překlad slova a 0 pro všechny ostatní. Tím bychom získali ihned one-hot matici předkladu věty.



Obrázek 28 Diagram masked multi-head attention. Attention se počítá pouze pro slovo samotné a pro slova, která se nachází před ním. Masked multi-head attention mechanismus je využíván v decoderu.

Důkaz č. 1

Důkaz je převzat z článku *„Rethinking Attention with Performers“*(Choromanski et al., 2020).

Pro všechny $x, y \in \mathbb{R}^d$ platí:

$$SM(x,y) = \exp(x^T y) = exp\left(-\frac{\|x\|^2}{2}\right) \cdot exp\left(\frac{\|x+y\|^2}{2}\right) \cdot exp\left(-\frac{\|y\|^2}{2}\right)$$

Následně mějme $w \in \mathbb{R}^d$ a využijme faktu, že:

Hustotní funkce mnohorozměrného Gaussova rozdělení je:

$$f(w) = \frac{1}{(2\pi)^{d/2}\Sigma^{1/2}} \cdot exp\left(-\frac{1}{2}(w-\mu)^T\Sigma^{-1}(w-\mu)\right)$$

Pokud $\Sigma = I$ a $\mu = c$, tak získáme zjednodušenou funkci hustoty:

$$f(w) = (2\pi)^{-d/2} exp\left(\frac{-\|w-c\|_2^2}{2}\right)$$

A z definice hustotní funkce platí, že:

$$\int (2\pi)^{-d/2} exp\left(\frac{-\|w-c\|_2^2}{2}\right) dw = 1$$

Integrál jakékoli funkce hustoty pravděpodobnosti (PDF) v celé její oblasti se musí rovnat 1, protože pravděpodobnost nějakého výsledku v celém prostoru musí být jistá. Tento integrál v podstatě shrnuje všechny pravděpodobnosti nalezení bodu w na libovolném místě v prostoru, jehož součet musí být podle definice pravděpodobnosti 1.

Poté pro libovolné $c \in \mathbb{R}^d$ odvodíme:

$$exp\left(\frac{\|x+y\|^2}{2}\right) = (2\pi)^{-d/2} exp\left(\frac{\|x+y\|^2}{2}\right) \int exp\left(\frac{-\|w-(x+y)\|^2}{2}\right) dw$$

$$= (2\pi)^{-d/2} \int exp\left(-\frac{\|w\|^2}{2} + w^T(x+y)\right) dw$$

$$= (2\pi)^{-d/2} \int exp\left(-\frac{\|w\|^2}{2}\right) \cdot exp(w^T x) \cdot exp(w^T y)\, dw$$

$$= \mathbb{E}_{w\sim\mathcal{N}(0,I_d)}[exp(w^T x) \cdot exp(w^T y)]$$

Po dosazení dostáváme:

$$SM(x,y) = exp\left(-\frac{\|x\|^2}{2}\right) \cdot \mathbb{E}_{w\sim\mathcal{N}(0,I_d)}[exp(w^T x) \cdot exp(w^T y)] \cdot exp\left(-\frac{\|y\|^2}{2}\right)$$

$$= \mathbb{E}_{w\sim\mathcal{N}(0,I_d)}\left[exp\left(w^T x - \frac{\|x\|^2}{2}\right) \cdot exp\left(w^T y - \frac{\|y\|^2}{2}\right)\right]$$

Tabulka estimátorů

$$K(x,y) \overset{\text{def}}{=} \mathbb{E}(\varphi(\mathrm{x})^T, \varphi(\mathrm{y})^T)$$

$$\varphi(\mathbf{x}) = \frac{h(\boldsymbol{x})}{\sqrt{m}} (f_1(w_1^T \boldsymbol{x}), \ldots, f_1(w_m^T \boldsymbol{x}), \ldots, f_l(w_1^T \boldsymbol{x}), \ldots, f_l(w_m^T \boldsymbol{x}))$$

Estimátory	$h(x)$	l	$f_1(x)$	$f_2(x)$	$\mathcal{D}$
$\widehat{SM_m^+}$	$\exp\left(-\frac{\|x\|^2}{2}\right)$	1	$\exp(x)$	None	$\mathcal{N}(0, I_d)$
$\widehat{SM_m^{hyp+}}$	$\frac{1}{\sqrt{2}}\exp\left(-\frac{\|x\|^2}{2}\right)$	2	$\exp(x)$	$\exp(-x)$	$\mathcal{N}(0, I_d)$

Tabulka 1 Estimátory popsány autory článku "Rethinking attention with performers"

Odkazy na zdrojové kódy modelů

Zdrojové kódy jsou dostupné na GitHubu autorů jednotlivých modelů:

- scBERT: https://github.com/TencentAILabHealthcare/scBERT
- scGPT: https://github.com/bowang-lab/scGPT
- TOSICA: https://github.com/JackieHanLab/TOSICA

Tabulka datasetů:

Dataset	Cell populations	Cell number	Accession number
Zheng68k	11	68 450	SRP073767
Baron	14	8 569	GSE84133
Muraro	9	2 122	GSE85241
Xin	4	1 449	GSE81608
Segerstolpe	13	2 133	E-MTAB-506
MacParland	20	8 444	GSE115469
myeloid	16	138 161	GSE154763
MS	16	22 327	E-HCAD-35
hPancreas	15	14 818	Tabulka 3
hArtery	10	50 202	GSE159677
hBone	7	26 140	GSE152805
mBrain	10	56 195	Tabulka 4
mPancreas	21	36 351	GSE132188
mAtlas	155	356 213	GSE132042

Tabulka 2 Dodatečné informace k datasetům zmíněných v textu. Pro vyhledání kompletního popisu datasetu je možné použít accession number.

hPancreas:

Dataset	Accession number
Baron	GSE84133
Muraro	GSE85241
xin	GSE81608
segerstolpe	E-MTAB-5061
Lawlor	GSE86473

Tabulka 3 Datasety, které jsou obsaženy v datasetu hPancreas.

mBrain:

Dataset	Accession number
Saunders	GSE116470
Tabula muris	GSE109774
Rosenberg	GSE110823
Zeisel	GSE60361

Tabulka 4 Datasety, které jsou obsaženy v datasetu mBrain.

Methods	Description	Isolation process	Applicability	Throughput (cells per run)	Cost	Merits	Limitations
Limiting dilution	Application of hand pipettes or pipetting robots to isolate single cells through dilution of the cell suspension	Manual/semi-automatic	Suspension cells	Low (< 100)	Low	Simple operation	Low specificity
							Low efficiency
							Low precision
							Cell loss
							Low work capacity (< 100)
Micromanipulation	Application of inverted microscope combined with micropipettes to select and isolate single cells	Manual	Suspension cells	Low (< 100)	Low	Simple operation	Low efficiency
						Flexible sampling	Mechanical injury
						Visualized operation	High difficulty
							Low work capacity (< 100)
LCM	Application of infrared laser under a microscope to isolate single cell or cell compartments from solid tissue samples	Manual	Tissue samples	Low (< 100)	High	Maintain integrity of sample	Nuclear damage
							Genetic material loss
							RNA pollution
							High difficulty
							Low work capacity (< 100)
FACS	Application of fluorescence labeling specific molecules on the cell surface to sort cells	Semi-automatic	Suspension cells	High (> 1000)	High	·High specificity	Mechanical injury
						·High accuracy	Large sample amount
						·High sensitivity	Cannot process cells less than 1000
Traps-based microfluidics	Application of microfluidic chips to separate single cells through traps	Semi-automatic	Suspension cells	High (> 1000)	High	Flexible operation	Low specificity
						Efficient cell pairing and fusion	Partial stimulation on cells
Valves-based microfluidics	Application of microfluidic chips to separate single cells through valves	Semi-automatic	Suspension cells	High (> 1000)	High	High sensitivity	Difficult and time-consuming fabrication
						High automation	Not portable
						Low sample volume	
Droplet-based microfluidics	Application of microfluidic chips to separate single cells through droplets	Semi-automatic	Suspension cells	High (1000–10,000)	High	High sensitivity	Random encapsulation
						High specificity	Complex equipment
						Noise-free	

Obrázek 29 Přehled metod využívaných pro separaci buněk v single-cell analýze. Přehled převzat z článku „Microfluidics applications for high-throughput single cell sequencing“.