



NATIONAL TECHNICAL UNIVERSITY OF ATHENS
SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING

DIVISION OF COMPUTER SCIENCE

**Cell2Text: Multimodal LLM for Generating Textual
Descriptions from Single-Cell RNA-Seq Profiles**

DIPLOMA THESIS

Aris Markogiannakis

Supervisor: Georgios Stamou
Professor, NTUA

Athens, November 2025



National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

Cell2Text: Multimodal LLM for Generating Textual Descriptions from Single-Cell RNA-Seq Profiles

DIPLOMA THESIS

Aris Markogiannakis

Supervisor: Georgios Stamou
Professor, NTUA

Approved by the three-member examining committee on November 5, 2025.

.....
Georgios Stamou
Professor, NTUA

.....
Michail Vazirgiannis
Professor, École Polytechnique

.....
Athanasios Voulodimos
Assistant Professor, NTUA

Athens, November 2025

.....

Άρης Μαρκογιαννάκης
Διπλωματούχος Ηλεκτρολόγος Μηχανικός και Μηχανικός Υπολογιστών, Ε.Μ.Π.

Σημείωμα Πνευματικών Δικαιωμάτων

Copyright © Aris Markogiannakis, 2025.

Απαγορεύεται η αντιγραφή, αποθήκευση και διανομή της παρούσας Εργασίας, εξ ολοκλήρου ή τμήματος αυτής, για εμπορικό σκοπό. Επιτρέπεται η ανατύπωση, αποθήκευση και διανομή για σκοπό μη κερδοσκοπικό, εκπαιδευτικής ή ερευνητικής φύσης, υπό την προϋπόθεση να αναφέρεται η πηγή προέλευσης και να διατηρείται το παρόν μήνυμα. Ερωτήματα που αφορούν τη χρήση της Εργασίας για κερδοσκοπικό σκοπό πρέπει να απευθύνονται προς τον συγγραφέα.

Σημείωμα Αποποίησης Ευθυνών

Οι απόψεις και τα συμπεράσματα που περιέχονται σε αυτό το έγγραφο εκφράζουν τον συγγραφέα και δεν πρέπει να ερμηνευθεί ότι αντιπροσωπεύουν τις επίσημες θέσεις του Εθνικού Μετσόβιου Πολυτεχνείου.

Abstract

Single-cell RNA sequencing has revolutionized biological research by enabling gene expression measurement at cellular resolution, revealing diverse cell types, states, and disease contexts. Recent single-cell foundation models can learn generalizable representations directly from expression data, improving downstream classification and clustering tasks. However, such models typically rely on fixed label spaces that limit their ability to express cellular diversity.

This thesis presents *Cell2Text*, a multimodal generative framework that transforms single-cell transcriptomic profiles into structured natural language descriptions. By integrating pretrained single-cell encoders with large language models through learnable projection modules, Cell2Text generates coherent summaries describing cellular identity, tissue of origin, disease relevance, and biological pathway activity. Experimental results show that Cell2Text achieves higher accuracy than baseline models, maintains strong ontological consistency through PageRank-based similarity metrics, and produces semantically faithful text outputs. Overall, the proposed approach highlights the potential of combining biological and linguistic representations for scalable and informative single-cell characterization.

Keywords: Deep Learning, Single-cell RNA sequencing, Foundation Models, Multimodal Learning, Gene Expression, Natural Language Generation, Large Language Models

Περίληψη

Η μονοκυτταρική ανάλυση *RNA* έχει μεταμορφώσει τη βιολογική έρευνα, επιτρέποντας τη μέτρηση της γονιδιακής έκφρασης σε επίπεδο κυττάρου και αποκαλύπτοντας ποικιλία κυτταρικών τύπων, καταστάσεων και παθολογικών συνθηκών. Τα πρόσφατα μοντέλα θεμελίωσης για μονοκυτταρικά δεδομένα μπορούν να μαθαίνουν γενικεύσιμες αναπαραστάσεις απευθείας από τα δεδομένα έκφρασης, βελτιώνοντας την ταξινόμηση και την ομαδοποίηση. Ωστόσο, τα υπάρχοντα μοντέλα περιορίζονται σε προκαθορισμένες ετικέτες, γεγονός που δυσχεραίνει την περιγραφή της πλήρους κυτταρικής ποικιλομορφίας.

Η παρούσα εργασία παρουσιάζει το *Cell2Text*, ένα πολυτροπικό παραγωγικό πλαίσιο που μετατρέπει προφίλ μονοκυτταρικής μεταγραφωμικής σε δομημένες περιγραφές φυσικής γλώσσας. Συνδυάζοντας προεκπαιδευμένους μετασχηματιστές μονοκυτταρικών δεδομένων με μεγάλα γλωσσικά μοντέλα, το *Cell2Text* παράγει συνεκτικές περιγραφές που αποτυπώνουν τον κυτταρικό τύπο, την προέλευση ιστού, τη συσχέτιση με ασθένειες και τη δραστηριότητα βιολογικών οδών. Τα πειραματικά αποτελέσματα δείχνουν ότι το *Cell2Text* υπερτερεί των βασικών μοντέλων ως προς την ακρίβεια και τη σημασιολογική συνέπεια, αναδεικνύοντας τη δυναμική των πολυτροπικών μοντέλων για την ανάλυση μονοκυτταρικών δεδομένων.

Λέξεις-Κλειδιά: Βαθιά Μάθηση, Μονοκυτταρική αλληλούχηση *RNA*, Μοντέλα Θεμελίωσης, Πολυτροπική Μάθηση, Γονιδιακή Έκφραση, Παραγωγή Φυσικής Γλώσσας, Μεγάλα Γλωσσικά Μοντέλα

Acknowledgements

I would like to express my sincere gratitude to Professor Michalis Vazirgiannis for the opportunity to conduct this thesis in his laboratory. I also wish to thank my professor at the National Technical University of Athens, Professor Georgios Stamou, whose courses inspired me and introduced me to the field of Artificial Intelligence. Special thanks go to Dr. Michalis Chatziannastasis and PhD candidate Xiao Fei for their help, insightful discussions, and valuable contribution to the completion of this work. Finally, I warmly thank my family for their constant support and patience throughout these years.

Ευχαριστίες

Θα ήθελα να εκφράσω την ειλικρινή μου ευγνωμοσύνη στον καθηγητή κ. Μιχάλη Βαζιργιάννη για την ευκαιρία να πραγματοποιήσω την εργασία αυτή στο εργαστήριό του. Ευχαριστώ επίσης τον καθηγητή μου στο Εθνικό Μετσόβιο Πολυτεχνείο, κ. Γεώργιο Στάμου, του οποίου τα μαθήματα αποτέλεσαν για εμένα πηγή έμπνευσης και εισαγωγή στον χώρο της Τεχνητής Νοημοσύνης. Ιδιαίτερες ευχαριστίες θα ήθελα να εκφράσω στον Δρ. Μιχάλη Χατζηαναστάση και στον υποψήφιο Διδάκτορα Xiao Fei για τη βοήθειά τους, τις χρήσιμες συζητήσεις και τη συμβολή τους στην ολοκλήρωση αυτής της εργασίας. Τέλος, ευχαριστώ θερμά την οικογένειά μου για τη συνεχή υποστήριξη και υπομονή τους όλα αυτά τα χρόνια.

Contents

Abstract	5
Περίληψη	7
Acknowledgements	9
Ευχαριστίες	11
Contents	13
Contents	16
List of Figures	17
List of Figures	17
List of Tables	18
List of Tables	19
1 Εκτεταμένη Περίληψη στα Ελληνικά	20
1.1 Εισαγωγή	20
1.2 Θεωρητικό Υπόβαθρο	21
1.3 Κατασκευή Πολυτροπικού Συνόλου Δεδομένων	24
1.4 <i>Cell2Text</i> : Αρχιτεκτονική, Εκπαίδευση και Αξιολόγηση	28
1.5 Αποτελέσματα	32
1.6 Συμπεράσματα και Μελλοντικές Επεκτάσεις	35
2 Introduction	37
2.1 Motivation	37
2.2 Problem Statement	38
2.3 Contributions	38
2.4 Thesis Structure	39

3	Theoretical Background	40
3.1	Single-cell Biology	40
3.1.1	Introduction	40
3.1.2	scRNA-seq Technologies	40
3.1.3	Data Characteristics and Challenges	41
3.1.4	Quality Control and Normalization	41
3.1.5	Dimensionality Reduction and Clustering	42
3.1.6	Extensions and Multimodal Data	42
3.1.7	Biological Ontologies and Standardized Annotations	42
3.1.8	Pathway-based Annotation	43
3.2	Machine Learning	43
3.2.1	Introduction	43
3.2.2	Learning Paradigms	44
3.2.3	Key Concepts and Algorithms	44
3.2.4	Model Evaluation and Validation	45
3.3	Deep Learning and Neural Networks	45
3.3.1	Optimization in Deep Neural Networks	45
3.3.2	Feedforward Neural Networks	46
3.3.3	Convolutional Neural Networks (CNNs)	47
3.3.4	Recurrent Neural Networks (RNNs)	48
3.3.5	Graph Neural Networks (GNNs)	48
3.3.6	Attention Mechanisms and Transformers	49
3.4	Single-cell Gene Expression Data Representation Learning	50
3.4.1	Early Deep Learning Models	51
3.4.2	Transformer-Based Foundation Models	51
3.5	Large Language Models	55
3.5.1	Evolution of LLM Architectures	56
3.5.2	LLaMA	56
3.5.3	Pretraining Paradigms	57
3.5.4	Finetuning Techniques	58
3.5.5	Multimodal Large Language Models	59
3.5.6	Applications and Challenges	60
3.6	Inspiration for Cell2Text and Similar Frameworks	61
3.6.1	Prot2Text	61
3.6.2	CellWhisperer	62
3.6.3	Cell2Sentence	62
3.6.4	InstructCell	63
3.6.5	scMMGPT	63
4	Multimodal Dataset Construction and Sampling Methodology	64
4.1	Overview	64
4.2	Data Source and Initial Filtering	64
4.2.1	Quality Control and Exclusion Criteria	64

4.3	Stratified Sampling Strategy	65
4.3.1	Distribution-based Sampling with Tissue Adjustment	65
4.3.2	Cell Type Diversity Sampling	66
4.3.3	Disease Representation Strategy	66
4.3.4	Donor Diversity and Rare Category Sampling	66
4.4	Dataset Splitting Strategy	66
4.5	Diversity Metrics and Quality Assessment	66
4.6	Multimodal Text Description Generation	67
4.6.1	Ontology Integration and Pathway Enrichment	68
4.6.2	Natural Language Description Architecture	69
4.6.3	Text Description Quality Metrics	70
4.7	Validation and Quality Assurance	70
4.8	Final Dataset Characteristics	70
5	Cell2Text: Architecture, Training, Evaluation	76
5.1	Model Architecture	76
5.1.1	Cell Encoder: Geneformer	77
5.1.2	Projection Module	77
5.1.3	Text Decoder: LLaMA	79
5.1.4	Forward Pass	79
5.2	Training Strategy	79
5.2.1	Frozen Cell Encoder	80
5.2.2	Projection Module Training	80
5.2.3	Decoder Finetuning Strategies	80
5.2.4	Decoder Input Integration	81
5.2.5	Training Objective	82
5.3	Evaluation Metrics	82
5.3.1	Ontology-Based Pagerank Cell Type Similarity Metric	82
5.3.2	Evaluation Metrics for Classification	86
5.3.3	Evaluation Metrics for Text Generation Quality	89
6	Experiments & Results	92
6.1	Experimental Setup	92
6.1.1	Hardware and Software Environment	92
6.1.2	Model Architecture	92
6.1.3	Training Configuration	93
6.1.4	Fine-tuning Strategies	94
6.2	Results	94
6.2.1	Text Generation Quality	94
6.2.2	Biological Entity Recognition	96
6.2.3	Pathway Activity Identification	98

7 Conclusion &Future Work	100
7.1 Conclusion	100
7.2 Future Work	101
Bibliography	103

List of Figures

1.1	Επισκόπηση της κατανομής τύπων κυττάρων στο σύνολο δεδομένων. Για σαφήνεια, εμφανίζονται μόνο οι 30 πιο συχνές κατηγορίες.	27
1.2	Επισκόπηση της κατανομής κατηγοριών ασθενειών στο σύνολο δεδομένων. Για σαφήνεια, εμφανίζονται μόνο οι 30 πιο συχνές κατηγορίες.	27
1.3	Επισκόπηση της κατανομής κατηγοριών ιστών στο σύνολο δεδομένων. Για σαφήνεια, εμφανίζονται μόνο οι 30 πιο συχνές κατηγορίες.	28
1.4	Σχηματική απεικόνιση της αρχιτεκτονικής του μοντέλου <i>Cell2Text</i>	29
3.1	Transformer Architecture (Conceptual Diagram)	50
3.2	Timeline of Large Language Model Evolution.	58
4.1	Distribution of tokens in gene expression sequences	72
4.2	Distribution of tokens in natural language descriptions	72
4.3	Overview of the distribution of cell types in the dataset. For clarity, only the 30 most abundant categories are shown.	73
4.4	Overview of the distribution of disease categories in the dataset. For clarity, only the 30 most abundant categories are shown.	73
4.5	Overview of the distribution of tissue (general) categories in the dataset. For clarity, only the 30 most abundant categories are shown.	74
4.6	Overview of the distribution of pathway categories in the dataset.	74
4.7	Overview of the distribution of assay categories in the dataset.	75
5.1	High-level architecture of the Cell2Text model.	77
5.2	Overview of the Q-Former architecture (adapted from BLIP-2). The Q-Former bridges the cell encoder outputs to the decoder space via learnable query tokens that <i>summarize</i> the biological information through cross-attention.	78
5.3	Distribution of similarity scores across all cell type pairs in the Cell Ontology, plotted on a logarithmic frequency scale to emphasize the heavy-tailed nature.	85
5.4	Cumulative distribution function (CDF) of similarity scores, showing the rapid accumulation of low similarity values.	86

List of Tables

1.1	Μετρικές ποιότητας παραγόμενου κειμένου ανά μοντέλο. Υψηλότερες τιμές υποδηλώνουν καλύτερη ποιότητα παραγωγής κειμένου.	32
1.2	Απόδοση αναγνώρισης τύπου κυττάρου. <i>Ont. Sim.</i> : Οντολογική ομοιότητα	33
1.3	Απόδοση πρόβλεψης ασθένειας και ιστού.	33
1.4	Απόδοση πρόβλεξης ενεργών βιολογικών οδών. <i>Sub. Acc.</i> : Ακρίβεια πλήρους συνόλου, <i>Jac. Sim.</i> : Ομοιότητα <i>Jaccard</i> , $W - F1$: Σταθμισμένο $F1$	34
4.1	Sampling strategy allocation and objectives	65
4.2	Normalized Shannon diversity before and after applying the sampling strategy.	67
4.3	Number of Cell of each Dataset Split	71
5.1	Summary statistics of the PageRank-based similarity scores across all cell type pairs in the Cell Ontology, highlighting the heavy-tailed distribution.	85
6.1	Training hyperparameters for each Cell2Text configuration.	93
6.2	Text generation quality metrics across model configurations. Higher scores indicate better generation quality. Metrics: BLEU (Bilingual Evaluation Understudy), ROUGE (Recall-Oriented Understudy for Gisting Evaluation), BBT (BioBERT F1-Score), RBT (RoBERTa F1-Score).	95
6.3	Cell type recognition performance. Cell Type Accuracy measures exact matches, while Ont Sim. (Ontology Similarity) captures semantic closeness in the Cell Ontology hierarchy. The other metric is weighted F1 score.	96
6.4	Disease and tissue recognition performance across model configurations. Metrics include Acc (Accuracy) and weighted F1 score. Bold indicates best performance, <u>underline</u> indicates second best.	97

6.5	Pathway prediction performance metrics. Sub. Acc. (Subset Accuracy) requires exact match of both predicted pathways, Jac. Sim. (Jaccard Similarity) measures set overlap allowing partial matches, and W-F1 (Weighted F1) accounts for pathway frequency in the dataset. Bold indicates best performance, <u>underline</u> indicates second best.	99
-----	---	----

Κεφάλαιο 1

Εκτεταμένη Περίληψη στα Ελληνικά

1.1 Εισαγωγή

Κίνητρο

Η τεχνική της αλληλούχισης *RNA* σε μεμονωμένα κύτταρα (*scRNA-seq*) επιτρέπει να μετράμε τη γονιδιακή έκφραση σε κάθε κύτταρο ξεχωριστά, αποκαλύπτοντας έτσι την ετερογένεια που χάνεται στις μαζικές αναλύσεις. Η τεχνολογία αυτή έχει αλλάξει ριζικά τη βιοϊατρική έρευνα, επιτρέποντας τη μελέτη ανάπτυξης, ανοσολογικών αποκρίσεων και παθολογικών μηχανισμών.

Ωστόσο, τα δεδομένα *scRNA-seq* είναι υψηλής διαστατικότητας, αραιά και θορυβώδη, γεγονός που καθιστά δύσκολη την ανάλυσή τους. Οι περισσότερες υπάρχουσες προσεγγίσεις βασίζονται στη μείωση διαστάσεων και στη χρήση ταξινομητών με προκαθορισμένες κατηγορίες κυττάρων, μια διαδικασία που είναι δύσκολο να κλιμακωθεί και περιορίζει την εκφραστικότητα του μοντέλου.

Πρόσφατα θεμελιώδη μοντέλα, όπως τα *Geneformer* [1] και *scGPT* [2], μαθαίνουν πλούσιες αναπαραστάσεις κυττάρων απευθείας από δεδομένα έκφρασης. Ωστόσο, η χρήση προκαθορισμένων κεφαλών ταξινόμησης ως έξοδο περιορίζει την εκφραστικότητα και τη δυνατότητα γενίκευσης των μοντέλων.

Πρόβλημα και στόχος

Αν και τα σύγχρονα θεμελιώδη μοντέλα μαθαίνουν πλούσιες ενσωματώσεις από δεδομένα έκφρασης, η χρήση κατηγορικών εξόδων περιορίζει τη δυνατότητά τους να αποτυπώσουν τη λειτουργική ποικιλία των κυττάρων.

Η παρούσα διατριβή προτείνει ένα πλαίσιο που συνδυάζει θεμελιώδη μοντέλα κυτταρικών δεδομένων με μεγάλα γλωσσικά μοντέλα, ώστε να αποδίδει τις ενσωματώσεις κυττάρων σε περιγραφές φυσικής γλώσσας, ενισχύοντας την ερμηνευσιμότητα και τη

βιολογική συνέπεια των αποτελεσμάτων.

Συνεισφορές

Προτείνουμε το *Cell2Text*, ένα πολυτροπικό γενετικό πλαίσιο που συνδυάζει θεμελιώδη μοντέλα μεμονωμένων κυττάρων και μεγάλα γλωσσικά μοντέλα (*LLMs*) για τη μετατροπή προφίλ γονιδιακής έκφρασης σε φυσική γλώσσα. Οι κύριες συνεισφορές είναι:

- **Πλαίσιο:** Αντιστοίχιση των ενσωματώσεων κυττάρων με *LLMs* για παραγωγή βιολογικά τεκμηριωμένων περιγραφών.
- **Αποτελεσματικότητα:** Ανταγωνιστικά αποτελέσματα σε ταξινομήσεις κυτταρικού τύπου, ιστού και βιολογικών οδών, με διατήρηση οντολογικής συνέπειας.
- **Ερμηνευσιμότητα:** Παραγωγή περιγραφών που αποτυπώνουν ταυτότητα, ιστικό πλαίσιο και λειτουργική δραστηριότητα.
- **Δεδομένα:** Δημιουργία εκτεταμένου πολυτροπικού συνόλου από το *CELLxGENE*, εμπλουτισμένου με οντολογίες και μεταδεδομένα.

1.2 Θεωρητικό Υπόβαθρο

Βιολογία Μεμονωμένων Κυττάρων

Τα κύτταρα αποτελούν τις θεμελιώδεις μονάδες της ζωής, και η μοριακή τους κατάσταση καθορίζει τη λειτουργία τους μέσα σε ιστούς και οργανισμούς. Οι παραδοσιακές τεχνικές μοριακής ανάλυσης, όπως η μαζική αλληλούχιση *RNA* (*bulk RNA-seq*), μετρούν τον μέσο όρο της γονιδιακής έκφρασης σε μεγάλο πληθυσμό κυττάρων, αποκρύπτοντας την κυτταρική ετερογένεια και τις σπάνιες καταστάσεις. Η μονοκυτταρική αλληλούχιση *RNA* (*scRNA-seq*), που εισήχθη από τους *Tang et al.* [3], επιτρέπει την ποσοτική ανάλυση του μεταγραφώματος σε επίπεδο μεμονωμένων κυττάρων, αποκαλύπτοντας διαφοροποιήσεις σε τύπους, καταστάσεις και αναπτυξιακές γραμμές. Αυτή η δυνατότητα έχει αλλάξει ριζικά πεδία όπως η αναπτυξιακή βιολογία, η ανοσολογία και η ογκολογία, συμβάλλοντας σε πρωτοβουλίες όπως ο Άτλας Ανθρώπινων Κυττάρων.

Οι τεχνολογίες *scRNA-seq* διακρίνονται σε μεθόδους βασισμένες σε πλάκες (π.χ. *SMART-seq2*), που προσφέρουν πλήρη μεταγραφήματα υψηλής ευαισθησίας, και σε μεθόδους βασισμένες σε σταγόνες (π.χ. *Drop-seq*, *10x Genomics*) για υψηλή απόδοση. Οι μοναδικοί μοριακοί αναγνωριστές (*UMIs*) βελτιώνουν την ποσοτική ακρίβεια, ενώ τα δεδομένα χαρακτηρίζονται από υψηλή διαστατικότητα, αραιότητα και υπερδιασπορά, απαιτώντας μεθόδους κανονικοποίησης όπως η *SCTransform*.

Πριν από την ανάλυση, εφαρμόζεται ποιοτικός έλεγχος για την αφαίρεση χαμηλής ποιότητας κυττάρων και γονιδίων, ενώ η μείωση διαστατικότητας με *PCA*, *t-SNE* ή *UMAP*, ακολουθούμενη από συσταδοποίηση με τον αλγόριθμο *Leiden*, επιτρέπει την οπτικοποίηση και την ταυτοποίηση κυτταρικών πληθυσμών μέσω γονιδίων-δεικτών. Πρόσφατες επεκτάσεις ενσωματώνουν πολυτροπικά δεδομένα, όπως το *CITE-seq* για πρωτεΐνες επιφάνειας, το *scATAC-seq* για προσβασιμότητα χρωματίνης και τη χωρική μεταγραφωμική για τον εντοπισμό κυττάρων εντός ιστών. Οι βιολογικές οντολογίες (*Gene Ontology*, *Cell Ontology*) και οι βάσεις βιολογικών οδών (*KEGG*, *Reactome*, *MSigDB*) παρέχουν τυποποιημένα πλαίσια για σχολιασμό και εκτίμηση δραστηριότητας οδών, με εργαλεία όπως το *AUCell* που συνδέουν μοτίβα έκφρασης με λειτουργικές διεργασίες.

Μάθηση Αναπαράστασης για Δεδομένα Γονιδιακής Έκφρασης Μεμονωμένων Κυττάρων

Η αλληλούχιση μεμονωμένων κυττάρων (*scRNA-seq*) αποκάλυψε την κυτταρική ετερογένεια, αλλά τα δεδομένα είναι υψηλής διαστατικότητας, αραιά και θορυβώδη. Η μάθηση αναπαράστασης μέσω βαθιάς μάθησης στοχεύει στην εξαγωγή συμπαγών, εκφραστικών αναπαραστάσεων που συλλαμβάνουν κρυμμένα βιολογικά μοτίβα.

Πρώιμα μοντέλα όπως το *scVI* [4] βασίζονται σε παραλλαγές αυτοκωδικοποιητών (*VAEs*) με κατανομές αρνητικής διωνυμικής, ώστε να μοντελοποιούν τη στοχαστικότητα και την υπερδιασπορά των μετρήσεων. Το *scGNN* [5] εισάγει νευρωνικά δίκτυα γράφων (*GNNs*), κατασκευάζοντας γράφους κυττάρων και μαθαίνοντας αναπαραστάσεις που είναι ανθεκτικές σε θόρυβο και χρήσιμες για *clustering* και *imputation*.

Πιο πρόσφατα, μοντέλα τύπου *transformer* όπως το *Geneformer* [1] και το *scGPT* [2] αντιμετωπίζουν τα γονίδια ως *tokens*, μαθαίνοντας εξαρτήσεις έκφρασης μέσω αυτο-επιβλεπόμενης εκπαίδευσης σε εκατομμύρια κύτταρα. Αυτή η προσέγγιση, εμπνευσμένη από τη γλωσσική μοντελοποίηση, έχει αποδειχθεί αποτελεσματική σε εργασίες όπως *annotation*, πρόβλεψη *perturbations* και μεταφορά γνώσης μεταξύ οργανισμών.

Μεγάλα Γλωσσικά Μοντέλα

Τα Μεγάλα Γλωσσικά Μοντέλα (*LLMs*) έχουν αλλάξει ριζικά την επεξεργασία φυσικής γλώσσας, επιτρέποντας σε υπολογιστικά συστήματα να κατανοούν και να παράγουν κείμενο με ανθρώπινη ευχέρεια. Βασίζονται στην αρχιτεκτονική των *transformers* [6], η οποία αξιοποιεί μηχανισμούς αυτοπροσοχής για τη μοντελοποίηση μακρινών εξαρτήσεων, αντικαθιστώντας τις αναδρομικές δομές (*RNNs*, *LSTMs*). Από τα πρώτα μοντέλα όπως το *BERT* [7] και το *GPT* [8] έως τα σύγχρονα πολυτροπικά συστήματα (*GPT-4*, *PaLM*, *T5*), η αύξηση της κλίμακας και η βελτίωση της εκπαίδευσης οδήγησαν σε γενικευμένες ικανότητες κατανόησης, συλλογισμού και παραγωγής φυσικής γλώσσας.

Η οικογένεια *LLaMA* της *Meta* αντιπροσωπεύει την ανοικτή και αποδοτική κατεύθυνση των *LLMs*, με εκδόσεις σχεδιασμένες για διαφορετικές υπολογιστικές ανάγκες. Τα *LLaMA* 3.1 και 3.2 περιλαμβάνουν βελτιωμένες αρχιτεκτονικές *transformers*, εκτεταμένο λεξιλόγιο και παράθυρο συμφραζομένων 128K *tokens*, προσφέροντας ισχυρή επίδοση με χαμηλότερο υπολογιστικό κόστος. Ειδικά τα μικρά μοντέλα *LLaMA* 3.1 1B και 3B, που χρησιμοποιούνται στο παρόν έργο, συνδυάζουν υψηλή ποιότητα κειμενικής κατανόησης με αποδοτική εκτέλεση, καθιστώντας τα ιδανικά για ενσωμάτωση σε ερευνητικά πλαίσια όπως το προτεινόμενο *Cell2Text*.

Τα Μεγάλα Γλωσσικά Μοντέλα (*LLMs*) εκπαιδεύονται αρχικά με αυτοεπιβλεπόμενη μάθηση σε τεράστια μη επισημασμένα δεδομένα κειμένου, μαθαίνοντας τη δομή και τη σημασιολογία της γλώσσας. Η προεκπαίδευση μπορεί να βασίζεται σε μάσκες (*BERT*), σε αυτοπαλινδρόμηση (*GPT*) ή σε ενοποιημένα σχήματα μετατροπής κειμένου σε κείμενο (*T5*). Στη συνέχεια, τα μοντέλα εξειδικεύονται μέσω *fine-tuning*, δηλαδή προσαρμογής σε συγκεκριμένες εργασίες με μικρότερα, επισημασμένα σύνολα δεδομένων. Για να μειωθεί το υπολογιστικό κόστος, χρησιμοποιούνται μέθοδοι *Parameter – Efficient Fine – Tuning* όπως το *LoRA* [9], το οποίο δεν ενημερώνει όλους τους παραμέτρους αλλά μόνο μικρούς πίνακες χαμηλής τάξης μέσα στα στρώματα του μετασχηματιστή. Έτσι επιτυγχάνεται σημαντική μείωση των απαιτήσεων μνήμης και χρόνου εκπαίδευσης χωρίς απώλεια απόδοσης, καθιστώντας εφικτή την προσαρμογή μοντέλων (όπως τα *LLaMA* 3.1 1B και 3B) σε εξειδικευμένα δεδομένα.

Τα **πολυτροπικά Γλωσσικά Μοντέλα** επεκτείνουν αυτή τη φιλοσοφία ενσωματώνοντας πολλαπλές μορφές δεδομένων όπως κείμενο, εικόνες και βιολογικά δεδομένα για ενιαία κατανόηση και παραγωγή πληροφορίας. Μοντέλα όπως τα *BLIP-2* και *LLaVA* συνδυάζουν οπτικούς κωδικοποιητές με γλωσσικά μοντέλα μέσω μηχανισμών διασταυρούμενης προσοχής, επιτρέποντας εφαρμογές όπως περιγραφή εικόνων, οπτική ερώτηση-απάντηση και βιοϊατρική ανάλυση. Στο πεδίο της υπολογιστικής βιολογίας, οι πολυτροπικές προσεγγίσεις ανοίγουν τον δρόμο για τη συνδυαστική κατανόηση δεδομένων μεταγραφωμικής, μικροσκοπίας και κειμενικής πληροφορίας, ενισχύοντας την ερμηνευσιμότητα και τις δυνατότητες ανακάλυψης.

Σχετικά Πλαίσια και Επιρροές στον Σχεδιασμό του *Cell2Text*

Η ανάπτυξη του *Cell2Text* βασίστηκε σε πρόσφατες προσεγγίσεις πολυτροπικής μάθησης που συνδέουν βιολογικά δεδομένα υψηλής διάστασης με τη φυσική γλώσσα, διευκολύνοντας την ερμηνεία και την πρόσβαση στη γνώση. Το *Prot2Text* αποτέλεσε σημείο εκκίνησης, αποδεικνύοντας ότι οι πρωτεϊνικές ακολουθίες μπορούν να μετατραπούν σε φυσικές περιγραφές μέσω ενσωμάτωσης γραφημάτων και γλωσσικών μοντέλων. Παρόμοια, το *CellWhisperer* έδειξε πώς οι αναπαραστάσεις μεταγραφωμάτων μπορούν να ευθυγραμμιστούν με γλωσσικούς χώρους μέσω αντιθετικής εκπαίδευσης, επιτρέποντας αλληλεπιδραστική διερεύνηση δεδομένων *scRNA – seq*. Το *CellSentence* εισήγαγε τη μετατροπή γονιδιακής έκφρασης σε «προτάσεις κυττάρων», επιτρέποντας στα *LLMs* να μάθουν από συμβολικές αναπαραστάσεις γονι-

διακής έκφρασης.

Επιπλέον, το *InstructCell* εισήγαγε μηχανισμούς *instruction-tuning* και *cross-modal* ευθυγράμμιση, χρησιμοποιώντας *Query Transformers* για σύνδεση δεδομένων *scRNA-seq* με *LLMs* μέσω φυσικών ερωταποκρίσεων, βελτιώνοντας την ερμηνευσιμότητα και την αλληλεπίδραση. Τέλος, το *scMMGPT* ενοποίησε τα *scGPT* και *LLaMA-2* μέσω *Q-Formers* για αμφίδρομη μάθηση μεταξύ κυτταρικών και γλωσσικών αναπαραστάσεων, προσφέροντας πλούσια, βιολογικά συνεπή περιγραφικά κείμενα.

1.3 Κατασκευή Πολυτροπικού Συνόλου Δεδομένων

Πηγή Δεδομένων και Στρατηγική Δειγματοληψίας

Η βάση του πολυτροπικού μας συνόλου δεδομένων προέρχεται από το *CELLxGENE Census* [10], ένα τυποποιημένο αποθετήριο δεδομένων Μονοκυτταρικής Βιολογίας. Εξαγάγαμε κυτταρικά μεταδεδομένα συμπεριλαμβανομένων του αναπτυξιακού σταδίου, της κατάστασης ασθένειας, του τύπου πειράματος, της ιστικής προέλευσης, των σχολιασμών τύπου κυττάρου και των πληροφοριών δότη για δείγματα ανθρώπινων οργάνων.

Αναγκαιότητα Στρατηγικής Δειγματοληψίας

Η απλή τυχαία δειγματοληψία από μεγάλες βιολογικές βάσεις δεδομένων όπως το *CELLxGENE* οδηγεί σε σημαντικά προβλήματα αναπαράστασης. Συγκεκριμένοι ιστοί (όπως ο εγκέφαλος και το αίμα), τύποι κυττάρων (όπως τα *T* λεμφοκύτταρα) και ασθένειες (όπως η *COVID-19*) είναι δυσανάλογα υπερεκπροσωπημένοι λόγω της ευκολίας συλλογής δειγμάτων, της κλινικής σημασίας και ιστορικών ερευνητικών προτεραιοτήτων. Αυτή η μεροληψία θα οδηγούσε σε μοντέλα μηχανικής μάθησης που αποδίδουν άριστα σε κοινούς τύπους κυττάρων αλλά αποτυγχάνουν σε σπάνιες αλλά βιολογικά σημαντικές κατηγορίες.

Για την αντιμετώπιση των μεροληψιών και τη διασφάλιση αντιπροσωπευτικής και βιολογικά ποικιλόμορφης δειγματοληψίας, σχεδιάστηκε μια στρατηγική με πέντε συμπληρωματικές συνιστώσες. Ο συνδυασμός δειγματοληψίας βάσει κατανομής, ποικιλομορφίας τύπων κυττάρων και δοτών, ισορροπίας φυσιολογικών-παθολογικών καταστάσεων και στοχευμένης επιλογής σπάνιων κατηγοριών επιτρέπει τη ρεαλιστική αναπαράσταση του κυτταρικού πληθυσμού. Η προσέγγιση αυτή μειώνει τις υπερεκπροσωπήσεις κοινών ιστών, διατηρεί κρίσιμες πηγές βιολογικής μεταβλητότητας και εξασφαλίζει ευρεία κάλυψη του κυτταρικού και ιστικού ρεπερτορίου, ενισχύοντας έτσι τη γενίκευση των μοντέλων.

Εφαρμόσαμε επίσης μια διαδικασία διαχωρισμού εκπαίδευσης-επικύρωσης-δοκιμής με αναλογία 80/10/10, με διαχωρισμό σε επίπεδο δότη για τη διατήρηση της βιολογικής

ανεξαρτησίας και την αποφυγή διαρροής δεδομένων. Ο αλγόριθμος διασφαλίζει ότι όλες οι βιολογικές κατηγορίες που υπάρχουν στα σύνολα επικύρωσης και δοκιμής αντιπροσωπεύονται στο σύνολο εκπαίδευσης.

Δημιουργία Περιγραφών Φυσικής Γλώσσας

Η δεύτερη συνιστώσα του πολυτροπικού μας συνόλου δεδομένων αποτελείται από δομημένες περιγραφές φυσικής γλώσσας που δημιουργούνται από κυτταρικά μεταδεδομένα και λειτουργικούς σχολιασμούς.

Ενσωμάτωση Οντολογίας και Εμπλουτισμός Βιολογικών Οδών

Ενισχύσαμε τους κυτταρικούς σχολιασμούς μέσω δύο συμπληρωματικών προσεγγίσεων. Πρώτον, οι σχολιασμοί τύπου κυττάρου, ιστού και ασθένειας εμπλουτίστηκαν χρησιμοποιώντας τυποποιημένες βιολογικές οντολογίες όπως η *Cell Ontology (CL)*. Για κάθε οντολογικό όρο, εξαγάγαμε κανονικά ονόματα όρων, επίσημους ορισμούς όταν ήταν διαθέσιμοι, και συνώνυμα για ολοκληρωμένη κάλυψη.

Δεύτερον, υπολογίσαμε βαθμολογίες δραστηριότητας βιολογικών οδών χρησιμοποιώντας το *AUCell* [11], μια μέθοδο για την αξιολόγηση του εμπλουτισμού γονιδιακών υπογραφών σε επίπεδο μεμονωμένου κυττάρου. Μετά από καθολική προεπεξεργασία και κανονικοποίηση, ενσωματώσαμε τα 50 *Hallmark pathways* από το *MSigDB* και υπολογίσαμε βαθμολογίες εμπλουτισμού για κάθε κύτταρο. Φιλτράραμε βιολογικές οδούς ενεργές σε περισσότερο από 0.5% των κυττάρων και εντοπίσαμε τις δύο κορυφαίες εμπλουτισμένες βιολογικές οδούς ανά κύτταρο για συμπερίληψη στις κλειμενικές περιγραφές.

Αρχιτεκτονική Περιγραφών Φυσικής Γλώσσας

Κάθε περιγραφή κυττάρου ακολουθεί ένα δομημένο πρότυπο που ενσωματώνει τον τύπο κυττάρου με τον οντολογικό του ορισμό, την ιστική προέλευση, την κατάσταση ασθένειας, το φύλο, το αναπτυξιακό στάδιο και τις δύο κορυφαίες ενεργές βιολογικές οδούς. Το σύστημα χειρίζεται έξυπνα τις ελλείπουσες πληροφορίες, συνδυάζει πλεοναστικούς σχολιασμούς ιστών και προσαρμόζει τη γλώσσα βάσει της κατάστασης ασθένειας και των δημογραφικών χαρακτηριστικών. Για παράδειγμα, μια περιγραφή μπορεί να αναφέρει:

Example Description

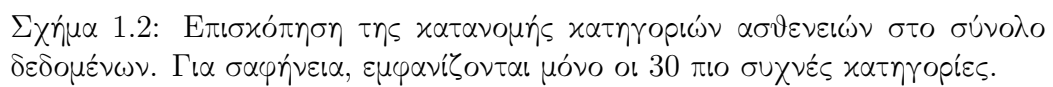
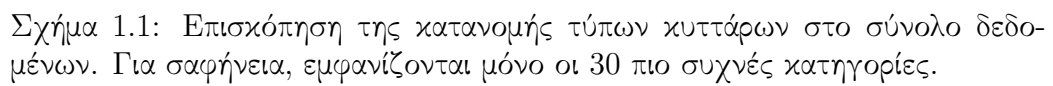
This sample consists of a ciliated columnar cell of tracheobronchial tree, multi-ciliated epithelial cell located in the trachea and bronchi, characterized by a columnar shape and motile cilia on its apical surface. These cilia facilitate mucociliary clearance by moving mucus and trapped particles toward the pharynx. It originates from the lung parenchyma of a normal male during elderly stage. This cell is associated with Genes mediating programmed cell death (apoptosis) by activation of caspases. Additionally, it involves Genes down-regulated in response to ultraviolet (UV) radiation.

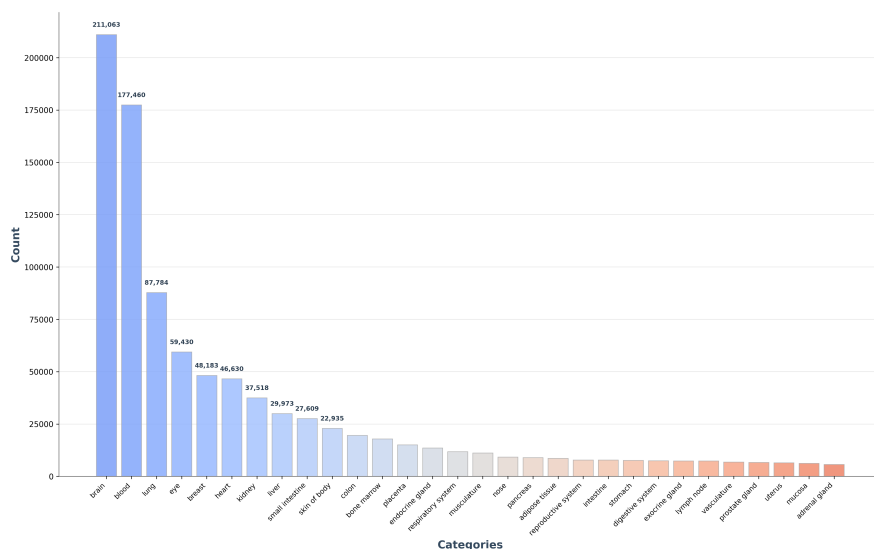
Τελικά Χαρακτηριστικά και Μεροληψίες Συνόλου Δεδομένων

Οι ακολουθίες γονιδιακής έκφρασης που τοκενοποιήθηκαν με τον τοκενοποιητή του *Geneformer* έχουν μέσο μήκος 1,843.2 τόκεν, ενώ οι περιγραφές φυσικής γλώσσας που τοκενοποιήθηκαν με τον τοκενοποιητή του *Llama – 3.2 – 1B – Instruct* είναι σημαντικά βραχύτερες, με μέσο όρο 104.3 τόκεν.

Η εξέταση των βιολογικών κατηγοριών αποκαλύπτει σημαντικές ανισορροπίες που αντανακλούν τις εγγενείς μεροληψίες στη συλλογή δεδομένων μονοκυτταρικής. Όπως φαίνεται στο Εικόνα 1.1, μεταξύ των 783 τύπων κυττάρων, τα γλοιακά κύτταρα όπως οι ολιγοδενδροκύτταροι κυριαρχούν λόγω δεδομένων που προέρχονται από εγκέφαλο, ενώ ανοσοκύτταρα όπως τα *T* και *B* λεμφοκύτταρα είναι επίσης εξέχοντα. Αντίθετα, στρωματικά και πλασματοκύτταρα παραμένουν υποεκπροσωπημένα. Στις κατηγορίες ασθενειών (Εικόνα 1.2), φυσιολογικά δείγματα και δείγματα *COVID – 19* κυριαρχούν μεταξύ 128 κατηγοριών, αντανακλώντας τον κεντρικό ρόλο των υγιών δοτών και την αύξηση της συλλογής δεδομένων κατά τη διάρκεια της πανδημίας. Οι κατανομές σε επίπεδο ιστών (Εικόνα 1.3) δείχνουν παρόμοιες ανισορροπίες, με εγκέφαλο και αίμα να κυριαρχούν μεταξύ των 347 κατηγοριών, ακολουθούμενους από πνεύμονα και μάτι, γεγονός που αντανακλά την εστίαση σε κλινικά σχετικούς και εύκολα προσβάσιμους ιστούς.

Συνολικά, αυτά τα χαρακτηριστικά υπογραμμίζουν ότι η σύνθεση του συνόλου δεδομένων αντανακλά όχι μόνο τη βιολογική συνάφεια αλλά και τη μεθοδολογική προσβασιμότητα, τις τεχνικές προτιμήσεις και τις ιστορικές ερευνητικές προτεραιότητες. Παρά τη στρατηγική στρωματοποιημένης δειγματοληψίας μας, αυτές οι εγγενείς μεροληψίες στα υποκείμενα δεδομένα επηρεάζουν το τελικό σύνολο δεδομένων και πρέπει να λαμβάνονται υπόψη κατά την ερμηνεία των αποτελεσμάτων του μοντέλου.





Σχήμα 1.3: Επισκόπηση της κατανομής κατηγοριών ιστών στο σύνολο δεδομένων. Για σαφήνεια, εμφανίζονται μόνο οι 30 πιο συχνές κατηγορίες.

1.4 *Cell2Text* : Αρχιτεκτονική, Εκπαίδευση και Αξιολόγηση

Αρχιτεκτονική

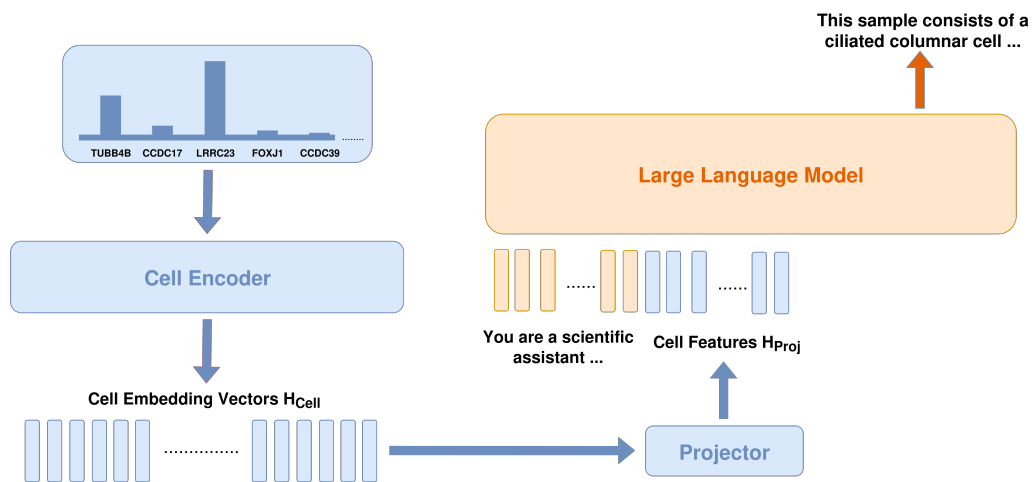
Το *Cell2Text* αποτελεί ένα πολυτροπικό παραγωγικό μοντέλο που στοχεύει στη μετατροπή δεδομένων γονιδιακής έκφρασης μεμονωμένων κυττάρων σε φυσική γλώσσα. Η αρχιτεκτονική του βασίζεται σε προσέγγιση *sequence-to-sequence (seq2seq)*, όπου η είσοδος είναι η ακολουθία γονιδίων ενός κυττάρου και η έξοδος μια παράγραφος που αποτυπώνει βιολογικές του ιδιότητες.

Σε υψηλό επίπεδο, το σύστημα αποτελείται από τρία βασικά υποσυστήματα:

- Έναν **κωδικοποιητή κυττάρων**, βασισμένο στο *Geneformer*, ο οποίος μετατρέπει τα δεδομένα έκφρασης γονιδίων σε ενσωματώσεις (*embeddings*) σε συνεχές λανθάνον χώρο.
- Ένα **μοντέλο προβολής**, το οποίο απεικονίζει τις ενσωματώσεις του κωδικοποιητή στον χώρο αναπαράστασης που απαιτεί ο αποκωδικοποιητής. Αποτελείται είτε από έναν πλήρως συνδεδεμένο νευρωνικό δίκτυο δύο επιπέδων είτε από έναν προσαρμοστικό μετασχηματιστή τύπου *Q-Former*, του οποίου ορισμένα βάρη (όπως τα αυτο-προσοχής) αρχικοποιούνται από ένα προ-εκπαιδευμένο μοντέλο γλώσσας (*PubmedBERT*).

- Έναν **αποκωδικοποιητή κειμένου**, βασισμένο στο *LLaMA*, ο οποίος παράγει φυσική γλώσσα με βάση τις απεικονισμένες ενσωματώσεις του κυττάρου και το κείμενο-πλαίσιο (*prompt*).

Η συνεργασία αυτών των τριών συστατικών επιτρέπει στο *Cell2Text* να «μεταφράζει» πολύπλοκες γονιδιακές εκφράσεις κυττάρων σε κατανοητές περιγραφές, γεφυρώνοντας το χάσμα μεταξύ μοριακής πληροφορίας και ανθρώπινης γλώσσας.



Σχήμα 1.4: Σχηματική απεικόνιση της αρχιτεκτονικής του μοντέλου *Cell2Text*.

Με αυτόν τον τρόπο, το *Cell2Text* λειτουργεί ως ένα πλήρως **παραγωγικό πολυτροπικό σύστημα**, ικανό να αποδίδει σε φυσική γλώσσα τη βιολογική πληροφορία που περιέχεται σε δεδομένα μονοκυτταρικής έκφρασης.

Εκπαίδευση

Η εκπαίδευση του *Cell2Text* έχει σχεδιαστεί ώστε να αξιοποιεί μεγάλα προεκπαιδευμένα μοντέλα με αποδοτικό τρόπο, προσαρμόζοντάς τα στο νέο έργο της παραγωγής φυσικής γλώσσας από δεδομένα γονιδιακής έκφρασης μεμονωμένων κυττάρων. Κατά την εκπαίδευση διαφορετικά μέρη του συστήματος παραμένουν παγωμένα ή προσαρμόζονται ανάλογα με τον ρόλο και τους διαθέσιμους πόρους.

Κωδικοποιητής Κυττάρων Ο κωδικοποιητής κυττάρων παραμένει παγωμένος σε όλη τη διάρκεια της εκπαίδευσης. Δεδομένου ότι έχει ήδη προεκπαιδευτεί σε εκατομμύρια γονιδιακές εκφράσεις μεμονωμένων κυττάρων, παρέχει πλούσιες και βιολογικά ουσιώδεις αναπαραστάσεις των γονιδίων. Κρατώντας τον σταθερού μειώνεται το υπολογιστικό κόστος και αποτρέπεται η απώλεια της βιολογικής γνώσης που έχει ήδη μάθει.

Μονάδα Προβολής Η μονάδα προβολής εκπαιδεύεται πλήρως, καθώς είναι υπεύθυνη για την ευθυγράμμιση των αναπαραστάσεων μεταξύ του κωδικοποιητή και του αποκωδικοποιητή.

Αποκωδικοποιητής Κειμένου Ο αποκωδικοποιητής (*LLaMA*) προσαρμόζεται με μία από τις ακόλουθες στρατηγικές:

- Πλήρης προσαρμογή (*full finetuning*): Ενημερώνονται όλες οι παράμετροι του *LLaMA*. Αυτή η επιλογή προσφέρει μέγιστη ευελιξία και δυνατότητα για λεπτομερή ειδική προσαρμογή στη νέα εργασία, αλλά απαιτεί μεγαλύτερους υπολογιστικούς πόρους και μνήμη.
- *Parameter – Efficient Finetuning (LoRA)*: Εισάγονται και εκπαιδεύονται μόνο μικρά, χαμηλής-τάξης *adapters* μέσα στους μηχανισμούς προσοχής του μετασχηματιστή, ενώ ο κύριος κορμός του *LLaMA* παραμένει παγωμένος. Η προσέγγιση αυτή μειώνει δραστικά τον αριθμό παραμέτρων που πρέπει να ενημερωθούν, επιτρέποντας ταχείες επαναλήψεις και εκπαίδευση σε περιορισμένο υλικό, χωρίς σημαντική υποβάθμιση στην απόδοση.

Ενσωμάτωση Εισόδου και Προτροπή (*prompt*) Κατά την εκπαίδευση, οι ενσωματώσεις του κυττάρου εισάγονται στο *LLaMA* μέσω ειδικών *placeholder tokens* στο κείμενο εισόδου. Το μοντέλο καθοδηγείται επιπλέον από μία σταθερή προτροπή:

«Είσαι επιστημονικός βοηθός εξειδικευμένος στην ανάλυση δεδομένων μονοκυτταρικής γονιδιακής έκφρασης και στη δημιουργία περιγραφικών περιλήψεων για την κατάσταση και τη λειτουργία των κυττάρων.»

Συνάρτηση Κόστους Η εκπαίδευση βασίζεται στη συνάρτηση κόστους αιτιώδους μοντελοποίησης γλώσσας (*causal language modeling loss*), δηλαδή τη διασταυρούμενη εντροπία (*cross – entropy loss*) μεταξύ των παραγόμενων και των πραγματικών *tokens*. Με αυτόν τον τρόπο, το μοντέλο μαθαίνει να προβλέπει το επόμενο *token* της περιγραφής, λαμβάνοντας υπόψη τόσο τα προηγούμενα όσο και τις ενσωματώσεις των κυττάρων.

Αξιολόγηση

Οντολογική Μετρική Ομοιότητας Τύπων Κυττάρου Βασισμένη στο *PageRank*

Για να αποτιμήσουμε τις προβλέψεις των μοντέλων μας με τρόπο που να λαμβάνει υπόψη τη βιολογική ιεραρχία των κυτταρικών τύπων, αναπτύσσουμε μια μετρική ομοιότητας βασισμένη στις σχέσεις “*is – a*” της οντολογίας *Cell Ontology (CL)*. Η *CL*

αναπαρίσταται ως ένας κατευθυνόμενος άκυκλικός γράφος, όπου οι κόμβοι αντιστοιχούν σε κυτταρικούς τύπους και οι ακμές περιγράφουν σχέσεις "is - a".

Για την ποσοτικοποίηση της συσχέτισης μεταξύ δύο κυτταρικών τύπων c_i και c_j , εφαρμόζουμε τον αλγόριθμο *Personalized PageRank (PPR)* στον γράφο της οντολογίας, οποίος έχει μετατραπεί σε μη κατευθυνόμενο. Η ιδέα είναι ότι ένας τυχαίος περίπατος ξεκινά από έναν κόμβο και κινείται κατά μήκος των σχέσεων της οντολογίας, με πιθανότητα επαναφοράς στον αρχικό κόμβο. Το αποτέλεσμα είναι ένα διάνυσμα κατανομής που αποτυπώνει πόσο "κοντά" βρίσκεται κάθε άλλος κόμβος στο σημείο εκκίνησης, λαμβάνοντας υπόψη όλη τη δομή της οντολογίας.

Η ομοιότητα μεταξύ δύο κυτταρικών τύπων ορίζεται ως:

$$S(c_i, c_j) \propto \log \left(1 + \frac{PPR(c_j | c_i)}{\tau} \right)$$

όπου τ είναι παράμετρος κλίμακας και $S(c_i, c_i) = 1$. Έτσι, οι τιμές $S(c_i, c_j)$ αντικατοπτρίζουν το βαθμό λειτουργικής και εξελικτικής συγγένειας μεταξύ των c_i και c_j βάσει της οντολογίας *CL*.

Μετρικές Ταξινόμησης

Για τα κλασικά έργα ταξινόμησης, όπως η πρόβλεψη τύπου κυττάρου, ιστού ή ασθένειας η απόδοση του μοντέλου αξιολογείται μέσω καθιερωμένων μετρικών όπως *accuracy*, *precision*, *recall* και το *F1 - score*. Οι μετρικές αυτές αποτυπώνουν τόσο τη συνολική ορθότητα όσο και την ισορροπία ανάμεσα στις κατηγορίες.

Μετρικές Πρόβλεψης Βιολογικών Οδών

Η αναγνώριση ενεργών βιολογικών οδών προσεγγίζεται ως ένα πρόβλημα πολλών ετικετών (*multi-label*), όπου κάθε κύτταρο μπορεί να σχετίζεται με περισσότερες από μία οδούς. Για την ολοκληρωμένη αξιολόγηση της απόδοσης, χρησιμοποιούμε τρεις συμπληρωματικές μετρικές: (1) την ακρίβεια υποσυνόλου (*subset accuracy*), που απαιτεί πλήρη ταύτιση μεταξύ προβλεπόμενων και πραγματικών συνόλων οδών, (2) την ομοιότητα *Jaccard*, η οποία αποδίδει μερική πίστωση για επικαλυπτόμενες προβλέψεις, και (3) το σταθμισμένο *F1 - score*, που υπολογίζεται ανά οδό και σταθμίζεται ως προς τη συχνότητά της, εξασφαλίζοντας δίκαιη αξιολόγηση τόσο για συχνές όσο και για σπάνιες βιολογικές οδούς.

Μετρικές Αξιολόγησης Παραγόμενου Κειμένου

Για την εκτίμηση της ποιότητας των περιγραφών που παράγει το μοντέλο, χρησιμοποιούμε ένα συνδυασμό λεξιλογικών και σημασιολογικών μετρικών. Οι *BLEU* και *ROUGE* μετρούν την επικάλυψη $n - gram$ και τη δομική ομοιότητα με τα κείμενα αναφοράς, αξιολογώντας τη ροή και την πληρότητα περιεχομένου. Για τη σημασιολογική πιστότητα, ιδιαίτερα κρίσιμη στη βιοϊατρική ορολογία, χρησιμοποιούμε το

BERTScore, βασισμένο σε προεκπαιδευμένα μοντέλα *RoBERTa* (γενικής γλώσσας) και *BioBERT* (βιοϊατρικού τομέα). Ο συνδυασμός αυτών των μετρικών εξασφαλίζει ότι οι παραγόμενες περιγραφές είναι όχι μόνο γλωσσικά ορθές και συνεκτικές, αλλά και επιστημονικά συναφείς.

1.5 Αποτελέσματα

Ποιότητα Παραγόμενου Κειμένου

Πίνακας 1.1: Μετρικές ποιότητας παραγόμενου κειμένου ανά μοντέλο. Υψηλότερες τιμές υποδηλώνουν καλύτερη ποιότητα παραγωγής κειμένου.

Μοντέλο	$B - 4$	$R - 2$	$R - L$	BBT	RBT
<i>LLaMA</i> – $3B + LoRA + MLP$	48.83	57.26	63.38	86.15	89.75
<i>LLaMA</i> – $3B + LoRA + Q - Former$	<u>57.46</u>	<u>65.57</u>	70.56	90.01	93.56
<i>LLaMA</i> – $1B + FSDP + MLP$	57.84	65.99	69.87	88.43	91.56
<i>LLaMA</i> – $1B + FSDP + Q - Former$	56.78	65.25	<u>70.47</u>	<u>89.85</u>	<u>93.53</u>

Η ποιότητα του παραγόμενου κειμένου αξιολογήθηκε μέσω μετρικών *BLEU*, *ROUGE* και *F1* βάσει βιοϊατρικών κειμενικών ενσωματώσεων. Το μοντέλο *LLaMA* – $3B + LoRA + Q - Former$ πέτυχε τη συνολικά καλύτερη σημασιολογική απόδοση (*BioBERT* – $F1 = 90.01\%$, *RoBERTa* – $F1 = 93.56\%$), δείχνοντας αυξημένη ικανότητα κατανόησης βιοϊατρικού περιεχομένου.

Αντιθέτως, το *LLaMA* – $1B + FSDP + MLP$ υπερείχε στις $n - gram$ μετρικές ($BLEU - 4 = 57.84\%$, $ROUGE - 2 = 65.99\%$), αποδεικνύοντας ότι η πλήρης εκπαίδευση όλων των παραμέτρων μπορεί να αντισταθμίσει την απλότητα του *MLP projector*. Γενικά, ο *Q - Former* προσφέρει πλεονεκτήματα σε σημασιολογικό επίπεδο, ιδίως όταν το μοντέλο εκπαιδεύεται με περιορισμένο αριθμό παραμέτρων (*LoRA*).

Αναγνώριση Τύπου Κυττάρου

Πίνακας 1.2: Απόδοση αναγνώρισης τύπου κυττάρου. *Ont. Sim.*: Οντολογική ομοιότητα

Μοντέλο	<i>Acc</i>	<i>F1</i>	<i>Ont. Sim.</i>
<i>Geneformer(Baseline)</i>	64.52	62.23	75.21
<i>scGPT(Baseline)</i>	66.55	65.01	75.63
<i>LLaMA – 3B + LoRA + MLP</i>	59.50	62.82	70.50
<i>LLaMA – 3B + LoRA + Q – Former</i>	67.83	68.52	78.52
<i>LLaMA – 1B + FSDP + MLP</i>	67.95	72.76	73.49
<i>LLaMA – 1B + FSDP + Q – Former</i>	64.21	65.66	<u>75.86</u>

Η αναγνώριση τύπου κυττάρου αποτελεί μία από τις πιο απαιτητικές εργασίες. Το *LLaMA – 1B + FSDP + MLP* πέτυχε την υψηλότερη ακρίβεια (*Acc*) (67.95%) και *F1* (72.76%), ξεπερνώντας και το ισχυρότερο *baseline* του *scGPT* (66.55% *Acc*, 65.01% *F1*).

Το *LLaMA – 3B + Q – Former* εμφάνισε τη μέγιστη οντολογική ομοιότητα (78.52%), ξεπερνώντας όλα τα μοντέλα, συμπεριλαμβανομένου του *scGPT* (75.63%), με σημαντική διαφορά. Είναι αξιοσημείωτο ότι το *LLaMA – 1B + FSDP + MLP*, παρότι πρώτο σε *Acc/F1*, υστερεί σε οντολογική ομοιότητα (73.49%) έναντι του *scGPT* (75.63%), αναδεικνύοντας έναν σαφή συμβιβασμό (*trade – off*) μεταξύ απόλυτης ακρίβειας και σημασιολογικής κατανόησης.

Πρόβλεψη Ασθένειας και Ιστού

Πίνακας 1.3: Απόδοση πρόβλεψης ασθένειας και ιστού.

Μοντέλο	Ασθένεια		Ιστός	
	<i>Acc</i>	<i>F1</i>	<i>Acc</i>	<i>F1</i>
<i>Geneformer(Baseline)</i>	62.50	59.82	59.63	58.31
<i>scGPT(Baseline)</i>	<u>63.15</u>	<u>62.12</u>	<u>60.56</u>	<u>60.25</u>
<i>LLaMA – 3B + LoRA + MLP</i>	61.13	66.32	59.23	64.17
<i>LLaMA – 3B + LoRA + Q – Former</i>	63.80	67.76	61.65	65.79
<i>LLaMA – 1B + FSDP + MLP</i>	65.34	70.45	62.39	67.88
<i>LLaMA – 1B + FSDP + Q – Former</i>	62.16	65.23	58.05	61.47

Η πρόβλεψη ασθένειας και ιστού επιβεβαιώνει τις τάσεις της προηγούμενης ενότητας. Το *LLaMA – 1B + FSDP + MLP* πέτυχε την καλύτερη συνολική απόδοση

(ασθένεια: 70.45% $\Phi 1$, ιστός: 67.88% $F1$), βελτιώνοντας το ισχυρότερο *baseline* του *scGPT* (ασθένεια: 62.12% $\Phi 1$, ιστός: 60.25% $F1$) κατά 7.5–8.5 ποσοστιαίες μονάδες. Η πρόβλεψη ασθένειας ήταν ευκολότερη από την πρόβλεψη ιστού, πιθανώς λόγω πιο διακριτών γονιδιακών προτύπων.

Το *LLaMA-3B+Q-Former* ακολουθεί με σταθερά δεύτερη επίδοση, επιβεβαιώνοντας ότι η συνδυαστική χρήση *LoRA* και *Q-Former* ενισχύει τη σημασιολογική γενίκευση, παρότι το μικρότερο μοντέλο με πλήρη εκπαίδευση υπερτερεί σε ακρίβεια.

Εντοπισμός Ενεργών Βιολογικών Οδών

Πίνακας 1.4: Απόδοση πρόβλεξης ενεργών βιολογικών οδών. *Sub. Acc.*: Ακρίβεια πλήρους συνόλου, *Jac. Sim.*: Ομοιότητα *Jaccard*, *W - F1*: Σταθμισμένο $F1$.

Μοντέλο	<i>Sub. Acc.</i>	<i>Jac. Sim.</i>	<i>W - F1</i>
<i>Geneformer(Baseline)</i>	35.55	49.44	56.61
<i>scGPT(Baseline)</i>	<u>36.13</u>	<u>50.13</u>	<u>58.35</u>
<i>LLaMA - 3B + LoRA + MLP</i>	30.02	42.77	51.34
<i>LLaMA - 3B + LoRA + Q - Former</i>	36.02	49.58	57.96
<i>LLaMA - 1B + FSDP + MLP</i>	36.42	50.67	59.10
<i>LLaMA - 1B + FSDP + Q - Former</i>	35.75	48.57	57.12

Ο εντοπισμός ενεργών βιολογικών οδών αποτελεί τη δυσκολότερη εργασία, με ακρίβειες γύρω στο 35–36%. Το *LLaMA - 1B + FSDP + MLP* διατήρησε την υπεροχή του, αν και με οριακή διαφορά από το ισχυρό *scGPT baseline* (*Sub. Acc.* = 36.42% έναντι 36.13%). Αντίθετα, το *3B + LoRA + MLP* παρουσίασε σημαντική πτώση, υπολειπόμενο κατά πολύ και των δύο βασελινες, υποδεικνύοντας ότι η περιορισμένη εκπαίδευση (*LoRA*) απαιτεί πιο εκφραστικούς προβολείς όπως ο *Q - Former*.

Το χάσμα μεταξύ *Subset Accuracy* και *Jaccard Similarity* δείχνει ότι τα μοντέλα συχνά εντοπίζουν τουλάχιστον μία σωστή οδό, αλλά δυσκολεύονται στην πλήρη ταύτιση των δύο πιο ενεργών. Παρ' όλα αυτά, η (οριακή) βελτίωση του κορυφαίου μοντέλου μας έναντι του *scGPT* δείχνει ότι το προτεινόμενο σύστημα μπορεί να αναπαραστήσει σύνθετα μεταβολικά πρότυπα σε μονοκύτταρα δεδομένα.

1.6 Συμπεράσματα και Μελλοντικές Επεκτάσεις

Συμπεράσματα

Η παρούσα εργασία παρουσίασε το *Cell2Text*, ένα πολυτροπικό γενετικό πλαίσιο που συνδέει τη μονοκυτταρική μεταγραφική ανάλυση με τη φυσική γλώσσα, παράγοντας επεξηγήσιμες περιγραφές κυττάρων απευθείας από γονιδιακά δεδομένα. Μέσω του συνδυασμού προεκπαιδευμένων βιολογικών μοντέλων με μεγάλα γλωσσικά μοντέλα το *Cell2Text* μετασχηματίζει την παραδοσιακή ταξινόμηση κυττάρων σε μια παραγωγική επεξηγήσιμη διαδικασία.

Τα αποτελέσματα έδειξαν ότι η γενετική εκπαίδευση με στόχο την παραγωγή φυσικής γλώσσας οδηγεί σε πλουσιότερες και πιο συνεκτικές βιολογικές αναπαραστάσεις σε σχέση με τις καθαρά διακριτικές προσεγγίσεις. Το μοντέλο δεν περιορίζεται σε μεμονωμένες προβλέψεις, αλλά μαθαίνει κοινή σημασιολογική αναπαράσταση που συνδέει έκφραση γονιδίων, κυτταρική λειτουργία, ιστικό πλαίσιο και ασθένειες.

Οι μετρικές οντολογικής ομοιότητας έδειξαν ότι το *Cell2Text* αναπτύσσει βιολογικά συνεπή κατανόηση των σχέσεων μεταξύ τύπων κυττάρων: οι περισσότερες λανθασμένες προβλέψεις αντιστοιχούν σε στενά συγγενείς τύπους, γεγονός που υποδηλώνει εσωτερίκευση της ιεραρχικής δομής της κυτταρικής οντολογίας.

Συνολικά, το *Cell2Text* αποδεικνύει ότι ο συνδυασμός θεμελιωδών βιολογικών μοντέλων και μεγάλων γλωσσικών μοντέλων αποτελεί μια επεκτάσιμη και ερμηνεύσιμη στρατηγική για τη βιοπληροφορική. Η προσέγγιση αυτή ανοίγει τον δρόμο για εφαρμογές σε ευρύτερες περιοχές, όπως η χωρική μεταγραφική ανάλυση, η πρόβλεψη φαρμακευτικών αποκρίσεων και η αυτόματη σύνθεση επιστημονικής γνώσης.

Μελλοντικές Επεκτάσεις

Παρότι το *Cell2Text* παρουσιάζει ιδιαίτερα ενθαρρυντικά αποτελέσματα, υπάρχουν αρκετές κατευθύνσεις για μελλοντική βελτίωση και διεύρυνση:

Ενσωμάτωση πρωτεϊνικής έκφρασης: Η επέκταση του πλαισίου ώστε να περιλαμβάνει δεδομένα *CITE-seq* θα επέτρεπε πιο πλήρη φαινοτυπική περιγραφή, βελτιώνοντας την ακρίβεια ειδικά σε ανοσολογικά κύτταρα.

Μάθηση με αναλυτική ιεραρχία (*Curriculum Learning*): Η εκπαίδευση με προοδευτική εισαγωγή κυτταρικών τύπων από γενικούς σε εξειδικευμένους μπορεί να ενισχύσει την κατανόηση των ιεραρχικών σχέσεων στην οντολογία.

Επέκταση σε χωρικά και χρονικά δεδομένα: Η ενσωμάτωση πληροφοριών χωρικής οργάνωσης ή χρονικής εξέλιξης θα επέτρεπε περιγραφές που αποτυπώνουν την κυτταρική θέση, ανάπτυξη ή παθολογική πορεία.

Διαδραστική ανατροφοδότηση: Η ενσωμάτωση σχολίων ειδικών (μέσω *RLHF*) θα μπορούσε να ευθυγραμμίσει τις παραγόμενες περιγραφές με την επιστημονική γνώση και πρακτικές ανάγκες των ερευνητών.

Δια-ειδική γενίκευση: Η αξιολόγηση της μεταφερσιμότητας του *Cell2Text* μεταξύ ειδών (π.χ. άνθρωπος–ποντίκι) θα δοκιμάσει την καθολικότητα των αναπαραστάσεών του και θα αναδεικνύει θεμελιώδεις αρχές κυτταρικής βιολογίας.

Οι παραπάνω κατευθύνσεις αναδεικνύουν το ευρύ δυναμικό εξέλιξης των πολυτροπικών προσεγγίσεων στη βιοπληροφορική, με το *Cell2Text* να θέτει τα θεμέλια για πιο έξυπνα, επεξηγήσιμα και συνεργατικά συστήματα τεχνητής νοημοσύνης στη μελέτη των κυτταρικών καταστάσεων.

Chapter 2

Introduction

2.1 Motivation

Single-cell RNA sequencing (scRNA-seq) has transformed biology by enabling measurement of gene expression at the level of individual cells. Unlike bulk RNA-seq, which averages signals across populations of cells, scRNA-seq reveals cellular heterogeneity, providing a powerful lens for studying development, tissue organization, disease mechanisms, and therapeutic response. Its applications span immunology, oncology, neuroscience, and regenerative medicine, making it a cornerstone of modern systems biology.

However, the interpretation of scRNA-seq data presents major challenges. Each cell is represented by thousands of noisy, sparse gene expression values, forming a high-dimensional space that is difficult to analyze directly. Current workflows rely on dimensionality reduction, clustering, and manual annotation with marker genes, a process that is labor-intensive, subjective, and difficult to scale to the millions of cells generated by contemporary studies. Furthermore, annotations typically collapse complex biology into rigid labels (e.g., “CD4 T cell”), discarding information about functional states, tissue specificity, disease relevance, and pathway activity.

Recent advances in deep learning have produced single-cell foundation models, such as Geneformer [1] and scGPT [2], which learn transferable representations of cells directly from gene expression profiles. These models improve performance on classification and clustering tasks, and their embeddings capture rich biological structure. Yet, their outputs remain constrained: they rely on downstream classifiers that map embeddings to predefined categories. This dependence on fixed label sets limits interpretability, prevents generalization to novel cell types, and makes model adaptation computationally expensive. In practice, the most biologically relevant information captured in embeddings is hidden from researchers, as outputs remain categorical and opaque.

2.2 Problem Statement

The central problem addressed in this thesis is the **mismatch between the richness of single-cell expression data and the limited, discrete outputs of existing computational models**. While single-cell foundation models capture complex representations of cellular identity, their reliance on classification heads collapses this information into narrow categories. This creates several bottlenecks:

- **Limited expressivity:** Cells are reduced to static labels, obscuring their state, tissue context, and pathway activity.
- **Poor generalization:** Models struggle with unseen or rare cell types, requiring retraining or fine-tuning.
- **Low interpretability:** Categorical outputs provide little explanatory context, making it difficult for biologists to derive mechanistic insights.

What is missing is a computational framework that can *translate* complex single-cell embeddings into flexible, interpretable, and biologically meaningful outputs. Natural language offers a powerful solution: it can describe multiple facets of cellular identity simultaneously, without being bound to rigid ontologies, and provides a medium that is inherently accessible to human interpretation.

2.3 Contributions

This thesis proposes **Cell2Text**, a multimodal generative framework that bridges single-cell foundation models and large language models (LLMs) to translate gene expression profiles into structured natural language descriptions. The main contributions are as follows:

- **Framework design:** We develop Cell2Text, a multimodal generative framework that aligns gene-level embeddings from single-cell foundation models with instruction-tuned LLMs, enabling the generation of coherent, biologically grounded natural language descriptions of cells.
- **Predictive performance:** We demonstrate that Cell2Text achieves competitive results across cell type, tissue, disease, and pathway classification tasks. Ontology-aware evaluation shows that even misclassifications maintain biological relevance, reflecting the model’s ability to capture semantic structure.
- **Interpretability through generation:** We show that Cell2Text produces high-quality natural language descriptions with strong semantic fidelity, capturing cellular identity, tissue origin, disease associations, and pathway activity in a way that is interpretable and scientifically meaningful.

- **Dataset construction:** We curate a large-scale multimodal dataset of over one million cells from CELLxGENE, enriched with ontology terms, tissue and disease metadata, and pathway annotations. This resource enables cross-modal training and evaluation at a scale not available in existing datasets.
- **Scalability and label efficiency:** We demonstrate that coupling single-cell embeddings with natural language enables scalable characterization of unseen cells, reducing dependence on large annotated reference atlases and costly retraining.

Taken together, these contributions show that moving beyond static labels toward natural language generation unlocks both stronger predictive performance and inherently interpretable outputs, offering a new paradigm for scalable and transparent single-cell analysis.

2.4 Thesis Structure

The remainder of this thesis is organized as follows. Chapter 3 provides the theoretical background, covering single-cell biology, machine learning fundamentals, foundation models for genomics, large language models, and related multimodal frameworks. Chapter 4 describes the construction of our multimodal dataset, including data collection, sampling strategies, and automated text generation. Chapter 5 presents the Cell2Text architecture, training methodology, and evaluation metrics. Chapter 6 reports experimental results across text generation quality, biological entity recognition, and pathway prediction tasks. Finally, Chapter 7 concludes with a summary of contributions and directions for future work.

Chapter 3

Theoretical Background

3.1 Single-cell Biology

3.1.1 Introduction

Cells are the fundamental units of life, and their molecular states determine their function within tissues and organisms. Traditional molecular profiling techniques, such as *bulk* RNA sequencing (RNA-seq), measure the average gene expression across a large population of cells [12]. While bulk RNA-seq has been invaluable for identifying differentially expressed genes between conditions, it cannot resolve cell-to-cell variability, and rare cell types or transient cellular states may be completely masked in the population average.

Single-cell RNA sequencing (scRNA-seq), first demonstrated by Tang *et al.* in 2009 [3], enables transcriptome-wide quantification at the resolution of individual cells. By capturing and sequencing RNA from individual cells, scRNA-seq reveals heterogeneity in cell types, states, and lineages that is invisible to bulk approaches. This capability has transformed fields such as developmental biology, immunology, and cancer research, and has been instrumental in large-scale efforts such as the Human Cell Atlas [13].

3.1.2 scRNA-seq Technologies

Multiple experimental protocols exist for scRNA-seq, each with trade-offs in throughput, sensitivity, and cost.

- **Plate-based methods** (e.g., SMART-seq2 [14]) capture full-length transcripts with high sensitivity, enabling detection of low-abundance genes, but are low-throughput and expensive.
- **Droplet-based methods** (e.g., Drop-seq [15], 10x Genomics [16]) encapsulate individual cells in microfluidic droplets with barcoded beads, enabling high-throughput profiling at the cost of 3'-end bias and sparser coverage.

- **Unique Molecular Identifiers (UMIs)** are short random sequences incorporated during library preparation to distinguish original RNA molecules from PCR duplicates, improving quantitative accuracy.

3.1.3 Data Characteristics and Challenges

The raw output of an scRNA-seq experiment is a cell-by-gene count matrix, where each entry represents the number of sequencing reads (or UMI counts) mapping to a particular gene in a particular cell. However, scRNA-seq data have several distinctive characteristics:

- **High dimensionality:** Tens of thousands of features (genes) per cell.
- **Sparsity:** Most entries are zero due to biological low expression and technical dropouts.
- **Batch effects:** Systematic differences caused by variations in experimental conditions, sequencing runs, or labs.
- **Count-based distribution:** Count-based distribution: Gene counts in scRNA-seq are non-negative integers and exhibit overdispersion, meaning the variance exceeds the mean—this makes the negative binomial distribution a more appropriate statistical model than the simple Poisson distribution [17].

These properties complicate downstream computational analysis and motivate the use of specialized normalization and representation learning methods.

3.1.4 Quality Control and Normalization

Before analysis, rigorous quality control (QC) is applied to remove low-quality cells and genes:

- **Cell filtering:** Removing cells with too few detected genes, too few counts, or high proportions of mitochondrial transcripts (indicative of stressed or dying cells).
- **Doublet detection:** Identifying droplets containing more than one cell using algorithms such as Scrublet [18].
- **Gene filtering:** Removing genes detected in only a few cells to reduce noise.

Normalization adjusts for differences in sequencing depth between cells. Common approaches include:

- **Library size normalization:** Counts per million (CPM), transcripts per million (TPM).
- **Log-transformation:** Stabilizes variance and reduces the effect of outliers.

- **Variance stabilization methods:** e.g., SCTransform [17] models counts with a regularized negative binomial regression.

3.1.5 Dimensionality Reduction and Clustering

Given the high dimensionality, dimensionality reduction is essential before visualization and clustering:

- **PCA** is often applied first to capture the main sources of variance.
- **t-SNE** and **UMAP** [19] provide non-linear embeddings for visualization.
- **Graph-based clustering** methods such as Louvain or Leiden [20] partition the cell graph into putative cell populations.

Cluster identities are assigned through **marker gene analysis**, where genes with significantly higher expression in one cluster are matched to known cell type markers.

3.1.6 Extensions and Multimodal Data

Recent advances integrate scRNA-seq with additional molecular measurements:

- **CITE-seq** [21] simultaneously measures transcriptomes and selected cell-surface proteins.
- **scATAC-seq** [22] captures chromatin accessibility at single-cell resolution.
- **Spatial transcriptomics** [23] preserves spatial location of cells within tissues while profiling gene expression.

These multimodal datasets offer richer biological context and create new opportunities for computational models capable of cross-modal translation.

3.1.7 Biological Ontologies and Standardized Annotations

Biological knowledge is inherently complex and hierarchical, requiring structured vocabularies called **biological ontologies** to provide standardized terminology and relationships between biological concepts. The **Open Biological and Biomedical Ontologies (OBO) Foundry** [24] maintains high-quality, interoperable ontologies including the Gene Ontology (GO) for molecular functions and biological processes, the Cell Ontology (CL) for standardized cell type definitions, and the Human Phenotype Ontology (HPO) for phenotypic descriptions.

In single-cell analysis, ontologies provide structured textual descriptions of cell types, functions, and biological processes, enabling systematic annotation beyond traditional manual curation approaches. These standardized vocabularies facilitate automated annotation methods and enable integration of diverse datasets through common semantic frameworks.

3.1.8 Pathway-based Annotation

In addition to ontologies, **biological pathways** provide structured representations of gene sets that collectively carry out specific cellular functions or participate in molecular signaling cascades. Pathway-based analysis enables the interpretation of single-cell gene expression profiles at a systems level, moving beyond individual marker genes to functionally coherent modules.

Several curated pathway databases are widely used in single-cell transcriptomics, including **KEGG** [25], **Reactome** [26], and the **Molecular Signatures Database (MSigDB)** [27]. Among these, the **Hallmark gene sets** from MSigDB provide a compact and interpretable representation of cellular states by summarizing canonical biological processes into 50 well-defined pathways.

To leverage such resources, computational methods are required to evaluate pathway activity at the resolution of individual cells. One widely adopted approach is **AUCell** [11], which estimates whether the genes of a given pathway are enriched among the most highly expressed genes of each cell. The algorithm proceeds in three steps:

- Rank each cell’s gene expression values, producing a per-cell ordered gene list.
- For a given gene set, compute the *Area Under the Curve (AUC)* of the recovery curve, which reflects the fraction of its genes detected within the highly expressed portion of the ranked list.
- Use the AUC value as a pathway activity score for that cell, enabling functional annotation at single-cell resolution.

Unlike bulk enrichment approaches, AUCell is specifically designed for the sparsity and heterogeneity of scRNA-seq data. It provides a robust means of identifying active regulatory programs in subpopulations of cells and is readily integrated into downstream analyses such as clustering, visualization, or pathway-based annotation. In this way, AUCell bridges curated pathway resources with single-cell functional interpretation, linking expression patterns to biological processes.

3.2 Machine Learning

3.2.1 Introduction

Machine learning is a branch of artificial intelligence that enables computers to learn patterns from data without being explicitly programmed for specific tasks. Rather than following pre-written instructions, machine learning algorithms automatically improve their performance on a given task through experience with training data.

3.2.2 Learning Paradigms

Machine learning approaches can be categorized based on the nature of the learning task and the availability of labeled data:

- **Supervised Learning:** Algorithms learn from labeled training examples to make predictions on new data. The dataset contains input-output pairs, and the goal is to approximate the mapping function. Common tasks include classification (predicting categories) and regression (predicting continuous values). Examples include linear regression, decision trees, and support vector machines.
- **Unsupervised Learning:** Algorithms discover hidden patterns in data without access to labels. These methods find structure in datasets where only input data is available. Key tasks include clustering (grouping similar data points), dimensionality reduction (finding lower-dimensional representations), and association rule learning.
- **Semi-supervised Learning:** Combines small amounts of labeled data with larger amounts of unlabeled data to improve learning performance. This approach is valuable when labeling data is expensive or time-consuming.
- **Reinforcement Learning:** Algorithms learn optimal actions through trial and error interactions with an environment, receiving rewards or penalties for different actions.

3.2.3 Key Concepts and Algorithms

Several fundamental concepts underpin machine learning:

- **Training and Testing:** Models are trained on a subset of data and evaluated on held-out test data to assess generalization performance.
- **Overfitting and Underfitting:** Overfitting occurs when a model learns the training data too specifically and fails to generalize. Underfitting happens when a model is too simple to capture underlying patterns.
- **Cross-validation:** A technique for assessing model performance by repeatedly splitting data into training and validation sets.
- **Feature Selection and Engineering:** The process of selecting relevant input variables and creating new features that better represent the underlying problem.

Common algorithms include:

- **Linear Models:** Linear regression, logistic regression, and regularized variants (Ridge, Lasso).

- **Tree-based Methods:** Decision trees, random forests, and gradient boosting machines.
- **Instance-based Learning:** k-nearest neighbors and kernel methods.
- **Clustering:** k-means, hierarchical clustering, and density-based methods.
- **Dimensionality Reduction:** Principal Component Analysis (PCA), t-SNE, and UMAP.

3.2.4 Model Evaluation and Validation

Proper evaluation is crucial for assessing machine learning model performance:

- **Classification Metrics:** Accuracy, precision, recall, F1-score, and area under the ROC curve (AUC).
- **Regression Metrics:** Mean squared error, mean absolute error, and coefficient of determination (R^2).
- **Cross-validation:** k-fold cross-validation provides robust performance estimates by averaging results across multiple train-test splits.

3.3 Deep Learning and Neural Networks

Deep learning has emerged as a dominant paradigm in machine learning, achieving state-of-the-art performance across domains including computer vision, natural language processing, and computational biology. Unlike traditional machine learning algorithms that rely heavily on manual feature engineering, deep learning methods learn hierarchical representations directly from raw data through multi-layered neural architectures [28].

In this section, we review the main building blocks of deep learning relevant to this work, covering fundamental architectures and optimization techniques.

3.3.1 Optimization in Deep Neural Networks

Training deep neural networks involves solving a high-dimensional, non-convex optimization problem, where the goal is to find parameters θ that minimize a task-specific loss function $\mathcal{L}(\theta)$ over data distribution $p(x, y)$:

$$\theta^* = \arg \min_{\theta} \mathbb{E}_{(x,y) \sim p} \mathcal{L}(f_{\theta}(x), y).$$

Unlike convex optimization, deep networks present highly non-linear landscapes with multiple local minima and saddle points [29]. Despite this, simple first-order methods have proven surprisingly effective in practice.

Gradient-based methods. The backbone of modern optimization is Stochastic Gradient Descent (SGD) [30], which updates parameters in the direction of the negative gradient of the loss, estimated from mini-batches of data:

$$\theta \leftarrow \theta - \eta \nabla_{\theta} \mathcal{L}(\theta; \mathcal{B}),$$

where η is the learning rate and $\mathcal{B}$ denotes a mini-batch. Variants of SGD include momentum [31], which accelerates convergence by accumulating a velocity term, and Nesterov momentum [32], which anticipates future gradients.

Adaptive methods. To cope with heterogeneous curvature in the parameter space, adaptive learning rate methods have been introduced. AdaGrad [33] rescales updates according to the historical accumulation of squared gradients, RMSProp [34] uses an exponential moving average, and Adam [35] combines momentum with adaptive learning rates, becoming one of the most widely adopted optimizers in practice.

Regularization and normalization. To improve generalization, various regularization techniques are applied during training. Dropout randomly zeroes out activations during training to prevent co-adaptation [36], while weight decay penalizes large parameter values. Batch Normalization [37] normalizes intermediate activations across mini-batches, stabilizing gradients and enabling faster training. Alternatives such as Layer Normalization [38] and Group Normalization [39] have been introduced for settings where batch statistics are less effective.

Learning rate schedules. The learning rate plays a crucial role in optimization. Fixed rates can lead to suboptimal convergence, motivating dynamic schedules such as step decay, cosine annealing [40], or cyclical learning rates [41], which allow faster convergence and help escape sharp minima. Recent work emphasizes the role of flat minima in improving generalization [42].

Overall, optimization in deep learning represents a delicate balance between algorithmic choices, network architecture, and data characteristics. The effectiveness of relatively simple first-order methods remains an open and fascinating aspect of deep learning theory.

3.3.2 Feedforward Neural Networks

The simplest form of deep learning models are feedforward neural networks, also known as multilayer perceptrons (MLPs). An MLP consists of layers of interconnected neurons, where each layer performs a linear transformation followed by a non-linear activation:

$$h^{(l)} = \sigma(W^{(l)}h^{(l-1)} + b^{(l)}),$$

where $h^{(l)}$ is the activation vector at layer l , $W^{(l)}$ and $b^{(l)}$ are learnable weights and biases, and σ is a non-linear activation function.

Expressivity. The universal approximation theorem shows that even a single hidden layer network with sufficient units can approximate any continuous function on compact subsets of $\mathbb{R}^n$ [43]. In practice, deeper architectures are more parameter-efficient and enable hierarchical representation learning, where successive layers capture increasingly abstract features.

Activation functions. Early networks used sigmoidal activations such as the logistic or hyperbolic tangent functions. These suffered from vanishing gradient issues, limiting the depth of trainable networks. The introduction of the Rectified Linear Unit (ReLU) [44] revolutionized deep learning by providing sparse activations and alleviating vanishing gradients. Variants such as Leaky ReLU, ELU, and GELU [45] have been proposed to further improve training stability.

Historical context and limitations. Feedforward networks date back to the perceptron model of Rosenblatt (1958) [46], which sparked early interest in neural computation. However, simple perceptrons were limited to linearly separable tasks, as highlighted by Minsky and Papert (1969) [47]. The introduction of backpropagation [48] enabled training of multilayer perceptrons, rekindling interest in neural networks. Despite their flexibility, MLPs do not exploit structure in data such as locality (images), sequentiality (text), or relational dependencies (graphs), motivating specialized architectures such as CNNs, RNNs, and GNNs.

Feedforward neural networks remain a fundamental component of deep learning research, serving as the building blocks for more complex architectures and as versatile baselines for structured and unstructured data alike.

3.3.3 Convolutional Neural Networks (CNNs)

Convolutional Neural Networks (CNNs) are designed for data with grid-like structure, such as images, by exploiting local spatial correlations. A convolutional layer applies a set of learnable kernels K that are convolved with the input X , producing feature maps:

$$h_{i,j}^{(k)} = \sigma \left(\sum_{u,v} K_{u,v}^{(k)} \cdot X_{i+u,j+v} + b^{(k)} \right),$$

where (i, j) indexes spatial location and k denotes the feature channel. This operation introduces two crucial concepts: *local receptive fields*, which restrict each neuron to only a small neighborhood of the input, and *parameter sharing*, which dramatically reduces the number of learnable weights compared to fully connected layers. Pooling operations (e.g., max pooling, average pooling) further introduce translational invariance and dimensionality reduction.

CNNs were originally inspired by the hierarchical organization of the visual cortex [49]. Their modern resurgence began with the success of AlexNet [50] on

the ImageNet challenge, which demonstrated the effectiveness of deep convolutional architectures trained on GPUs. Subsequent developments such as VGG networks [51] and residual networks (ResNets) [52], enabled much deeper and more efficient CNNs. While most commonly applied to image recognition, CNNs have also been adapted to one-dimensional genomic sequences, chromatin accessibility profiles, and spatial transcriptomics data.

3.3.4 Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs) extend neural networks to handle sequential data by introducing hidden states that evolve over time. Given an input sequence $\{x_t\}_{t=1}^T$, the hidden state h_t at time t is updated according to:

$$h_t = \sigma(W_h h_{t-1} + W_x x_t + b),$$

where W_h and W_x are learnable weight matrices. The recursive definition enables parameter sharing across time steps, allowing the network to capture dependencies in variable-length sequences. However, standard RNNs suffer from vanishing and exploding gradients during backpropagation through time, making it difficult to learn long-range dependencies [53].

To address this, gated architectures were introduced. The Long Short-Term Memory (LSTM) network [54] incorporates input, output, and forget gates that control the flow of information through a memory cell, enabling robust modeling of long sequences. RNNs dominated natural language processing prior to the advent of Transformers, powering models for language modeling, translation, and speech recognition.

3.3.5 Graph Neural Networks (GNNs)

Graph Neural Networks (GNNs) generalize neural architectures to data that can be represented as graphs, i.e., sets of nodes V and edges E . A common paradigm is message passing, where node representations are iteratively updated by aggregating features from neighbors. For a node v , its updated embedding at layer $l + 1$ is given by:

$$h_v^{(l+1)} = \sigma \left(W^{(l)} \cdot \text{AGG}(\{h_v^{(l)}\} \cup \{h_u^{(l)} : u \in \mathcal{N}(v)\}) \right),$$

where $\mathcal{N}(v)$ is the neighborhood of v and AGG is a permutation-invariant aggregation function (e.g., sum, mean, attention). This allows GNNs to learn representations that incorporate both node attributes and graph structure.

The foundational idea was introduced by Scarselli et al. [55], but modern scalable formulations emerged with Graph Convolutional Networks (GCNs) [56], which approximate spectral convolutions on graphs, and Graph Attention Networks

(GATs) [57], which incorporate attention mechanisms for adaptive weighting of neighbors. GNNs have achieved strong results in social network analysis, chemistry, and knowledge graphs.

3.3.6 Attention Mechanisms and Transformers

Motivation. While RNNs and CNNs provided powerful tools for sequence modeling, they suffer from inherent limitations. RNNs process inputs sequentially, leading to difficulties in capturing long-range dependencies due to vanishing gradients and limited parallelism. CNNs, although more parallelizable, require large receptive fields (i.e., many stacked layers) to model long-distance relationships. Attention mechanisms [58] were introduced as a remedy, enabling models to learn dynamic, content-based weighting over input elements, rather than relying solely on fixed-size context windows.

Attention Mechanism. The basic attention formulation takes a *query* vector q , a set of *keys* $\{k_i\}$, and corresponding *values* $\{v_i\}$. The attention output is a weighted sum of the values:

$$\text{Attention}(q, K, V) = \sum_{i=1}^n \alpha_i v_i, \quad \alpha_i = \frac{\exp\left(\frac{q \cdot k_i}{\sqrt{d_k}}\right)}{\sum_{j=1}^n \exp\left(\frac{q \cdot k_j}{\sqrt{d_k}}\right)},$$

where d_k is the dimension of the key vectors. This softmax-normalized dot product assigns higher weights to keys more similar to the query. The mechanism allows the model to “attend” to relevant parts of the input when generating each output.

Self-Attention and Multi-Head Attention. The Transformer [6] extends this idea through *self-attention*, where queries, keys, and values are learned linear projections of the same input sequence. This enables each element in the sequence to attend to all others, capturing global dependencies. To enrich representation capacity, *multi-head attention* is introduced:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O,$$

where each head computes an independent scaled dot-product attention with different learned projections (W_i^Q, W_i^K, W_i^V). Positional encodings are added to the inputs to provide information about order, compensating for the lack of recurrence.

Transformer Architecture. The Transformer consists of stacked encoder and decoder blocks. Each block applies multi-head attention, feedforward networks, residual connections, and layer normalization. The encoder maps an input sequence to contextualized embeddings, while the decoder autoregressively generates outputs, attending to both encoder representations and previously generated tokens.

This design removes sequential dependencies, enabling massive parallelization and scalability. Key improvements such as pre-layer normalization [59], efficient attention variants [60], and sparse attention [61] further extend applicability to longer sequences and larger models.

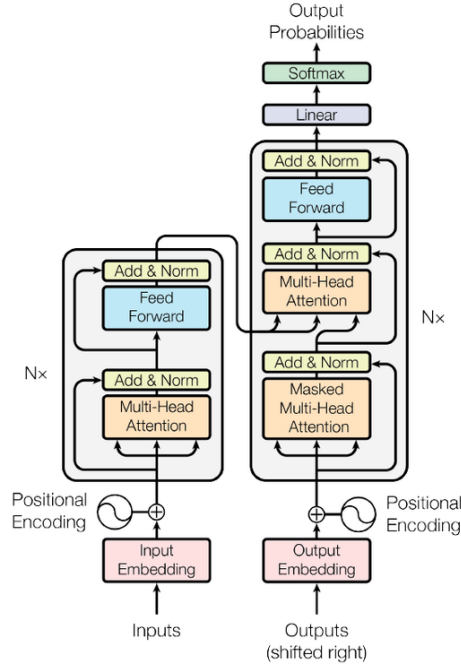


Figure 3.1: Transformer Architecture (Conceptual Diagram)

3.4 Single-cell Gene Expression Data Representation Learning

Single-cell RNA sequencing (scRNA-seq) has revolutionized our understanding of cellular heterogeneity by enabling the measurement of gene expression at the individual cell level. However, the high-dimensional, sparse, and noisy nature of scRNA-seq data poses significant challenges for downstream analyses such as cell type identification, trajectory inference, and perturbation prediction. Representation learning, particularly through deep learning models, has emerged as a powerful approach to extract meaningful low-dimensional embeddings from these data. These embeddings capture latent patterns in gene expression, facilitating improved performance in various tasks. Deep learning methods, including autoencoders, graph neural networks (GNNs), and transformer-based architectures, have been widely adopted for this purpose. This section reviews deep learning models

for single-cell gene expression representation learning, focusing on foundational and transformer-based models, and includes additional relevant approaches for comprehensiveness.

3.4.1 Early Deep Learning Models

Before the emergence of large transformer-based foundation models, variational autoencoders (VAEs) formed the backbone of single-cell representation learning. scVI (single-cell Variational Inference) [4] is one of the most influential VAE-based frameworks, modeling gene expression counts with a negative binomial distribution to account for over-dispersion and technical noise. By learning a probabilistic latent space, scVI supports key downstream tasks such as batch correction, differential expression, clustering, and imputation. Its hierarchical Bayesian structure enables principled uncertainty estimation, which is especially valuable for noisy single-cell data. scVI is highly scalable, capable of handling datasets with millions of cells through efficient variational inference and GPU parallelization, and integrates seamlessly with tools like Scanpy for end-to-end analysis pipelines.

scGNN [5] is a graph neural network framework that aggregates cell-cell relationships and models heterogeneous gene expression patterns for imputation and clustering. It constructs a dynamically pruned k-nearest-neighbor cell graph (refined iteratively via outlier detection with Isolation Forest), using multi-modal autoencoders: a feature autoencoder (regularized by left-truncated mixture Gaussian for regulatory signals), graph autoencoder (two-layer GCNs for topological embeddings), and cluster-specific autoencoders for type-aware reconstruction. Trained with combined MSE, cross-entropy, and regularization losses, it provides scalable imputation tolerant to noise and superior clustering on benchmarks like Alzheimer’s disease studies. scGNN 2.0 extends this with improved imputation capabilities via enhanced graph refinement. Variants like scBiGNN employ bilevel graph learning to mine gene and cell relationships simultaneously.

3.4.2 Transformer-Based Foundation Models

Transformer architectures, inspired by large language models, have been adapted to scRNA-seq data by treating genes as tokens in a sequence. These models are often pretrained on large corpora of single-cell transcriptomes using self-supervised tasks like masked language modeling (MLM), enabling transfer learning to downstream applications.

3.4.2.1 scBERT

scBERT [62], introduced as a pioneering BERT-based deep language model for single-cell transcriptomics, addresses key challenges in cell type annotation, such as the scarcity of curated marker gene lists, batch effects, and limited gene-gene interaction knowledge. Built on the bidirectional encoder representations from

transformers (BERT) architecture, scBERT treats genes as tokens in a sequence, employing a gene-specific tokenization and embedding strategy that captures intricate regulatory relationships among genes. This approach allows the model to process high-dimensional gene expression profiles as natural language-like inputs, where each cell’s transcriptome is represented as a sequence of gene tokens ordered by expression levels. Pretrained on a massive unlabeled corpus exceeding 1,000,000 cells sourced primarily from the PanglaoDB repository, along with diverse datasets like Zheng68k and pancreatic atlases, scBERT utilizes masked language modeling for self-supervised learning. This pretraining enables the model to learn contextual gene dependencies without task-specific supervision, with innovations like gene-specific embeddings enhancing representational fidelity and robustness to technical variations in scRNA-seq data.

3.4.2.2 Geneformer

Geneformer [1] represents a pioneering transformer-based foundation model for single-cell transcriptomics, pretrained on a vast corpus of approximately 30 million single-cell transcriptomes, which has since been expanded to over 100 million cells to enhance its representational power. This large-scale pretraining enables Geneformer to learn robust, generalizable embeddings that capture the intricate dynamics of gene expression across diverse biological contexts. A key innovation in Geneformer’s architecture is its rank value encoding strategy, which reorders genes within each cell based on their expression levels, transforming raw count data into a sequence of gene tokens. These tokens are ranked and scaled according to corpus-wide expression statistics, effectively prioritizing genes that distinguish specific cell states while downweighting ubiquitously expressed housekeeping genes. This nonparametric encoding approach significantly reduces the model’s sensitivity to technical artifacts inherent in absolute expression counts, such as batch effects or sequencing depth variations, leading to more reliable and biologically interpretable representations.

The pretraining process employs a masked token prediction task, akin to masked language modeling in natural language processing, to learn contextual relationships among genes. Advanced variants, such as Geneformer-V2-316M, feature 316 million parameters, a 4096-token input sequence length, and a vocabulary of approximately 20,000 genes, allowing the model to process extensive transcriptomic profiles efficiently. Through this pretraining, Geneformer generates 1152-dimensional cell embeddings that encapsulate gene network dynamics and cellular identities, providing a versatile foundation for downstream analyses. These embeddings have demonstrated exceptional performance in a variety of tasks, including accurate cell type classification—even in challenging cross-species scenarios, such as applying a human-trained model to classify mouse cells—highlighting the model’s strong generalization capabilities.

One of Geneformer’s standout strengths is its efficacy in handling limited data through transfer learning, making it an invaluable tool for studying rare

cell types or under-represented disease states where annotated data is scarce. By leveraging pre-learned knowledge from massive corpora, the model achieves high performance with minimal fine-tuning, accelerating discoveries in precision medicine and developmental biology. Overall, Geneformer’s innovative encoding and scalable architecture position it as a cornerstone in the evolving landscape of single-cell foundation models, empowering researchers to unlock deeper insights into cellular heterogeneity and function.

3.4.2.3 scGPT

scGPT [2] is a groundbreaking foundation model for single-cell multi-omics, leveraging generative pretraining on an extensive dataset of over 33 million cells from diverse sources. By adapting the transformer architecture, scGPT jointly learns cell and gene representations through sophisticated gene expression modeling via a binning-based tokenization strategy and self-supervised tasks, effectively distilling complex biological insights into unified embeddings. The model employs a self-attention mechanism to capture intricate dependencies between genes and cells, optimizing both representations simultaneously through generative objectives. This approach enables scGPT to excel in a wide range of applications, including zero-shot and fine-tuned scenarios for tasks such as batch correction, cell type annotation, perturbation response prediction, and multi-omic integration.

The generative capabilities of scGPT facilitate *in silico* simulations of cellular states, allowing researchers to model hypothetical biological scenarios with high fidelity. The model demonstrates exceptional performance in predicting gene regulatory networks and reverse perturbations, providing valuable insights into cellular dynamics. Furthermore, variants like scGPT-spatial enhance its applicability by incorporating spatially resolved transcriptomics data. These variants utilize a mixture-of-experts (MoE) decoder to model modality-specific distributions, enabling precise analysis of spatial relationships within tissues. This makes scGPT-spatial particularly effective for studying complex tissue architectures and spatially dependent cellular interactions.

3.4.2.4 CellPLM

CellPLM [63] is an advanced transformer-based model designed to extend single-cell transcriptomics by incorporating cell-cell relational information, redefining cells as tokens and tissues as sentences within a novel pretraining framework. This innovative approach leverages a cell-level masked language modeling task to capture intricate biological relationships, enabling robust analysis of spatially-resolved transcriptomic data. The model integrates several key components: a gene expression embedder that aggregates gene embeddings weighted by their expression levels; a transformer encoder utilizing Flowformer for linear computational complexity and 2D sinusoidal positional encodings to account for spatial relationships; a Gaussian mixture model (GMM) latent space that employs a mixture of Gaussians prior

for variational encoding, optimized via evidence lower bound (ELBO) with reconstruction, conditional prior, and cluster terms; and a batch-aware decoder with feed-forward layers conditioned on batch embeddings to mitigate batch effects. This sophisticated architecture enables CellPLM to model complex cell group interactions and capture tissue-level dynamics effectively.

By treating cells within their relational and spatial contexts, CellPLM outperforms traditional single-cell models in tasks requiring an understanding of cellular interactions, such as tissue-level analyses, zero-shot clustering, denoising, and perturbation prediction. The incorporation of the GMM prior allows the model to capture distinct functional cell groups, enhancing its ability to identify biologically meaningful clusters without extensive labeled data. This makes CellPLM particularly effective in zero-shot and few-shot learning scenarios, where it can generalize to new datasets and mitigate challenges posed by limited data availability.

3.4.2.5 LangCell

LangCell [64] is an innovative framework that integrates single-cell transcriptomics with natural language processing through a multimodal language-cell pre-training approach. This model leverages both single-cell RNA sequencing (scRNA-seq) data and enriched textual descriptions, which include detailed cell identity information, to create a unified representation space. By jointly pretraining on these modalities, LangCell achieves robust zero-shot cell identity understanding, enabling the classification of cell types based solely on textual queries without requiring fine-tuning. This capability stems from its ability to align transcriptomic profiles with linguistic descriptions, facilitating cross-modal knowledge transfer that provides deep biological insights.

The framework employs contrastive learning techniques to align transcriptomic and linguistic representations effectively. This alignment enhances multimodal fusion, allowing LangCell to capture intricate relationships between gene expression patterns and their corresponding textual annotations. In benchmark evaluations, LangCell demonstrates superior performance in tasks such as automated cell annotation and query-based cell retrieval, surpassing existing state-of-the-art models. Its architecture is particularly adept at handling complex biological queries, making it a powerful tool for interpretive, text-guided single-cell analysis.

Moreover, LangCell exhibits exceptional flexibility in various learning scenarios. It excels in few-shot learning, where minimal labeled data is available, and maintains strong performance when fine-tuned for specific tasks. These characteristics make LangCell uniquely suited for applications in single-cell biology, where precise and interpretable analysis is critical. By bridging the gap between computational biology and natural language processing, LangCell paves the way for advanced, language-driven exploration of cellular data, offering researchers a versatile tool for uncovering novel biological discoveries.

3.4.2.6 scCello

scCello [65] is a sophisticated transcriptome foundation model that integrates cell ontology to guide its learning process, embedding structural relationships between cell types to enhance its predictive and analytical capabilities. Pretrained on an extensive dataset of 22 million cells, scCello leverages ontology-mapped labels to incorporate hierarchical biological knowledge into its framework. The model is optimized using a composite loss function that combines three key components: masked gene expression prediction, which refines the model’s ability to reconstruct gene expression profiles; cell-type coherence loss, which enforces distinct, type-specific representations; and ontology alignment loss, which embeds hierarchical relationships from the cell ontology graph. This multifaceted approach enables scCello to achieve superior generalization in cell type classification and facilitates the discovery of novel cell subtypes.

By incorporating prior biological knowledge through its ontology-guided architecture, scCello addresses critical limitations of unsupervised models, which often struggle to capture structured relationships in complex biological data. The model’s ability to embed hierarchical cell type relationships ensures more accurate and biologically meaningful representations, making it highly effective for diverse downstream tasks. In benchmark evaluations, scCello demonstrates competitive transferability, excelling in tasks such as predicting marker genes and forecasting cancer drug responses. Its robust performance extends to both supervised and unsupervised scenarios, offering researchers a versatile tool for exploring cellular diversity and function.

3.5 Large Language Models

Large Language Models (LLMs) represent a paradigm shift in natural language processing (NLP), enabling machines to understand, generate, and manipulate human language with unprecedented accuracy and fluency. These models, typically based on transformer architectures, have evolved from early contextual embeddings and encoder-only models to massive autoregressive decoders and instruction-tuned hybrids. Their scaling in parameters and training data has unlocked remarkable performance across diverse tasks, from machine translation to reasoning and code generation. The progression from ELMo and BERT to the GPT series, T5, PaLM, BLOOM, and LLaMA exemplifies this trajectory, highlighting advancements in pretraining strategies, model scaling, and efficient adaptation techniques. This section surveys the historical development, key architectures, finetuning methods including Parameter-Efficient Fine-Tuning (PEFT) approaches such as LoRA, applications, challenges, and future directions of LLMs, providing a comprehensive overview for their integration in computational biology and beyond.

3.5.1 Evolution of LLM Architectures

The transformer architecture [6], introduced in 2017, replaced recurrence with self-attention, enabling parallelization and better modeling of long-range dependencies. Shortly before this, ELMo [66] pioneered deep contextualized word representations using bidirectional LSTMs, showing that dynamic embeddings could outperform static vectors like Word2Vec or GloVe. However, transformers quickly surpassed LSTMs in scalability and efficiency, laying the foundation for today’s LLMs.

BERT [7] (Bidirectional Encoder Representations from Transformers), released in 2018, marked a turning point by introducing masked language modeling (MLM) and next-sentence prediction for bidirectional pretraining. Trained on BooksCorpus and English Wikipedia, BERT-base (110M parameters) and BERT-large (340M parameters) achieved state-of-the-art results on GLUE [67] and SQuAD [68], establishing bidirectional context as critical for NLP. Its success inspired derivatives like RoBERTa (optimized training) and XLNet [69], which combined autoregression with permutation-based objectives to overcome MLM’s limitations.

In parallel, generative approaches flourished. OpenAI’s GPT-1 (2018) demonstrated transfer learning from generative pretraining, while GPT-2 (2019) scaled to 1.5B parameters and showed emergent zero-shot capabilities on summarization and translation without explicit finetuning. The leap to GPT-3 [8], with 175B parameters, introduced few-shot in-context learning, rivaling task-specific models. More recently, GPT-4 (2023) added multimodal capabilities, handling both text and images with human-level reasoning and stronger alignment mechanisms.

Meanwhile, encoder-decoder models like T5 [70] reframed NLP tasks under a unified “text-to-text” paradigm, enabling translation, summarization, and classification with the same model. At larger scales, PaLM (540B parameters) showcased improved reasoning, multilingual ability, and chain-of-thought prompting. Open science efforts also emerged: BLOOM (176B) was trained collaboratively with multilingual coverage.

3.5.2 LLaMA

The LLaMA family of models from Meta AI exemplifies the push towards efficient, open-source large language models that prioritize accessibility and performance at various scales. Originating with LLaMA and LLaMA 2 (7B–70B parameters), the series evolved significantly with LLaMA 3, introducing native support for multilinguality, coding, reasoning, and tool usage through a dense Transformer architecture up to 405B parameters and 128K token context windows [71]. LLaMA 3.2, released in 2024, further advances this lineage with lightweight text-only models at 1B and 3B parameters, alongside vision-enabled variants at 11B and 90B parameters. These models feature an enhanced Transformer architecture with expanded vocabulary and embedding dimensions to boost performance in compact sizes. The text models are optimized for on-device use cases such as summa-

rization and tool calling, while the vision models employ a late-fusion architecture. This integrates image tokens from a pre-trained vision encoder (e.g., CLIP-like) via cross-attention layers in adapter modules, enabling advanced multimodal tasks like image reasoning and document understanding without compromising text-only capabilities.

Pretraining for LLaMA 3.2 emphasizes efficiency and alignment, building directly on LLaMA 3.1 foundations. The lightweight text models undergo knowledge distillation from larger LLaMA 3.1 variants (8B and 70B), using their logits as token-level targets during pretraining on vast multilingual corpora, which is then extended to a 128K context length. Post-training incorporates supervised fine-tuning (SFT), rejection sampling, and direct preference optimization (DPO), augmented by synthetic data generated with LLaMA 3.1 for tasks like instruction following, alongside safety mitigations. For the vision models, the process begins with frozen LLaMA 3.1 text backbones, adding trainable image encoders and adapters, followed by pretraining on large-scale noisy image-text pairs and refined with high-quality, domain-specific data. Alignment stages mirror the text models, including SFT and DPO with reward modeling for preference ranking, ensuring robust performance across vision-language tasks such as captioning and visual grounding. This methodology not only yields models competitive with larger counterparts but also facilitates fine-tuning and deployment in resource-constrained environments, making LLaMA 3.2 particularly suitable for framework integrations in computational biology.

3.5.3 Pretraining Paradigms

Pretraining is the cornerstone of LLMs, typically using self-supervised learning on massive unlabeled corpora to acquire broad language competence. Different paradigms emerged: BERT’s MLM masks 15% of tokens for bidirectional prediction; GPT’s autoregression predicts the next token to enable fluent generation; and XLNet’s permutation-based modeling blends the two. T5 further unified tasks into a sequence-to-sequence format, while recent instruction-tuned models are pretrained or finetuned on natural language instructions paired with responses, improving zero-shot usability.

Scaling laws, as demonstrated by GPT-3 and PaLM, reveal predictable performance gains with larger model size, data, and compute, motivating billion- and trillion-parameter systems. Yet challenges persist, including dataset bias, environmental cost of training, and reliance on specialized hardware like GPU/TPU clusters. The emergence of community-driven models such as BLOOM and LLaMA has partially democratized access, enabling broader research participation and responsible exploration of alignment, efficiency, and domain-specific adaptation.

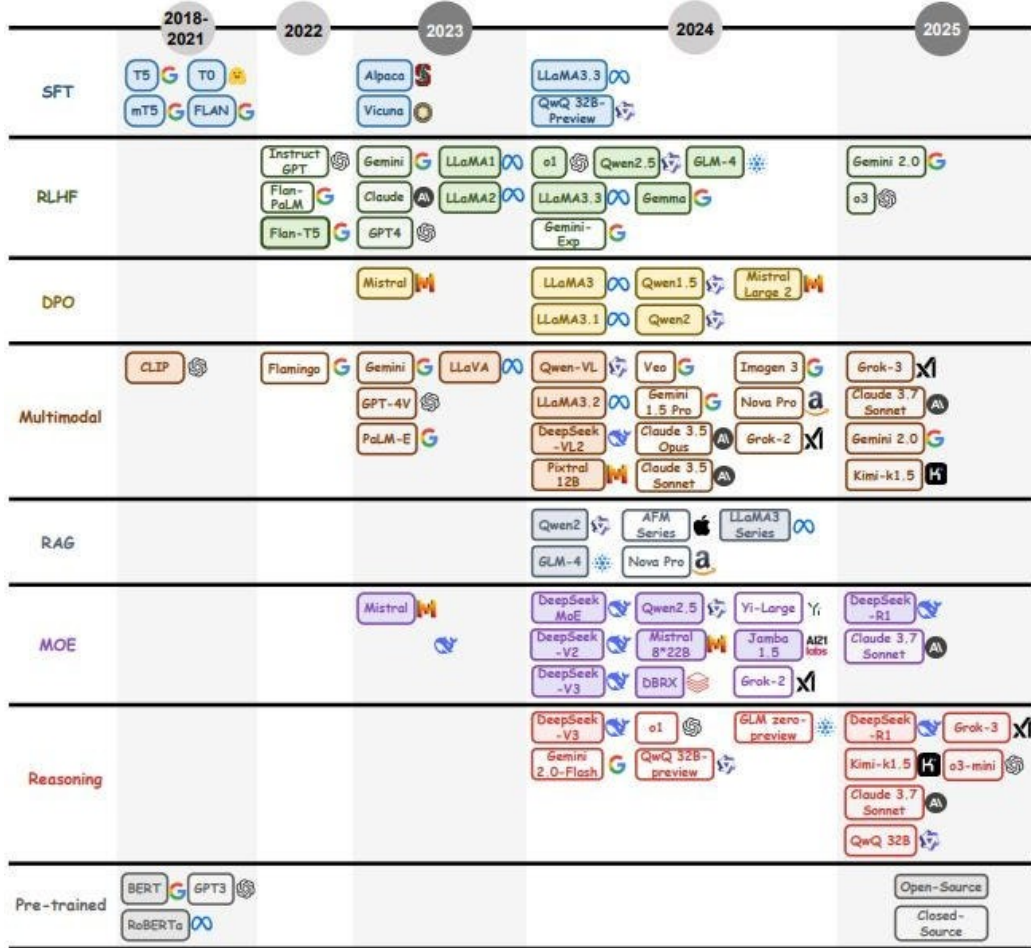


Figure 3.2: Timeline of Large Language Model Evolution.

3.5.4 Finetuning Techniques

Finetuning adapts pretrained LLMs to specific tasks by updating weights on labeled data, bridging the gap between general pretraining and domain-specific applications. Traditional full finetuning updates all parameters, effective but resource-intensive, and can risk catastrophic forgetting where the model loses general knowledge.

To address efficiency, Parameter-Efficient Fine-Tuning (PEFT) methods update only a fraction of parameters while reusing the frozen pretrained backbone. Techniques include adapter modules, which insert small trainable layers between transformer blocks; prompt tuning, where continuous soft embeddings replace task-specific prompts; and prefix tuning, which optimizes virtual tokens by modifying

the key/value states in attention layers. These approaches reduce memory footprint and training time while maintaining strong performance.

LoRA (Low-Rank Adaptation) [9] has become a prominent PEFT method due to its simplicity and scalability. Instead of updating the full weight matrix $W \in \mathbb{R}^{d \times k}$ of a linear layer (e.g., projection matrices in self-attention), LoRA freezes W and learns a low-rank decomposition ΔW :

$$W' = W + \Delta W, \quad \Delta W = AB,$$

where $A \in \mathbb{R}^{d \times r}$ and $B \in \mathbb{R}^{r \times k}$ with rank $r \ll \min(d, k)$. During training, only A and B are updated, while W remains fixed. This reduces the number of trainable parameters from $\mathcal{O}(dk)$ to $\mathcal{O}(r(d + k))$, often by several orders of magnitude. For example, with $d = k = 4096$ and $r = 8$, LoRA requires only 65k trainable parameters compared to 16M for the full matrix.

The forward computation for an input $x \in \mathbb{R}^k$ becomes:

$$h = Wx + \alpha A(Bx),$$

where α is a scaling factor (typically $\alpha = r$) ensuring that ΔW is initialized near zero, stabilizing training. Importantly, LoRA introduces no additional inference latency because AB can be merged into W after training.

Variants such as QLoRA further optimize efficiency by quantizing the frozen backbone weights to 4-bit precision and keeping LoRA adapters in higher precision, enabling finetuning of large models (e.g., 65B parameters) on a single consumer-grade GPU.

Beyond LoRA, alignment-focused strategies such as instruction finetuning (e.g., FLAN, Alpaca) use diverse instruction-response pairs to enhance generalization, while reinforcement learning from human feedback (RLHF) fine-tunes models to better reflect human preferences, as demonstrated in InstructGPT.

3.5.5 Multimodal Large Language Models

Multimodal models extend LLMs by integrating multiple data modalities, such as text, images, audio, and video, to enable holistic understanding and generation across sensory inputs. This fusion mimics human cognition, where perception combines visual, linguistic, and other cues. In computational biology, multimodal models are particularly promising for analyzing heterogeneous data like genomic sequences, microscopic images, and clinical notes. This subsection explores key architectures including LLaVA and BLIP-2’s Q-Former, other notable models, applications in computational biology, challenges, and future prospects, emphasizing their role in advancing interdisciplinary research.

Multimodal learning involves aligning representations from different modalities in a shared embedding space, often using contrastive losses or fusion mechanisms. Early models like CLIP [72] pretrained vision and language encoders contrastively

on image-text pairs, enabling zero-shot classification. Building on this, modern multimodal LLMs leverage frozen pretrained components for efficiency.

BLIP-2 (Bootstrapping Language-Image Pre-training) [73], introduced in 2023, employs a Q-Former (Querying Transformer) to bridge frozen image encoders (e.g., ViT) and LLMs (e.g., OPT or Flan-T5). The Q-Former, a lightweight transformer, uses learnable query embeddings to extract relevant visual features, trained in two stages: vision-language representation learning and generative learning. This approach achieves state-of-the-art on vision-language tasks like VQA and captioning with reduced parameters.

LLaVA (Large Language and Vision Assistant), also from 2023, combines a vision encoder (CLIP-ViT) with an LLM (Vicuna) via a projection layer, tuned on generated visual instruction data. Using GPT-4 to create instruction-following datasets from image captions, LLaVA excels in multimodal chat and reasoning, approaching GPT-4V performance. Extensions like LLaVA-Med adapt it for biomedical images.

Core components include modality-specific encoders (e.g., ViT for vision, BERT for text) and fusion modules like cross-attention or adapters. Training often involves bootstrapping: pre-aligning modalities then fine-tuning end-to-end. Instruction tuning with multimodal prompts enhances task adaptability.

In computational biology, multimodal LLMs enable integrated analysis of genomic data with imaging (e.g., predicting cell types from microscopy and transcriptomics) or literature mining with structural data (e.g., protein design). Challenges include data scarcity for specialized domains, alignment across noisy modalities, and computational demands. Future directions involve scaling to more modalities (e.g., genomics-specific encoders) and improving efficiency for edge deployment, fostering breakthroughs in personalized medicine and drug discovery.

3.5.6 Applications and Challenges

LLMs have transformed applications in NLP, including machine translation, sentiment analysis, question answering, and code generation. In computational biology, they aid in protein sequence prediction, drug discovery, and literature mining, with multimodal extensions enhancing analysis of integrated datasets.

However, challenges persist: hallucinations, where models generate plausible but incorrect information; ethical concerns like bias amplification; and environmental impact from high energy consumption. Future directions include sparse models, federated learning, and integration with knowledge graphs for enhanced reliability.

In summary, LLMs from BERT to GPT have redefined AI capabilities, with finetuning techniques like LoRA enabling efficient customization, paving the way for specialized applications in various domains.

3.6 Inspiration for Cell2Text and Similar Frameworks

The development of Cell2Text draws significant inspiration from recent advancements in multimodal machine learning frameworks that bridge high-dimensional biological data with natural language processing. These models address the challenge of translating complex genomic or proteomic information into interpretable, human-readable formats, thereby enhancing accessibility and accelerating insights in biomedical research. Pioneering works in this domain, such as Prot2Text for protein function prediction, have demonstrated the efficacy of integrating specialized encoders with large language models (LLMs) to generate semantically rich descriptions. Building on this foundation, subsequent single-cell-specific frameworks like CellWhisperer, Cell2Sentence, InstructCell, and scMMGPT have further refined multimodal alignments, incorporating contrastive learning, instruction tuning, and large-scale pre-training to handle transcriptomic data. By adapting these strategies, Cell2Text extends the paradigm to single-cell transcriptomics, leveraging gene embeddings and metadata-enriched datasets to produce comprehensive cell descriptions that integrate cellular identity, functional states, and contextual annotations.

3.6.1 Prot2Text

Prot2Text [74] served as a primary inspiration for Cell2Text by demonstrating the feasibility of multimodal models that generate natural language descriptions from biological sequences. In Prot2Text, the architecture integrates graph neural networks (GNNs) and transformers to process protein structures and sequences, enabling the generation of functional descriptions. The method begins with encoding protein graphs using a relational graph convolutional network (RGCN) to capture structural relationships, representing proteins as heterogeneous graphs derived from AlphaFold-predicted 3D structures with sequential, spatial, and hydrogen bond edges. Protein sequences are encoded into token embeddings using ESM2-35M. These token embeddings are fused with the mean-pooled graph representation through element-wise addition and projection layers, repeated across multiple blocks to enrich each amino acid token with structural information. The fused protein representations are then integrated into a GPT-2-based transformer decoder via multi-head cross-attention, facilitating the generation of coherent textual descriptions.

Building on this, Prot2Text v2 advances the framework with an encoder-decoder architecture and a two-stage training process for enhanced protein function prediction. The architecture features ESM-3B as the protein encoder, projecting amino acid sequences into embeddings via a nonlinear modality projector, which maps them into the semantic space of LLaMA-3.1-8B-Instruct. This integration avoids complex cross-attention, simplifying training while preserving semantic richness through structured templates that include optional metadata like protein names and taxonomy. Dataset creation refines SwissProt data into 256,690 non-redundant

proteins, with CD-HIT clustering ensuring low sequence identity between splits. Training involves contrastive alignment in the first stage and instruction-based fine-tuning in the second, using dropout on metadata to improve generalization. These methods inspired Cell2Text’s adapter module and LLM integration, adapting similar projection and fine-tuning strategies to bridge gene embeddings with textual cell descriptions.

3.6.2 CellWhisperer

CellWhisperer is a multimodal model enabling chat-based exploration of single-cell RNA-seq (scRNA-seq) data by bridging transcriptomics and natural language. It uses Geneformer to encode transcriptomes into 2048-dimensional embeddings from ranked gene sequences, with text encoded via BioBERT and both modalities aligned in a shared space using CLIP-like contrastive adapters (two-layer linear modules with ReLU and batch normalization). For generation, a two-layer adapter projects transcriptome embeddings into eight 4096-dimensional tokens for the Mistral 7B LLM, allowing interleaved cell-text processing with special tokens for cross-modal attention.

The dataset comprises over 1 million transcriptome-annotation pairs from bulk and pseudo-bulk scRNA-seq sources, with LLM-generated annotations using chain-of-thought prompts. Training has two stages: discriminative pre-training with InfoNCE loss, freezing Geneformer; and generative fine-tuning on synthetic Q&A pairs. Integrated into CELLxGENE Explorer, it supports zero-shot annotation, free-text search (e.g., intestinal T cells”), and interactive queries (e.g., Highly expressed genes?”), with UMAP visualizations. This chat-focused, adapter-driven design inspired Cell2Text’s prompts and projections for conversational single-cell analysis.

3.6.3 Cell2Sentence

Cell2Sentence v1 introduces a transformative approach by converting single-cell gene expression into "cell sentences" for LLM adaptation, inspiring Cell2Text’s gene-level embedding strategy. The method preprocesses scRNA-seq matrices by filtering low-quality cells and genes, normalizing counts, and applying a rank-order transformation to create sequences of gene names ordered by abundance. This preserves biological nuances through log-linear modeling of expression-rank relationships, allowing reversible conversion back to expression space.

Training involves fine-tuning LLMs on these cell sentences for tasks like summarization and perturbation prediction, with robustness evaluated via correlation metrics across datasets. Version 2 scales this by training on a billion-token corpus of cell sentences, using larger LLMs like Gemma for improved performance in single-cell tasks. These methods influenced Cell2Text’s dataset augmentation with pathways and ontology, fostering similar cross-modal learning for descriptive outputs.

3.6.4 InstructCell

InstructCell is a multi-modal AI copilot that enables intuitive natural language interactions for single-cell RNA sequencing (scRNA-seq) analysis, directly inspiring Cell2Text’s instruction-tuned LLM decoder for generating coherent cell descriptions. The architecture features a Querying Transformer (Q-Former) to encode scRNA-seq data into learnable query vectors via cross-attention, which are integrated with a pre-trained LLM backbone and a reconstruction module for decoding back to gene expression space. This setup supports tasks such as cell type annotation, pseudo-cell generation, and drug response prediction, ensuring quantitative fidelity alongside textual outputs.

The dataset is built from 299,155 scRNA-seq samples from public sources like the Human Cell Atlas, with GPT-4o synthesizing diverse instruction-response pairs incorporating user personas for varied dialogues. Training uses supervised fine-tuning on these pairs, emphasizing cross-modal alignment through contrastive losses to synchronize transcriptomic and textual embeddings. This conversational focus enhances accessibility and interpretability, paralleling Cell2Text’s structured prompts for professional genomic insights.

3.6.5 scMMGPT

scMMGPT fuses single-cell transcriptomics with textual knowledge to advance multimodal representation learning, influencing Cell2Text’s cross-modal alignment and description generation. It employs scGPT for deriving cell embeddings from normalized top-2,048 expressed genes and LLaMA-2 for text processing, connected by bidirectional Q-Former projectors (32 queries) that enable cell-to-text and text-to-cell flows while preserving expression fidelity through tokenization.

Curated from 27 million cells in CellxGene across tissues like brain and lung, the dataset pairs transcriptomes with ontology-based (OBO) and narrative (Wikipedia) descriptions to ground data in interpretable language. Training features two stages: discriminative pre-training with InfoNCE for alignment, followed by generative pre-training emphasizing bidirectional tasks—cell-to-text autoregressive summary generation via causal language modeling to capture functional states, and text-to-cell reconstruction with MSE loss for simulating perturbations.

Chapter 4

Multimodal Dataset Construction and Sampling Methodology

4.1 Overview

This section describes the systematic approach employed for constructing a large-scale multimodal dataset consisting of 1M cells, combining single-cell RNA sequencing (scRNA-seq) gene expression data with associated textual metadata. The dataset was carefully sampled from the CELLxGENE Census database to ensure comprehensive representation across multiple biological dimensions while maintaining statistical rigor for downstream machine learning applications.

4.2 Data Source and Initial Filtering

The foundation of our multimodal dataset originates from the CELLxGENE Census [10], a standardized corpus of single-cell transcriptomic data. We extracted cellular metadata including developmental stage, disease status, assay type, tissue origin, cell type annotations, and donor information for *Homo sapiens* samples.

4.2.1 Quality Control and Exclusion Criteria

To ensure data quality and consistency, we implemented a multi-stage filtering pipeline:

1. **Metadata Completeness:** Cells were required to have complete annotations for tissue type, disease status, and cell type classifications.
2. **Primary Data Selection:** Only primary experimental data was retained, removing duplicate entries.

3. **Assay Technology Filtering:** We excluded rare or highly specialized sequencing protocols to maintain technical consistency:

- Full-length protocols: Smart-seq variants
- Low-prevalence methods: Quartz-seq, MARS-seq, SORT-seq
- Proprietary technologies: GEXSCOPE, BD Rhapsody
- Targeted approaches: 10x gene expression flex

After initial filtering, the dataset comprised `N_filtered` cells meeting our quality criteria.

4.3 Stratified Sampling Strategy

To construct a representative subset while maintaining biological diversity, we implemented a multi-component stratified sampling strategy. The target dataset size was set to `target_total` cells, distributed across five complementary sampling approaches:

Table 4.1: Sampling strategy allocation and objectives

Sampling Component	Allocation (%)	Primary Objective
Distribution-based	30%	Overall population representation
Cell type diversity	25%	Comprehensive cell type coverage
Disease representation	20%	Pathological condition diversity
Donor diversity	15%	Inter-individual variation
Rare categories	10%	Low-frequency tissue/stage preservation

4.3.1 Distribution-based Sampling with Tissue Adjustment

The distribution-based component aims to reflect the overall population structure while correcting for known overrepresentation biases in the source database. We identified that brain and blood tissues were disproportionately represented and implemented weighted sampling to achieve more realistic biological proportions:

$$w_i = \begin{cases} \frac{p_{\text{target}}^{(t)}}{p_{\text{observed}}^{(t)}} & \text{if tissue}(i) = t \text{ and } p_{\text{observed}}^{(t)} > p_{\text{target}}^{(t)} \\ 1 & \text{otherwise} \end{cases} \quad (4.1)$$

where $p_{\text{target}}^{(t)}$ and $p_{\text{observed}}^{(t)}$ represent target and observed proportions for tissue type t , respectively.

4.3.2 Cell Type Diversity Sampling

To ensure comprehensive representation of cellular diversity, we implemented a capped sampling strategy across cell types:

$$n_{\text{sample}}^{(c)} = \min \left(n_{\text{available}}^{(c)}, \lfloor 0.03 \times n_{\text{cell_type}} \rfloor \right) \quad (4.2)$$

where $n_{\text{sample}}^{(c)}$ represents the number of cells sampled for cell type c , capped at 3% of the total cell type allocation to prevent dominance by highly abundant cell types.

4.3.3 Disease Representation Strategy

Disease sampling employed differential capping to balance normal and pathological conditions while preventing overrepresentation of specific diseases:

$$n_{\text{disease}}^{(d)} = \begin{cases} \min(n_{\text{available}}^{(d)}, 0.05 \times n_{\text{disease}}) & \text{if } d = \text{COVID-19} \\ \min(n_{\text{available}}^{(d)}, 0.70 \times n_{\text{disease}}) & \text{if } d = \text{normal} \\ \min(n_{\text{available}}^{(d)}, 0.03 \times n_{\text{disease}}) & \text{otherwise} \end{cases} \quad (4.3)$$

4.3.4 Donor Diversity and Rare Category Sampling

Donor diversity sampling ensured broad representation across individuals by capping contributions per donor at 2% of the donor allocation. Rare category sampling targeted tissues and developmental stages in the bottom 10th percentile of frequency distribution to preserve representation of biologically important but numerically rare categories.

4.4 Dataset Splitting Strategy

We implemented a robust train-validation-test splitting procedure designed to prevent data leakage while ensuring comprehensive category representation in the training set. The splitting follows a 80/10/10 ratio with donor-level partitioning to maintain biological independence.

4.5 Diversity Metrics and Quality Assessment

To quantify the representational quality of our sampling strategy, we employed the normalized Shannon diversity index across multiple biological dimensions:

Algorithm 1 Robust Dataset Splitting with Category Preservation

```
1: Input: Dataset  $D$ , stratification columns  $S$ , target ratios  $(r_{\text{train}}, r_{\text{val}}, r_{\text{test}})$ 
2: Initialize  $D_{\text{train}} \leftarrow \emptyset$ ,  $D_{\text{val}} \leftarrow \emptyset$ ,  $D_{\text{test}} \leftarrow \emptyset$ 
3: for each column  $s \in S$  do
4:   for each category  $c$  in column  $s$  do
5:     if category  $c$  not represented in training donors then
6:       Select minimal donor  $d$  covering category  $c$ 
7:       Add  $d$  to training donor set
8:     end if
9:   end for
10: end for
11: Greedily assign remaining donors to validation and test sets based on
    target sizes
12: Assign remaining donors to training set
13: return  $(D_{\text{train}}, D_{\text{val}}, D_{\text{test}})$ 
```

$$H'_{\text{norm}}(X) = \frac{-\sum_{i=1}^k p_i \log_2(p_i)}{\log_2(k)} \quad (4.4)$$

where p_i represents the proportion of category i in variable X , and k is the total number of categories. This metric provides a standardized measure of diversity ranging from 0 (complete homogeneity) to 1 (perfect evenness).

Variable	Before Sampling	After Sampling
Cell Type	0.7470	0.8431
Tissue (general)	0.6106	0.7035
Disease	0.3479	0.4957

Table 4.2: Normalized Shannon diversity before and after applying the sampling strategy.

4.6 Multimodal Text Description Generation

The second component of our multimodal dataset consists of structured natural language descriptions generated from cellular metadata and functional annotations. This text generation pipeline transforms categorical biological metadata into coherent, contextually rich descriptions suitable for cross-modal learning.

4.6.1 Ontology Integration and Pathway Enrichment

We enhanced cellular annotations through two complementary approaches:

Biological Ontology Integration Cell type, tissue, disease, and other categorical annotations were enriched using standardized biological ontologies including the Cell Ontology (CL), Human Phenotype Ontology (HPO), and Experimental Factor Ontology (EFO). For each ontological term, we extracted:

- Canonical term names
- Formal definitions when available
- Synonyms for comprehensive coverage

Functional Pathway Analysis We computed pathway activity scores using AUCell [11], a method for assessing gene signature enrichment at the single-cell level. The analysis pipeline consisted of:

1. **Global Preprocessing:** Normalization and highly variable gene (HVG) selection across the entire dataset to ensure consistent feature space
2. **Gene Signature Loading:** Integration of Hallmark pathways from MSigDB v2025.1 containing curated biological process signatures
3. **Batch-wise AUCell Computation:** Processing cells in chunks to manage memory constraints while calculating enrichment scores:

$$\text{AUC}_{i,s} = \frac{1}{|G_s|} \sum_{g \in G_s} \frac{r_{i,g}}{N} \quad (4.5)$$

where $\text{AUC}_{i,s}$ represents the area under the curve for cell i and signature s , G_s is the set of genes in signature s , $r_{i,g}$ is the rank of gene g in cell i , and N is the total number of genes.

Pathway Selection and Assignment From the computed AUCell scores, we:

- Filtered pathways active in $>0.5\%$ of cells to focus on biologically relevant processes
- Identified the top two most enriched pathways per cell for inclusion in text descriptions
- Mapped pathway names to human-readable descriptions via web scraping of pathway URLs

4.6.2 Natural Language Description Architecture

Our text generation system employs a rule-based approach with sophisticated linguistic templates that adapt to biological context:

Developmental Stage Bucketing Age-related metadata was standardized into biologically meaningful categories:

$$\text{AgeBucket}(s) = \begin{cases} \text{early embryonic stage} & \text{if } s \in \{9-16 \text{ weeks post-fertilization}\} \\ \text{late embryonic stage} & \text{if } s \in \{17-31 \text{ weeks post-fertilization}\} \\ \text{infancy} & \text{if } s \in [0, 2) \text{ years} \\ \text{early childhood} & \text{if } s \in [2, 5) \text{ years} \\ \text{childhood} & \text{if } s \in [5, 15) \text{ years} \\ \text{adolescence} & \text{if } s \in [15, 20) \text{ years} \\ \text{young adulthood} & \text{if } s \in [20, 30) \text{ years} \\ \text{adulthood} & \text{if } s \in [30, 50) \text{ years} \\ \text{middle age} & \text{if } s \in [50, 65) \text{ years} \\ \text{elderly stage} & \text{if } s \in [65, 80) \text{ years} \\ \text{advanced elderly stage} & \text{if } s \geq 80 \text{ years} \end{cases} \quad (4.6)$$

Template-based Description Generation Each cell description follows a structured template:

"This sample consists of a [CELL_TYPE], [CELL_DEFINITION]. It originates from the [TISSUE] of a [DISEASE_STATUS] [SEX] during [DEVELOPMENTAL_STAGE]. This cell is associated with [PATHWAY_1]. Additionally, it involves [PATHWAY_2]."

Example Description

This sample consists of a ciliated columnar cell of tracheobronchial tree, multi-ciliated epithelial cell located in the trachea and bronchi, characterized by a columnar shape and motile cilia on its apical surface. These cilia facilitate mucociliary clearance by moving mucus and trapped particles toward the pharynx. It originates from the lung parenchyma of a normal male during elderly stage. This cell is associated with Genes mediating programmed cell death (apoptosis) by activation of caspases. Additionally, it involves Genes down-regulated in response to ultraviolet (UV) radiation.

The system intelligently handles missing information, combines redundant tissue annotations, and adapts language based on disease status and demographic characteristics.

4.6.3 Text Description Quality Metrics

Generated descriptions exhibit several desirable properties:

- **Completeness:** Integration of all available metadata dimensions
- **Biological Accuracy:** Incorporation of formal ontological definitions
- **Functional Context:** Inclusion of pathway activity information
- **Linguistic Coherence:** Natural language flow with appropriate biological terminology
- **Consistency:** Standardized format across all samples while preserving biological specificity

4.7 Validation and Quality Assurance

Our sampling methodology underwent rigorous validation to ensure:

1. **No Data Leakage:** Complete donor-level separation between training, validation, and test sets
2. **Category Completeness:** All biological categories present in validation and test sets are represented in the training set
3. **Diversity Preservation:** Maintenance of biological diversity across all dataset partitions
4. **Statistical Balance:** Appropriate representation of rare and common biological conditions

The resulting multimodal dataset provides a robust foundation for training and evaluating machine learning models that integrate gene expression and textual information while maintaining the biological complexity inherent in single-cell transcriptomic data.

4.8 Final Dataset Characteristics

The dataset exhibits distinct statistical and distributional patterns across its various dimensions. Gene expression sequences tokenized with the Geneformer tokenizer (Figure 4.1) have an average length of 1,843.2 tokens, with the distribution peaking between 1,000–1,500 tokens and a noticeable spike at the 4,096-token maximum, reflecting truncation or padding effects. In contrast, natural language descriptions tokenized with the Llama-3.2-1B-Instruct tokenizer (Figure 4.2) are much shorter, averaging 104.3 tokens, with most clustering around 100–150 tokens and steadily

fewer exceeding 200 tokens. Examining biological categories reveals strong skews: among 783 annotated cell types (Figure 4.3), glial cells such as oligodendrocytes dominate due to brain-derived data, while immune cells like T and B cells are also prominent, whereas stromal and plasma cells remain underrepresented, highlighting biases introduced by sampling accessibility and methodological challenges. A similar pattern is observed in disease categories (Figure 4.4), where normal (592,932) and COVID-19 (61,961) samples dominate among 128 categories, reflecting the central role of healthy controls in baseline studies and the surge of data collection during the pandemic. Tissue-level distributions (Figure 4.5) reinforce these imbalances: brain (211,063) and blood (177,460) top the 347 annotated categories, followed by lung (87,784) and eye (59,430), consistent with the focus on clinically relevant and accessible tissues. Assay representation (Figure 4.7) is similarly skewed, with droplet-based single-cell RNA-seq protocols 10x v3 (480,361) and 10x v2 (242,493) far outnumbering others such as scRNA-seq3 (86,205) and 10x transcription profiling (53,623), underscoring the dominance of scalable standardized methods. Finally, pathway distributions (Figure 4.6) are shaped by these sampling and tissue biases: hallmark TNF- α via NF κ B (302,664) and hallmark UV response DN (205,134) lead, alongside immune-related pathways such as interferon- α response (167,957) and allograft rejection (151,116), while tissue-specific pathways like spermatogenesis (3,109) remain rare. Together, these statistics highlight that the dataset’s composition reflects not only biological relevance but also methodological accessibility, technical preferences, and historical research priorities.

Dataset Split	Number of Cells
Training Set	799,711 cells
Validation Set	100,083 cells
Test Set	100,090 cells

Table 4.3: Number of Cell of each Dataset Split

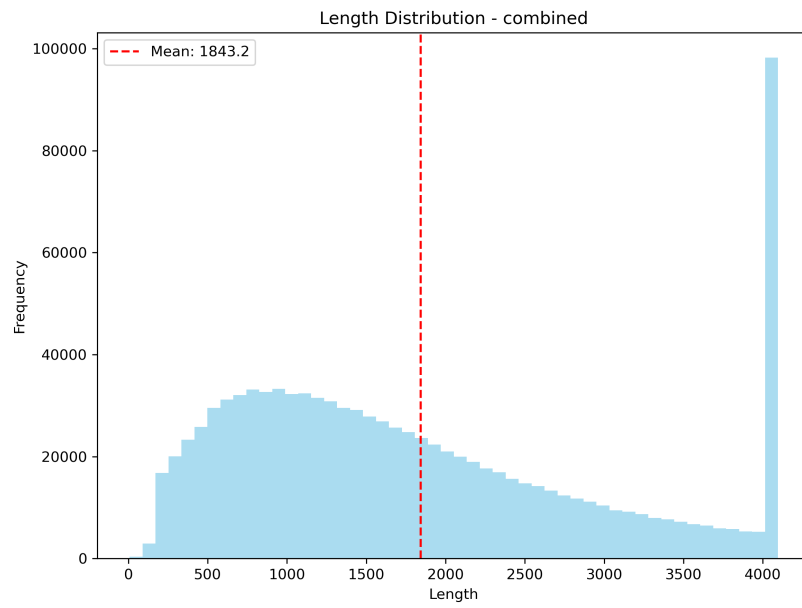


Figure 4.1: Distribution of tokens in gene expression sequences

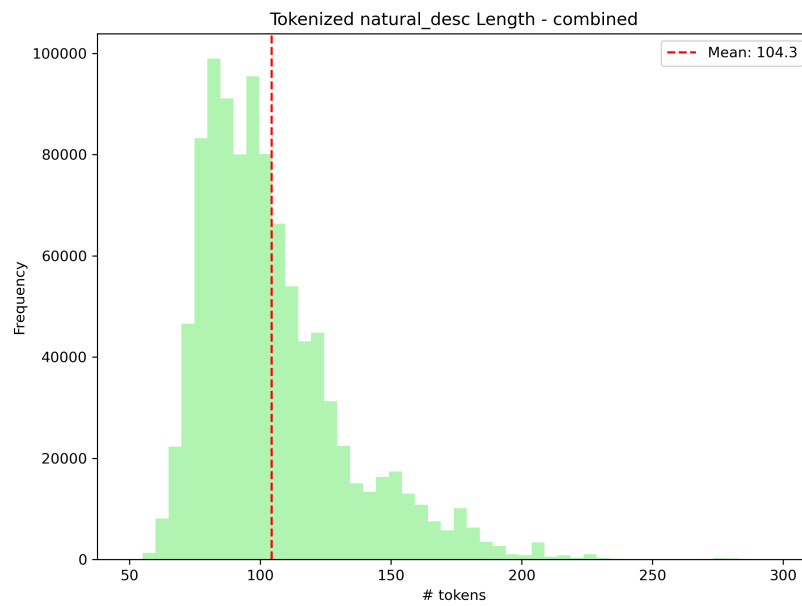


Figure 4.2: Distribution of tokens in natural language descriptions

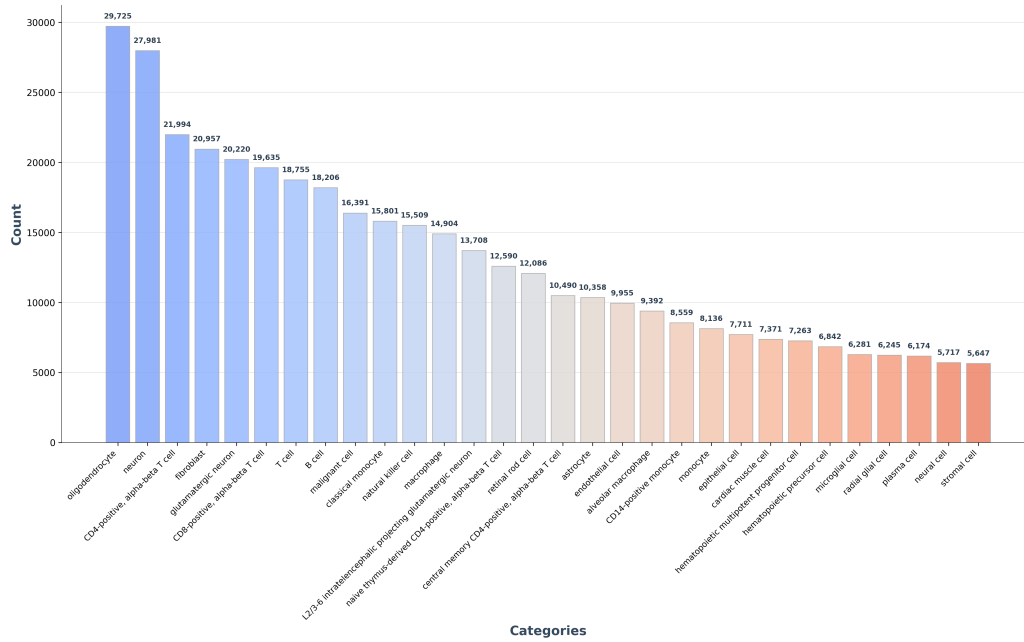


Figure 4.3: Overview of the distribution of cell types in the dataset. For clarity, only the 30 most abundant categories are shown.

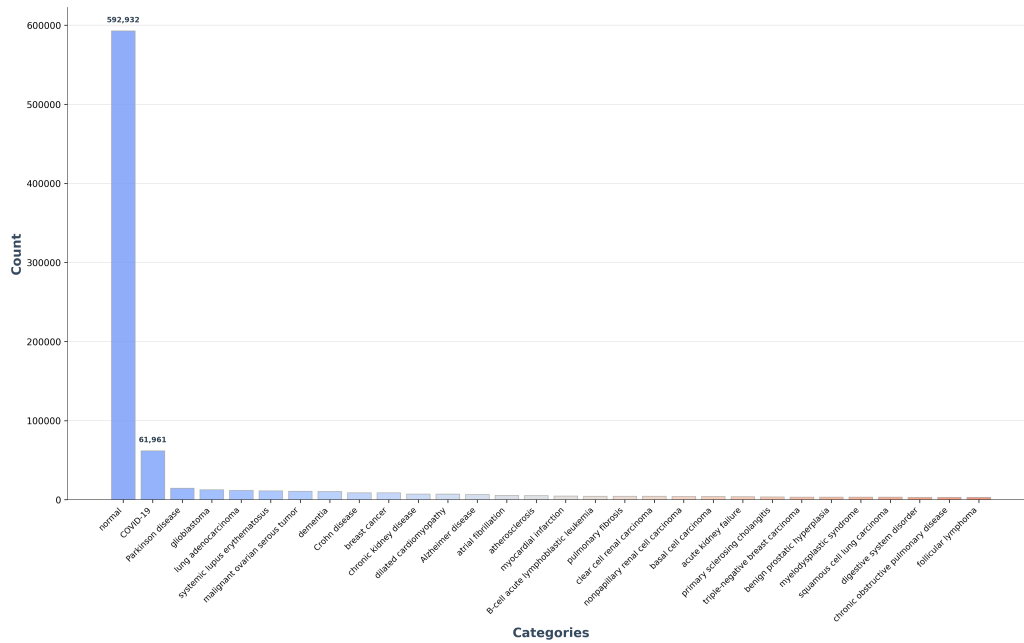
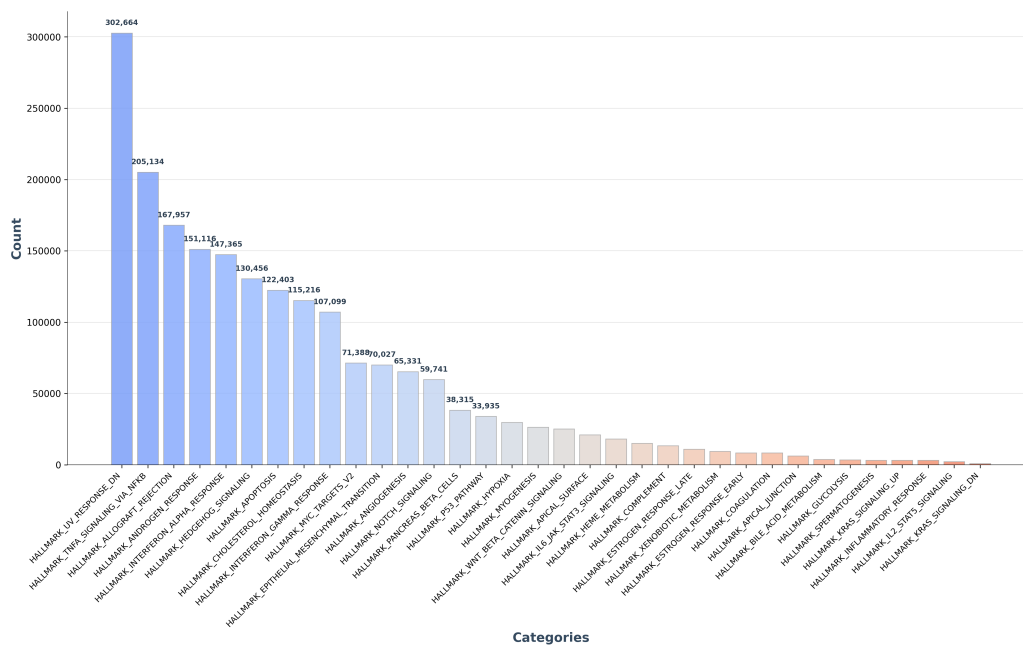
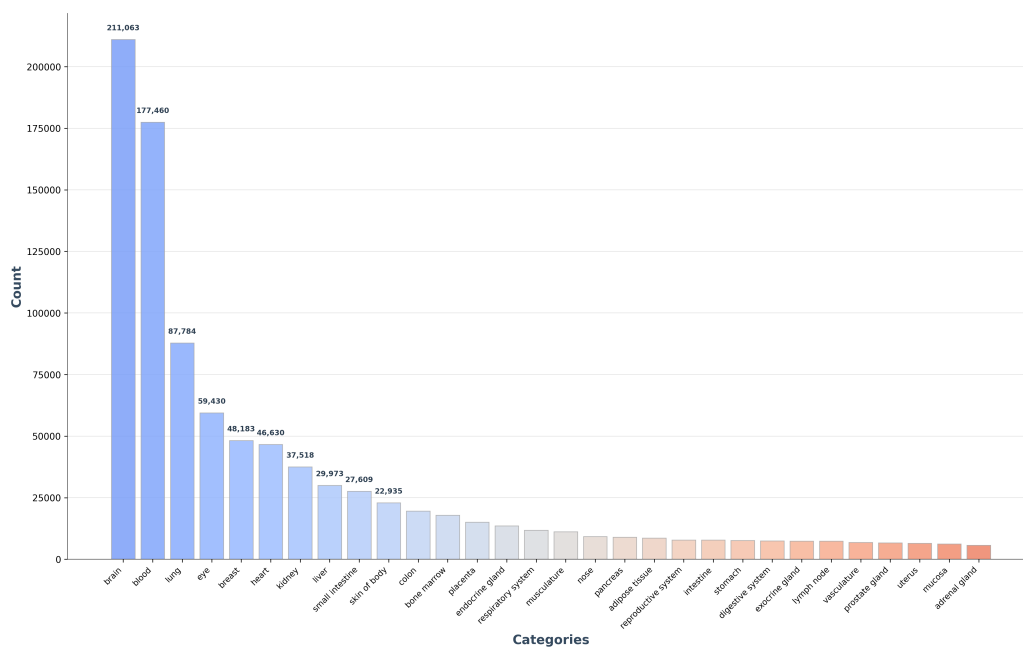


Figure 4.4: Overview of the distribution of disease categories in the dataset. For clarity, only the 30 most abundant categories are shown.



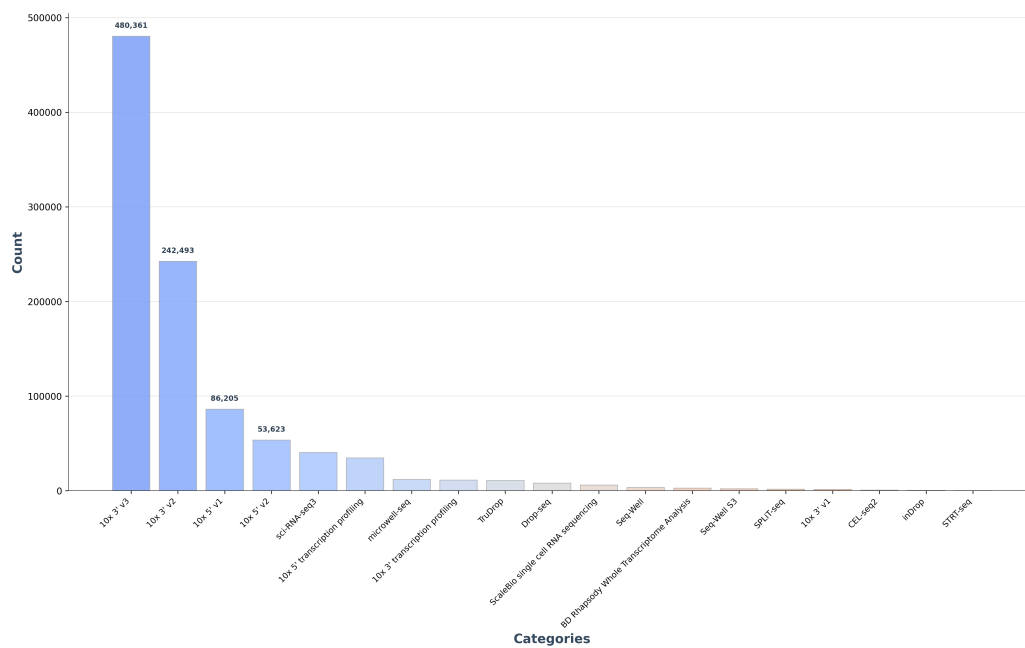


Figure 4.7: Overview of the distribution of assay categories in the dataset.

Chapter 5

Cell2Text: Architecture, Training, Evaluation

5.1 Model Architecture

The `Cell2Text` model is a multimodal generative framework designed to transform single-cell gene expression data into natural language descriptions. At a high level, the system combines:

- A **cell encoder**, Geneformer, which embeds raw gene expression tokens into a continuous latent representation.
- A **projection module**, which maps the encoder’s tokens into the embedding space expected by the decoder.
- A **text decoder**, based on LLaMA, which generates natural language conditioned on the projected cell embeddings.

This architecture allows the system to jointly reason over high-dimensional cell expression profiles and produce coherent textual descriptions, bridging the gap between molecular-level input and human-readable output.

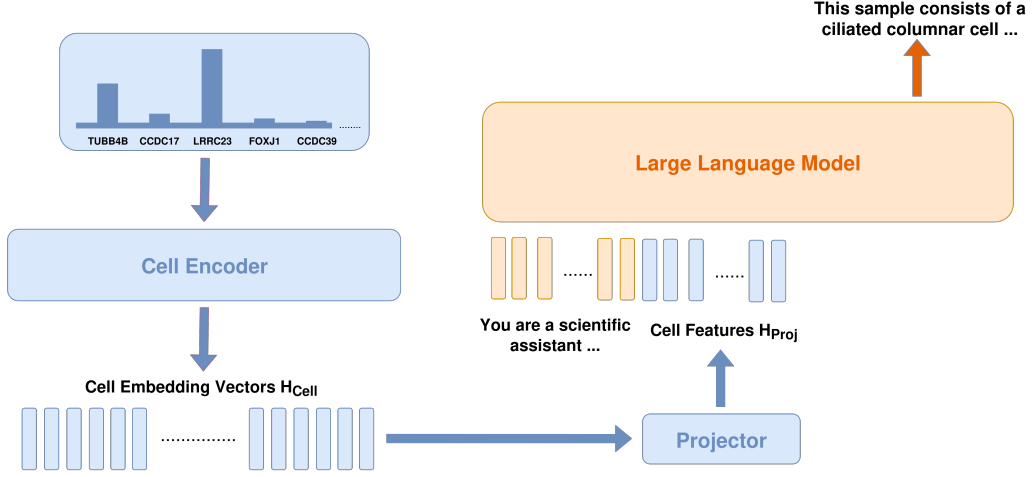


Figure 5.1: High-level architecture of the Cell2Text model.

5.1.1 Cell Encoder: Geneformer

The encoder of Cell2Text is built on top of Geneformer, a transformer-based model trained on 100 millions of single-cell expression profiles. Geneformer tokenizes single-cell gene expression data starting from a count matrix, where it performs normalization using a pre-calculated non-zero median derived from its pretraining corpus. The resulting token sequence consists of rank-valued gene expressions, sorted by their values, which capture the relative expression levels across genes in a cell.

Given this input sequence of gene expression tokens (where each token corresponds to a gene or transcript identifier), Geneformer produces contextualized embeddings that capture co-expression relationships, biological pathways, and regulatory dependencies.

The encoder outputs a tensor of shape (B, L, H) , where B is the batch size, L is the number of input tokens (e.g., up to 4096), and H is the hidden dimension.

Cell2Text frames the overall task as a **sequence-to-sequence (seq2seq)** problem, transforming the sequence of sorted, rank-valued gene expressions into natural language text.

5.1.2 Projection Module

The encoder embeddings and the decoder input embeddings typically live in spaces of different dimensionalities. To align them, Cell2Text employs a **projection module**. This projector is responsible for mapping the cell encoder hidden states (H_{enc}) into the decoder hidden dimension (H_{dec}).

This projection can be approached with two distinct philosophies. The first is

direct token-level translation, where every output token from the cell encoder (e.g., $L = 4096$) is individually mapped to the decoder's space. This preserves the full sequence length but requires the decoder to handle a very long input of biological tokens. The second philosophy is **fixed-length summarization**, which acts as an information bottleneck. This approach uses a small, fixed number of learnable "query" tokens to distill or summarize the entire long sequence of cell embeddings into a short, fixed-length representation (e.g., 32 tokens). We explored both approaches:

1. **MLP Projection Layer:** This implements the *direct translation* philosophy. It is a lightweight 2-layer perceptron with a hidden dimension of 2048, GELU activation, layer normalization, and dropout. It processes each of the L input tokens independently and is computationally efficient.
2. **QFormer Projector:** This implements the *fixed-length summarization* philosophy. Inspired by BLIP-2 and adapted for biomedical applications, this module uses a PubMedBERT-base backbone. Specifically, the self-attention layers and standard feed-forward networks (FFNs) are initialized from the pre-trained PubMedBERT weights, enabling the query tokens to leverage biomedical textual knowledge. In contrast, the cross-attention layers (inserted every 2 layers, allowing the 256 queries to attend to the $L = 4096$ cell embeddings) and query-specific FFNs are randomly initialized. This architecture forces the model to learn a compact, summary representation of the cell. For a visual overview, see Figure 5.2.

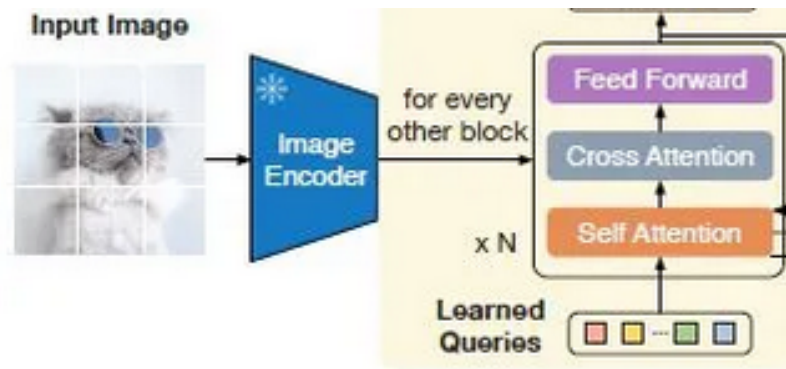


Figure 5.2: Overview of the Q-Former architecture (adapted from BLIP-2). The Q-Former bridges the cell encoder outputs to the decoder space via learnable query tokens that *summarize* the biological information through cross-attention.

5.1.3 Text Decoder: LLaMA

For natural language generation, Cell2Text employs a modified LLaMA decoder. The decoder is a large language model pretrained on textual data, with the following modifications:

- A **placeholder token** mechanism, which allows inserting projected cell embeddings directly into the token stream.
- A **custom input preparation step**, where placeholder tokens in the prompt are replaced by projected cell embeddings before decoding.
- Standard autoregressive generation capabilities (beam search, temperature sampling, top- p filtering).

This design enables the decoder to seamlessly blend cell-derived embeddings with text prompts, producing cell-specific natural language outputs.

5.1.4 Forward Pass

The forward pass through Cell2Text can be summarized as:

1. Encode gene expression tokens with Geneformer to obtain cell embeddings H_{cell} .
2. Project H_{cell} into the decoder space using the selected projector (MLP or QFormer) to get H_{proj} .
3. Insert projected embeddings H_{proj} into the decoder input sequence at placeholder positions.
4. Generate text autoregressively with LLaMA.

Formally, given expression tokens X_{cell} and a textual prompt X_{text} :

$$H_{cell} = \text{Geneformer}(X_{cell}),$$

$$H_{proj} = \text{Projector}(H_{cell}),$$

$$Y = \text{LLaMA}(X_{text}, H_{proj}),$$

where Y denotes the generated natural language sequence.

5.2 Training Strategy

The training of the Cell2Text model is designed to efficiently adapt large pretrained models for the novel task of generating natural language descriptions from single-cell expression data. The overall strategy is modular, with different components either frozen or finetuned depending on their role and the computational resources available.

5.2.1 Frozen Cell Encoder

The **Geneformer** encoder is kept frozen throughout training. Since Geneformer is already pretrained on millions of single-cell profiles, it provides rich contextualized embeddings of expression tokens. Freezing the encoder reduces the computational cost and prevents catastrophic forgetting of its biological representations.

5.2.2 Projection Module Training

The **projection module** is fully trained as it is responsible for aligning the encoder and decoder embedding spaces. Depending on the experiment, we use one of two architectures:

- **MLP projector:** a lightweight 2-layer feedforward network with GELU activations and dropout.
- **QFormer projector:** query tokens interact with the encoder outputs through cross-attention layers, inspired by BLIP-2, allowing parameter-efficient representation learning.

In all cases, the projector is fully trainable.

5.2.3 Decoder Finetuning Strategies

The **LLaMA decoder** is initialized from a pretrained checkpoint and adapted for our task. We explore multiple finetuning strategies that balance computational efficiency with model expressiveness:

5.2.3.1 Full Finetuning with FSDP

For scenarios requiring maximum model capacity and where computational resources permit, we employ **Fully Sharded Data Parallel (FSDP)** training. FSDP is a distributed training paradigm that shards model parameters, gradients, and optimizer states across multiple GPUs, enabling training of models that exceed single-GPU memory capacity.

Key characteristics of our FSDP implementation:

- **Parameter sharding:** Model parameters are partitioned across all available GPUs, with each device holding only a fraction of the full model. During forward and backward passes, parameters are gathered on-demand via all-gather operations.
- **Gradient accumulation:** To simulate larger effective batch sizes, gradients are accumulated over multiple micro-batches before performing optimizer steps. Gradient clipping with a maximum norm of 0.1 is applied to stabilize training.

- **Backward prefetching:** The `BACKWARD_PRE` prefetch strategy is employed to overlap communication and computation during the backward pass, improving training throughput.
- **Synchronization:** The `sync_module_states` flag ensures all processes start with identical model parameters, and `use_orig_params` maintains compatibility with standard PyTorch optimizers.

FSDP enables full finetuning of all decoder parameters while maintaining memory efficiency. This approach achieves maximum flexibility and model expressiveness but requires multi-GPU infrastructure and careful hyperparameter tuning.

5.2.3.2 Parameter-Efficient Finetuning with LoRA and DDP

For resource-constrained settings or rapid experimentation, we employ **Low-Rank Adaptation (LoRA)** combined with **Distributed Data Parallel (DDP)** training. LoRA injects trainable low-rank matrices into the attention layers while keeping the majority of LLaMA parameters frozen.

Key characteristics of our LoRA implementation:

- **Selective adaptation:** Only low-rank adapters are trained, reducing the number of trainable parameters.
- **Rank selection:** We use low-rank bottlenecks to capture task-specific adaptations while maintaining parameter efficiency.
- **DDP parallelism:** Unlike FSDP’s parameter sharding, DDP replicates the full model on each GPU and synchronizes gradients after each backward pass. This is more memory-intensive per device but simpler to implement and debug.
- **Memory footprint:** By freezing the vast majority of decoder parameters, LoRA enables training on single-GPU setups or smaller GPU clusters.

The choice between FSDP and LoRA depends on the available computational budget: FSDP provides maximum model capacity at the cost of implementation complexity, while LoRA offers rapid iteration and lower hardware requirements.

5.2.4 Decoder Input Integration

To integrate cell embeddings into the decoder, placeholder tokens in the input sequence are replaced with projected encoder embeddings during training. The decoder is additionally conditioned with a fixed **system prompt**:

“You are a scientific assistant specialized in analyzing single-cell gene expression data and generating descriptive summaries of cell states and functions.”

This prompt enforces domain-specific generation behavior and guides the model toward scientifically accurate outputs.

5.2.5 Training Objective

We use the standard **causal language modeling loss** (cross-entropy loss) over the generated tokens. Given the input cell embeddings H_{proj} and a target sequence of tokens Y , the objective is:

$$\mathcal{L} = - \sum_{t=1}^T \log P(y_t \mid y_{<t}, H_{proj}),$$

where T is the output sequence length.

5.3 Evaluation Metrics

5.3.1 Ontology-Based Pagerank Cell Type Similarity Metric

To evaluate cell type classification beyond simple accuracy, we employ a similarity metric that captures the biological and hierarchical relatedness between cell types. This metric accommodates "near misses" where a predicted cell type is incorrect but closely related to the true label (e.g., predicting 'T cell' instead of 'CD4+ T cell').

Inspired by SCCELLO [65], we utilize the structured knowledge of the Cell Ontology (CL) [24] to compute these similarities, enabling a nuanced assessment of how well predictions align with true cell types in terms of biological meaning. Unlike SCCELLO, which applies ontology-based similarity for contrastive learning to train a single-cell encoder, our approach uses this metric to evaluate predictions, providing deeper insight into the model's understanding of cell type relationships and its ability to navigate the hierarchical structure of cell biology.

5.3.1.1 Ontology Graph Construction

The Cell Ontology (CL) is represented as a directed acyclic graph (DAG), where nodes correspond to cell types and edges correspond to **is_a** relations (parent-child hierarchical relationships).

Formally, we define the ontology as a graph:

$$\mathcal{G} = (\mathcal{V}, \mathcal{E}), \tag{5.1}$$

where:

- $\mathcal{V}$ is the set of vertices (cell type terms),

- $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ is the set of directed edges such that $(u, v) \in \mathcal{E}$ if u is a parent of v via an `is_a` relation.

For similarity purposes, we treat the graph as *undirected*, allowing information to flow both from parents to children and from children to parents. This choice reflects the intuition that ontological similarity is bi-directional.

5.3.1.2 Personalized PageRank for Cell Types

We employ **Personalized PageRank (PPR)** [75] as the foundation of our similarity metric. Imagine a random walker starting at a cell type (source node) exploring the graph but biased to return to the starting node (personalization). This bias ensures higher scores for biologically related cell types, like subtypes or parents, while distant ones score lower.

For a given source node $s \in \mathcal{V}$, the personalized PageRank vector $\pi^{(s)}$ assigns a stationary distribution over all nodes, indicating how strongly each node is connected to s through the ontology structure. The PageRank vector is defined as the solution to the recurrence:

$$\pi^{(s)} = (1 - \alpha) \mathbf{e}_s + \alpha \pi^{(s)} P, \quad (5.2)$$

where:

- $\alpha \in (0, 1)$ is the damping factor (typically $\alpha = 0.9$),
- $\mathbf{e}_s$ is the personalization vector, with $(\mathbf{e}_s)_i = 1$ if $i = s$, and 0 otherwise,
- P is the transition probability matrix of the random walk on $\mathcal{G}$, with entries

$$P_{ij} = \begin{cases} \frac{1}{\deg(i)} & \text{if } (i, j) \in \mathcal{E}, \\ 0 & \text{otherwise.} \end{cases}$$

Equation (5.2) admits a closed-form solution:

$$\pi^{(s)} = (1 - \alpha) (I - \alpha P)^{-1} \mathbf{e}_s. \quad (5.3)$$

Intuitively, $\pi_j^{(s)}$ reflects the probability that a random walk starting from node s is found at node j in the stationary distribution, where the walk occasionally “teleports” back to s with probability $(1 - \alpha)$.

5.3.1.3 Similarity Definition and Biological Asymmetry

For a cell type c_i , we compute a Personalized PageRank vector with personalization centered on c_i , where $\text{PPR}(c_j|c_i)$ measures relatedness to c_j . The similarity score between two cell types c_i and c_j is defined as:

$$S(c_i, c_j) \propto \log \left(1 + \frac{\text{PPR}(c_j|c_i)}{\tau} \right), \quad (5.4)$$

where τ is a small threshold constant that scales scores before normalization to $[0, 1]$. The diagonal elements are defined as $S(c_i, c_i) = 1$ for all $c_i \in \mathcal{V}$.

5.3.1.4 Distributional Analysis

The similarity matrix $S \in \mathbb{R}^{|\mathcal{V}| \times |\mathcal{V}|}$ is dense but exhibits structure. We analyze its distribution to understand semantic relationships:

- The **diagonal** corresponds to perfect self-similarity: $S_{ii} = 1$.
- The **off-diagonal** values capture semantic closeness between different cell types.
- Many pairs exhibit near-zero similarity (sparsity), reflecting unrelated ontology branches.

Empirically, we observe that similarity scores are heavy-tailed: a small fraction of pairs exhibit high similarity, while the majority are close to zero. This reflects the hierarchical structure of the ontology.

5.3.1.5 Distribution and Statistics of the Similarity Metric

The similarity scores, computed using Personalized PageRank (PPR) on the undirected Cell Ontology graph, exhibit a characteristic distribution that underscores their effectiveness for cell type annotation. As illustrated in Figure 5.3, the distribution of similarity scores across all pairs of cell types is highly skewed, with the vast majority of pairs having near-zero similarity. This reflects the hierarchical and sparse nature of the Cell Ontology, where most cell types are distantly related, and only a small fraction share close ontological relationships.

Quantitatively, the similarity scores range from 0 to 1, with a mean of 0.049 and a median of 0.016, indicating strong right skewness. Notably, 95% of cell type pairs have similarity scores below 0.215, while only the top 1% of pairs achieve similarities above 0.438. These statistics, summarized in Table 5.1, confirm the heavy-tailed nature of the distribution, where high similarity scores are rare and correspond to biologically meaningful relationships (e.g., between “T Cell” and “CD4+ T-cell”).

The cumulative distribution function (CDF), shown in Figure 5.4, further highlights this property. The sharp rise in the CDF at low similarity values demonstrates that a large fraction of pairwise similarities are small, enabling effective discrimination between related and unrelated cell types. This heavy-tailed distribution is crucial for our application, as it ensures that the similarity metric can distinguish closely related cell types (e.g., within the same ontological branch) from unrelated ones, thereby supporting accurate evaluation of LLM predictions.

Statistical analysis confirms the heavy-tailed nature of our similarity distribution. Log-log regression analysis shows strong linearity ($R^2 = 0.862$) with a power-law exponent of $\alpha = 0.67$, while rank-frequency analysis demonstrates excellent fit quality ($R^2 = 0.930$). These results validate that our PageRank-based similarities exhibit the expected heavy-tailed characteristics of hierarchical biological networks, ensuring effective discrimination between closely related and distant cell types

The skewed distribution has important implications for cell type annotation. High similarity scores indicate semantically close cell types, allowing the metric to assign partial credit to “near misses” in LLM predictions (e.g., predicting “T Cell” instead of “CD4+ T-cell”). Conversely, the prevalence of low similarity scores ensures that unrelated cell types (e.g., “T Cell” vs. “Neuron”) are clearly separated, reducing false positives in annotation tasks. This property makes the metric particularly well-suited for bridging the structured knowledge of the Cell Ontology with the statistical predictions of LLMs.

Table 5.1: Summary statistics of the PageRank-based similarity scores across all cell type pairs in the Cell Ontology, highlighting the heavy-tailed distribution.

Statistic	Value
Mean	0.049
Median	0.016
95th Percentile	0.215
Top 1% Threshold	0.438

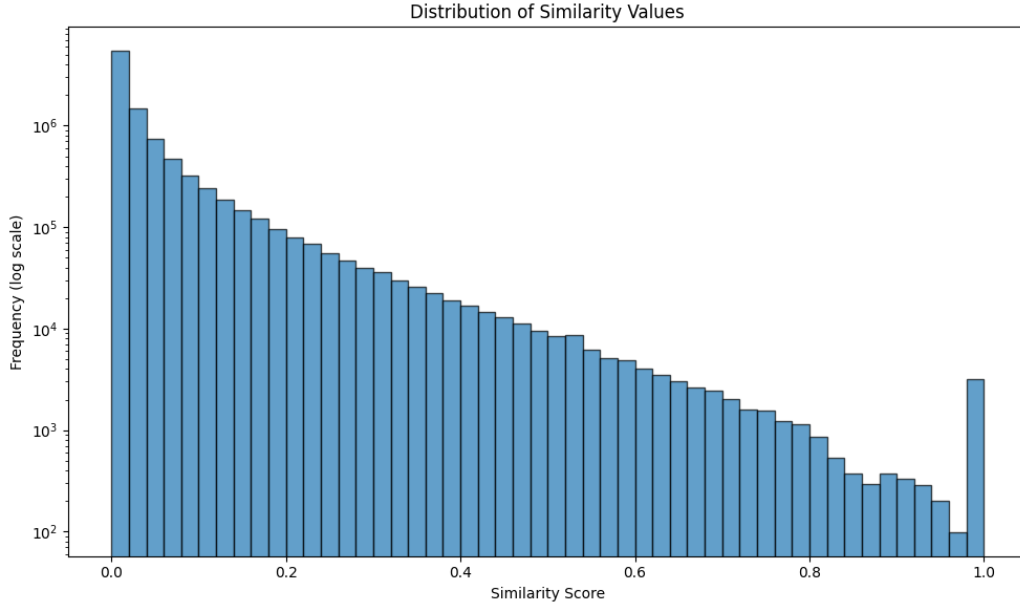


Figure 5.3: Distribution of similarity scores across all cell type pairs in the Cell Ontology, plotted on a logarithmic frequency scale to emphasize the heavy-tailed nature.

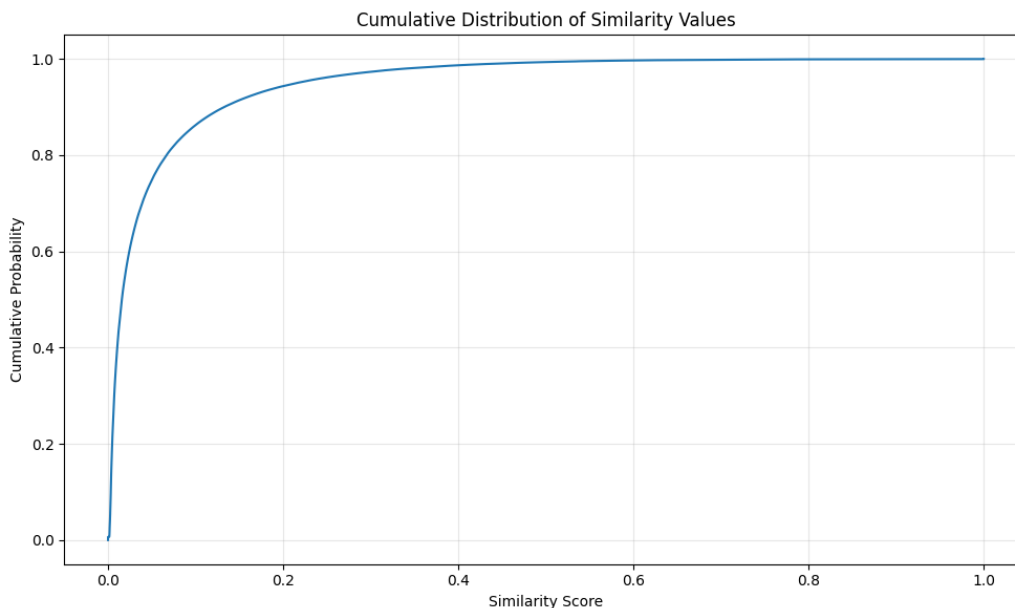


Figure 5.4: Cumulative distribution function (CDF) of similarity scores, showing the rapid accumulation of low similarity values.

5.3.1.6 Application to Cell Type Annotation

This similarity metric provides a principled way to compare candidate cell type labels suggested by LLMs against ontology-defined cell types. Given a predicted label ℓ (normalized to ontology terms), we can rank ontology cell types by $\text{sim}(\ell, \cdot)$, enabling:

- Nearest-neighbor retrieval of most semantically similar cell types,
- Resolution of ambiguous or incomplete predictions,
- Quantitative scoring for annotation confidence.

Thus, the ontology-based PageRank similarity serves as a **bridge between symbolic ontology structure and statistical LLM predictions**.

5.3.2 Evaluation Metrics for Classification

To rigorously assess the performance of our Cell2text models in downstream classification tasks, we employ a comprehensive suite of evaluation metrics tailored to the unique challenges of single-cell data analysis. These metrics span standard classification measures, hierarchical similarity scores, and multi-label evaluation frameworks. Below, we provide a detailed exposition of each metric, including their mathematical formulations and rationale for their selection in the context of cellular attribute prediction and pathway activity identification.

5.3.2.1 Basic Classification Metrics

For the prediction of core cellular metadata, such as cell type, tissue of origin, associated disease, and donor sex, we primarily report *accuracy*, a fundamental metric that quantifies the proportion of correct predictions. Given a dataset with N samples and corresponding true labels $y_i \in \mathcal{Y}$ and predicted labels $\hat{y}_i \in \mathcal{Y}$ for $i = 1, \dots, N$, where $\mathcal{Y}$ denotes the set of possible classes, accuracy is defined as:

$$\text{Accuracy} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(y_i = \hat{y}_i),$$

where $\mathbb{I}(\cdot)$ is the indicator function that equals 1 if the argument is true and 0 otherwise. This metric is particularly suitable for balanced multi-class problems, providing an intuitive measure of overall correctness. However, in high-cardinality settings like cell type classification (where $|\mathcal{Y}|$ can exceed hundreds), accuracy alone may not capture nuances such as class imbalance or hierarchical relationships, motivating the use of complementary metrics. To address potential imbalances and provide a more granular assessment, we also compute *precision*, *recall*, and *F1-score* on a per-class basis, followed by macro-averaging across classes. For a specific class $c \in \mathcal{Y}$, precision Prec_c measures the proportion of true positives among all positive predictions:

$$\text{Prec}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FP}_c},$$

where TP_c is the number of true positives (correctly predicted instances of class c) and FP_c is the number of false positives (instances incorrectly predicted as class c). Similarly, recall Rec_c evaluates the model's ability to identify all actual instances of class c :

$$\text{Rec}_c = \frac{\text{TP}_c}{\text{TP}_c + \text{FN}_c},$$

with FN_c denoting false negatives (actual instances of class c missed by the model). The F1-score harmonizes these by taking their harmonic mean:

$$\text{F1}_c = 2 \cdot \frac{\text{Prec}_c \cdot \text{Rec}_c}{\text{Prec}_c + \text{Rec}_c}.$$

Macro-averaged variants aggregate these per-class scores equally:

$$\text{Prec}_{\text{macro}} = \frac{1}{|\mathcal{Y}|} \sum_{c \in \mathcal{Y}} \text{Prec}_c, \quad \text{Rec}_{\text{macro}} = \frac{1}{|\mathcal{Y}|} \sum_{c \in \mathcal{Y}} \text{Rec}_c, \quad \text{F1}_{\text{macro}} = \frac{1}{|\mathcal{Y}|} \sum_{c \in \mathcal{Y}} \text{F1}_c.$$

These metrics are essential for imbalanced datasets, as they prevent majority classes from dominating the evaluation.

5.3.2.2 Multi-Label Pathway Classification Metrics

Pathway activity identification is inherently multi-label, as multiple pathways may be active in a given cell, and we focus on predicting the top- k (here, $k = 2$) enriched pathways from generated text. Evaluation requires metrics that handle set-based predictions and partial overlaps.

Subset Accuracy This is the strictest measure, requiring exact set agreement between the predicted set $\hat{P}_i \subseteq \mathcal{P}$ (with $|\mathcal{P}| = 34$ Hallmark pathways) and the true set P_i for each sample i :

$$\text{Subset Acc} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(\hat{P}_i = P_i).$$

It penalizes any deviation, making it ideal for assessing precise pathway prioritization.

Jaccard Similarity To credit partial matches, we use the Jaccard index, which quantifies set overlap:

$$J(\hat{P}_i, P_i) = \frac{|\hat{P}_i \cap P_i|}{|\hat{P}_i \cup P_i|},$$

averaged across samples:

$$J_{\text{avg}} = \frac{1}{N} \sum_{i=1}^N J(\hat{P}_i, P_i).$$

For $k = 2$, this assigns 0.5 credit if one pathway matches, balancing rigor with leniency.

Weighted F1-Score For a per-pathway view, we compute binary F1-scores treating each pathway $p \in \mathcal{P}$ as a separate classification task. Let $y_{i,p} \in 0, 1$ indicate true activity of pathway p in sample i , and $\hat{y}_{i,p} \in 0, 1$ the prediction (derived by thresholding extracted pathways). The per-pathway F1 is:

$$\text{F1}_p = 2 \cdot \frac{\text{Prec}_p \cdot \text{Rec}_p}{\text{Prec}_p + \text{Rec}_p},$$

with $\text{Prec}_p = \frac{\sum_i \mathbb{I}(y_{i,p}=1, \hat{y}_{i,p}=1)}{\sum_i \mathbb{I}(\hat{y}_{i,p}=1)}$ and analogously for recall. The weighted F1 aggregates these, weighting by pathway prevalence $\pi_p = \frac{1}{N} \sum_i y_{i,p}$:

$$\text{F1}_{\text{weighted}} = \sum_{p \in \mathcal{P}} \pi_p \cdot \text{F1}_p.$$

This accounts for imbalance, emphasizing frequently active pathways.

5.3.3 Evaluation Metrics for Text Generation Quality

To evaluate the quality of the natural language descriptions generated by our Cell2text models, we employ a suite of metrics that assess both lexical overlap and semantic fidelity. These metrics are crucial for ensuring that the generated text is not only fluent and structurally sound but also scientifically accurate within the biomedical domain. Below, we describe the metrics used, including their mathematical formulations and their relevance to evaluating generative performance in the context of single-cell data.

5.3.3.1 Lexical Metrics

Lexical metrics focus on surface-level similarity between generated and reference texts, capturing aspects like fluency and content coverage. We use BLEU and ROUGE to evaluate these properties.

BLEU (Bilingual Evaluation Understudy) BLEU measures the precision of n-gram overlaps between the generated text $G = \{g_1, g_2, \dots, g_m\}$ and the reference text $R = \{r_1, r_2, \dots, r_n\}$. Specifically, we compute BLEU-2 (B-2) and BLEU-4 (B-4), which consider bigrams and 4-grams, respectively. For an n-gram order k , the BLEU score is defined as:

$$\text{BLEU-}k = \text{BP} \cdot \exp \left(\sum_{i=1}^k w_i \log p_i \right), \quad (5.5)$$

where p_i is the precision of i-grams, calculated as:

$$p_i = \frac{\sum_{\text{i-gram} \in G} \min(\text{Count}_G(\text{i-gram}), \text{Count}_R(\text{i-gram}))}{\sum_{\text{i-gram} \in G} \text{Count}_G(\text{i-gram})}, \quad (5.6)$$

and BP is the brevity penalty to penalize overly short outputs:

$$\text{BP} = \begin{cases} 1 & \text{if } |G| > |R|, \\ \exp \left(1 - \frac{|R|}{|G|} \right) & \text{otherwise,} \end{cases} \quad (5.7)$$

with $|G|$ and $|R|$ denoting the lengths of the generated and reference texts. The weights w_i are typically uniform (e.g., $w_i = 1/k$). BLEU emphasizes precision and fluency, making it suitable for assessing whether the generated descriptions use terminology consistent with reference texts.

ROUGE (Recall-Oriented Understudy for Gisting Evaluation) ROUGE measures recall by evaluating the overlap of n-grams or sequences between generated and reference texts. We report ROUGE-1 (R-1), ROUGE-2 (R-2), and ROUGE-L (R-L, for longest common subsequence). For a given n-gram order, ROUGE-N is:

$$\text{ROUGE-N} = \frac{\sum_{\text{n-gram} \in R} \min(\text{Count}_G(\text{n-gram}), \text{Count}_R(\text{n-gram}))}{\sum_{\text{n-gram} \in R} \text{Count}_R(\text{n-gram})}, \quad (5.8)$$

where the numerator counts matched n-grams, and the denominator is the total number of n-grams in the reference. ROUGE-L extends this to the longest common subsequence (LCS), capturing structural similarity:

$$\text{ROUGE-L} = \frac{\text{LCS}(G, R)}{|R|}, \quad (5.9)$$

where $\text{LCS}(G, R)$ is the length of the longest common subsequence. ROUGE prioritizes content coverage, ensuring the generated text captures key elements of the reference, which is critical for scientific accuracy in our domain.

5.3.3.2 Semantic Metrics

Lexical metrics alone cannot capture the scientific correctness of generated descriptions, as paraphrasing may alter surface forms while preserving meaning. To address this, we use BERTScore, which leverages contextual embeddings to measure semantic similarity, with variants tailored to general and biomedical contexts.

BERTScore BERTScore computes similarity between generated and reference texts using contextual embeddings from pre-trained language models. For each token in the generated text $g_j \in G$ and reference text $r_k \in R$, embeddings are obtained from a model (e.g., RoBERTa or BioBERT). The similarity is calculated via cosine similarity:

$$\text{sim}(g_j, r_k) = \frac{\mathbf{e}_{g_j} \cdot \mathbf{e}_{r_k}}{\|\mathbf{e}_{g_j}\| \|\mathbf{e}_{r_k}\|}, \quad (5.10)$$

where $\mathbf{e}_{g_j}$ and $\mathbf{e}_{r_k}$ are the embeddings. Precision, recall, and F1-scores are computed as:

$$P_{\text{BERT}} = \frac{1}{|G|} \sum_{g_j \in G} \max_{r_k \in R} \text{sim}(g_j, r_k), \quad (5.11)$$

$$R_{\text{BERT}} = \frac{1}{|R|} \sum_{r_k \in R} \max_{g_j \in G} \text{sim}(g_j, r_k), \quad (5.12)$$

$$F1_{\text{BERT}} = 2 \cdot \frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}, \quad (5.13)$$

averaged over all samples. We report:

- **RoBERTa F1 (RBT-f1)**: Uses RoBERTa-large, a robustly optimized BERT model pre-trained on diverse general-domain corpora, to capture semantic similarity in natural language. This provides a baseline for general semantic correctness independent of domain-specific terminology.
- **BioBERT F1 (BBT-f1)**: Uses BioBERT, fine-tuned on biomedical corpora (PubMed abstracts and PMC full-text articles), to assess domain-specific semantic correctness. This metric is critical for evaluating the scientific validity of generated descriptions, as it is sensitive to biomedical terminology and concepts.

BERTScore’s use of contextual embeddings ensures robustness to paraphrasing, making it ideal for verifying that generated texts preserve the biological meaning of reference descriptions. The combination of RoBERTa and BioBERT variants provides both general linguistic quality assessment and domain-specific accuracy evaluation.

These metrics collectively provide a comprehensive evaluation of text generation quality. BLEU and ROUGE ensure lexical and structural fidelity, while BERTScore (with RoBERTa and BioBERT variants) validates semantic and domain-specific accuracy, aligning with the needs of biomedical text generation.

Chapter 6

Experiments & Results

6.1 Experimental Setup

6.1.1 Hardware and Software Environment

All experiments were conducted on a distributed computing setup utilizing $2 \times$ NVIDIA A6000 GPUs, each with 48GB of memory. The models were trained using PyTorch with distributed data parallel (DDP) for LoRA-based fine-tuning and fully sharded data parallel (FSDP) for full fine-tuning configurations.

6.1.2 Model Architecture

The Cell2Text framework consists of three main components: a GeneFormer encoder, a projector module, and a LLaMA decoder. We experimented with two projector architectures:

- **MLP Projector:** A two-layer multi-layer perceptron with hidden dimension of 2048. The architecture maps from the GeneFormer output dimension (1152) to 2048, and then projects to the LLaMA input dimension (2048 for LLaMA-3.2-1B and 3072 for LLaMA-3.2-3B).
- **Q-Former Projector:** A query-based transformer architecture with 512 learnable query tokens, adapting the representation between encoder and decoder modalities.

We evaluated two LLaMA decoder variants: LLaMA-3.2-1B and LLaMA-3.2-3B, resulting in four model configurations in total.

6.1.3 Training Configuration

6.1.3.1 Dataset

The dataset consisted of 200,000 single-cell samples with corresponding textual descriptions. We used an 80/10/10 split for training, validation, and testing, resulting in 160,000 training samples, 20,000 validation samples, and 20,000 test samples.

6.1.3.2 Baseline: Geneformer with Classification Head

For the baseline Geneformer model, we added a classification head on top of the pretrained encoder and fine-tuned it for cell type, tissue, and disease prediction tasks. Training was conducted using the AdamW optimizer with weight decay of 0.01 and an initial learning rate of 5×10^{-5} . The model was trained for 3 epochs with a batch size of 64 per GPU, employing gradient clipping at a maximum norm of 1.0 and automatic mixed precision (AMP) to enhance stability and efficiency. Training was performed using distributed data parallelism (DDP) across two NVIDIA A6000 GPUs.

6.1.3.3 Baseline: scGPT with Classification Head

For the scGPT baseline, we fine-tuned the pretrained scGPT-human model on downstream classification tasks. The model was trained for 3 epochs using the Adam optimizer with an initial learning rate of 1×10^{-4} and a StepLR scheduler with a decay factor of 0.9 per epoch. We used a batch size of 16 per GPU, enabling automatic mixed precision (AMP) and gradient clipping at a maximum norm of 1.0. All training was conducted using distributed data parallelism (DDP).

6.1.3.4 Cell2Text Model Configurations

All Cell2Text configurations were trained for 3 epochs across two NVIDIA A6000 GPUs. Due to varying memory requirements of different architectures, we adjusted batch sizes and gradient accumulation steps to maintain consistent effective batch sizes while maximizing GPU utilization. Table 6.1 summarizes the training hyperparameters for each configuration.

Table 6.1: Training hyperparameters for each Cell2Text configuration.

Configuration	Batch/GPU	Grad Acc	Eff. Batch	LR
LLaMA-3B + LoRA + MLP	2	16	64	1×10^{-4}
LLaMA-3B + LoRA + Q-Former	4	8	64	1×10^{-4}
LLaMA-1B + FSDP + MLP	2	16	64	1×10^{-7}
LLaMA-1B + FSDP + Q-Former	4	8	64	1×10^{-7}

The Q-Former configurations use larger per-GPU batch sizes (4 vs. 2) with correspondingly smaller gradient accumulation steps (8 vs. 16) to maintain an effective batch size of 64 across all experiments. This adjustment accounts for the Q-Former’s more efficient memory usage: the Q-Former compresses the input into 512 learnable query tokens that are fed to the LLaMA decoder, whereas the MLP projector passes through a larger number of tokens from the full GeneFormer output sequence. This difference in sequence length results in higher memory consumption for MLP configurations during the decoder forward pass. All configurations maintain identical effective batch sizes to ensure fair comparison.

6.1.3.5 Optimization

All Cell2Text models were trained using the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 1 \times 10^{-8}$. We employed a linear learning rate schedule with warmup: the learning rate linearly increases from 0 to the specified maximum over the first 10 training steps, then linearly decays to 0 over the remaining steps. Gradient clipping with a maximum norm of 1.0 was applied to stabilize training.

The LoRA configurations use a significantly higher learning rate (1×10^{-4}) with zero weight decay, as parameter-efficient fine-tuning typically requires stronger learning signals and benefits from less regularization. In contrast, the FSDP configurations employ a much lower learning rate (1×10^{-7}) with weight decay of 0.01, as full parameter fine-tuning of large language models requires more conservative optimization to prevent catastrophic forgetting of pretrained knowledge.

6.1.4 Fine-tuning Strategies

- **LoRA (Low-Rank Adaptation):** Applied to the LLaMA-3.2-3B model, with rank $r = 128$ and $\alpha = 256$ for the decoder layers, enabling parameter-efficient fine-tuning by adapting low-rank decomposition matrices. Training was distributed using DDP.
- **FSDP (Fully Sharded Data Parallel):** Applied to the LLaMA-3.2-1B model, enabling full parameter fine-tuning with memory-efficient sharding across GPUs.

6.2 Results

6.2.1 Text Generation Quality

We evaluated the quality of generated cell descriptions using automated text generation metrics. Table 6.2 presents the performance of all four model configurations across BLEU score and embedding-based similarity metrics.

Table 6.2: Text generation quality metrics across model configurations. Higher scores indicate better generation quality. Metrics: **BLEU** (Bilingual Evaluation Understudy), **ROUGE** (Recall-Oriented Understudy for Gisting Evaluation), **BBT** (BioBERT F1-Score), **RBT** (RoBERTa F1-Score).

Model Configuration	B-4	R-2	R-L	BBT	RBT
LLaMA-3B + LoRA + MLP	48.83	57.26	63.38	86.15	89.75
LLaMA-3B + LoRA + Q-Former	<u>57.46</u>	<u>65.57</u>	70.56	90.01	93.56
LLaMA-1B + FSDP + MLP	57.84	65.99	69.87	88.43	91.56
LLaMA-1B + FSDP + Q-Former	56.78	65.25	<u>70.47</u>	<u>89.85</u>	<u>93.53</u>

The text generation results reveal several interesting patterns across model configurations and training strategies. The LLaMA-3.2-3B model with LoRA fine-tuning and Q-Former projector achieves the best performance on semantic similarity metrics, with BioBERT F1 of 90.01% and RoBERTa F1 of 93.56%, demonstrating superior ability to capture biomedical semantic content despite being parameter-efficient. This configuration also obtains the highest ROUGE-L score (70.56%), indicating better preservation of the longest common subsequences with reference texts and suggesting more coherent long-range text generation.

Notably, the introduction of the MLP projector to the 3B LoRA configuration results in a substantial performance drop across all metrics, with BioBERT F1 decreasing by 3.86 points to 86.15% and RoBERTa F1 dropping to 89.75%. This indicates that for larger models with parameter-efficient training, the Q-Former’s learnable query mechanism is crucial for effectively bridging the gap between GeneFormer’s biological representations and LLaMA’s text generation capabilities.

Surprisingly, the smaller LLaMA-3.2-1B model with full FSDP fine-tuning and MLP projector achieves the highest scores on BLEU-4 (57.84%) and ROUGE-2 (65.99%), outperforming even the larger 3B model with Q-Former on these n-gram-based metrics. This suggests that full parameter updates combined with simpler projection may produce text with better surface-level n-gram overlap with references. The fully fine-tuned 1B model appears to benefit from complete parameter adaptation, compensating for the simpler projector architecture.

The performance difference between Q-Former and MLP projectors appears to depend critically on both model size and training strategy. For the 1B model with full fine-tuning, the MLP projector shows competitive or superior performance across n-gram metrics compared to Q-Former, with substantially fewer parameters in the projection layer. However, the Q-Former still provides advantages for the 1B model on semantic metrics (89.85% vs 88.43% BioBERT F1). For the 3B model restricted to LoRA training, the Q-Former becomes essential, suggesting that when decoder parameters are frozen or rank-limited, a more sophisticated projector is needed to compensate.

The consistent gap between embedding-based metrics (BioBERT, RoBERTa)

and n-gram-based metrics (BLEU, ROUGE) suggests that all models capture biomedical semantics reasonably well (86-93% F1 range), but differ in their exact lexical choices and surface realization. The high RoBERTa F1 scores (89-93%) across all configurations indicate that generated descriptions are semantically similar to references even when exact word choices differ, which is particularly appropriate for scientific text generation where multiple valid phrasings exist for describing the same biological phenomena.

6.2.2 Biological Entity Recognition

We assessed the models’ ability to correctly identify and extract biological entities from generated descriptions. Tables 6.3 and 6.4 present detailed performance metrics for cell type, disease, and tissue recognition respectively.

6.2.2.1 Cell Type Annotation

Table 6.3: Cell type recognition performance. Cell Type Accuracy measures exact matches, while **Ont Sim.** (**Ontology Similarity**) captures semantic closeness in the Cell Ontology hierarchy. The other metric is weighted **F1** score.

Model Configuration	Acc	F1	Ont. Sim.
Geneformer (Baseline)	64.52	62.23	75.21
scGPT (Baseline)	66.55	65.01	75.63
LLaMA-3B + LoRA + MLP	59.50	62.82	70.50
LLaMA-3B + LoRA + Q-Former	<u>67.83</u>	<u>68.52</u>	78.52
LLaMA-1B + Full + MLP	67.95	72.76	73.49
LLaMA-1B + Full+ Q-Former	64.21	65.66	<u>75.86</u>

Cell type annotation represents one of the most challenging aspects of single-cell analysis due to the hierarchical nature of cell type taxonomies. The Cell Ontology contains thousands of cell types with complex parent-child relationships, where subtle distinctions between closely related types require precise understanding of gene expression patterns.

Several Cell2Text configurations demonstrate improved performance over the **strongest baseline, scGPT (66.55% accuracy, 65.01% F1)**, validating the hypothesis that natural language generation can enhance cell type prediction. Notably, the LLaMA-3.2-3B + LoRA + MLP configuration underperforms **both baselines** with 59.50% accuracy and 62.82% F1, highlighting that parameter-efficient training with simple projection is insufficient for this fine-grained classification task.

The LLaMA-3.2-1B + FSDP + MLP configuration achieves the best overall performance with 67.95% accuracy and 72.76% F1 score, representing improvements

of **1.40** and **7.75** percentage points over the **scGPT** baseline. This result is particularly notable given the smallest model size and simplest projector, suggesting that full fine-tuning enables effective direct mapping of gene expression patterns to cell type identity.

The ontology similarity metrics provide crucial insights into error patterns. Overall ontology similarity scores (70.50-78.52%) substantially exceed exact match accuracies (59.50-67.95%), indicating that prediction errors typically involve semantically related cell types. For example, confusing "vasa recta ascending limb cell" with "vasa recta descending limb cell" represents a near-miss rather than fundamental misunderstanding.

The LLaMA-3.2-3B + Q-Former achieves the highest overall ontology similarity (78.52%), **surpassing all other configurations and baselines by a significant margin (vs. 75.63% for scGPT)**. This indicates superior capture of hierarchical cell type relationships, though at the cost of lower F1 (68.52%). **Interestingly, the 1B + FSDP + MLP model, which was best for Acc/F1, scores lower on this semantic metric (73.49%) than the scGPT baseline (75.63%), highlighting a clear trade-off between exact-match accuracy and deeper semantic understanding.**

The training strategy proves critical: the 3B model with LoRA benefits substantially from Q-Former over MLP (68.52% vs. 62.82% F1, 78.52% vs. 70.50% ontology similarity), while the 1B model with full FSDP fine-tuning achieves best results with the simpler MLP projector. This indicates that when decoder parameters are heavily constrained via LoRA, sophisticated projection mechanisms become essential for performance.

6.2.2.2 Disease and Tissue Prediction

Table 6.4: Disease and tissue recognition performance across model configurations. Metrics include **Acc** (Accuracy) and weighted **F1** score. **Bold** indicates best performance, underline indicates second best.

Model Configuration	Disease		Tissue	
	Acc	F1	Acc	F1
Geneformer (Baseline)	62.50	59.82	59.63	58.31
scGPT (Baseline)	63.15	62.12	60.56	60.25
LLaMA-3B + LoRA + MLP	61.13	66.32	59.23	64.17
LLaMA-3B + LoRA + Q-Former	<u>63.80</u>	<u>67.76</u>	<u>61.65</u>	<u>65.79</u>
LLaMA-1B + Full + MLP	65.34	70.45	62.39	67.88
LLaMA-1B + Full + Q-Former	62.16	65.23	58.05	61.47

Disease and tissue prediction results, presented in Table 6.4, follow similar patterns to those observed in cell type annotation, further validating the architectural insights from the previous section. The LLaMA-3.2-1B model with FSDP training and MLP projector again achieves the best performance across both tasks, with disease prediction accuracy of 65.34% and F1 of 70.45%, and tissue prediction accuracy of 62.39% and F1 of 67.88%. These results represent substantial improvements over the strongest baseline, **scGPT** (2.19 percentage points on disease accuracy and 8.33 points on F1; 1.83 points on tissue accuracy and 7.63 points on F1), further validating the effectiveness of the text generation approach. The consistent superiority of the 1B + FSDP + MLP configuration across all three entity types (cell type, disease, tissue) suggests that full fine-tuning with direct projection provides a robust and generalizable solution for diverse single-cell annotation tasks.

Disease prediction shows the strongest overall performance among the three entity types, with most Cell2Text configurations achieving F1 scores above 65% (ranging from 65.23-70.45%), compared to cell type F1 scores ranging from 62.82-72.76%. This superior performance may be attributed to the relatively smaller and more distinct label space for diseases. The 3B + LoRA + Q-Former configuration achieves competitive performance (63.80% accuracy, 67.76% F1), maintaining its position as the second-best configuration. The 3B + LoRA + MLP configuration also shows improved performance over the **scGPT** baseline on disease F1 (**66.32% vs. 62.12%**), though it still lags behind the Q-Former variant, reinforcing that LoRA-trained models benefit from sophisticated projection mechanisms.

Tissue prediction proves to be the most challenging task among the three entity types, with accuracies ranging from 58.05% to 62.39%. This increased difficulty likely stems from multiple factors: tissue origin can be inherently ambiguous for certain cell types. **The performance gap between our best model and the scGPT baseline is slightly smaller for tissue (e.g., 7.63 points F1 gain) than for disease (8.33 points F1 gain)**, suggesting that tissue information may be less directly or consistently encoded in the gene expression patterns. Interestingly, the 1B + Q-Former configuration performs below both baselines on tissue accuracy (**58.05% vs. 60.56%** for scGPT), which may indicate that the query-based projection introduces additional complexity that hinders learning when model capacity is limited and the prediction signal is weaker. Nevertheless, the text generation approach still provides meaningful improvements across all configurations on the F1 metric, demonstrating the framework’s ability to extract and represent diverse biological attributes from single-cell transcriptomics data.

6.2.3 Pathway Activity Identification

Pathway prediction represents a multi-label classification task where each cell is associated with multiple active biological pathways simultaneously. In our framework, we predict the top-2 most active pathways for each cell based on AUCell scores from the reference annotations, transforming this into a fixed-size multi-label

problem. Table 6.5 presents performance metrics that capture different aspects of multi-label prediction quality.

Table 6.5: Pathway prediction performance metrics. **Sub. Acc.** (**Subset Accuracy**) requires exact match of both predicted pathways, **Jac. Sim.** (**Jaccard Similarity**) measures set overlap allowing partial matches, and **W-F1** (**Weighted F1**) accounts for pathway frequency in the dataset. **Bold** indicates best performance, underline indicates second best.

Model Configuration	Sub. Acc.	Jac. Sim.	W-F1
Geneformer (Baseline)	35.55	49.44	56.61
scGPT (Baseline)	<u>36.13</u>	<u>50.13</u>	<u>58.35</u>
LLaMA-3B + LoRA + MLP	30.02	42.77	51.34
LLaMA-3B + LoRA + Q-Former	36.02	49.58	57.96
LLaMA-1B + Full + MLP	36.42	50.67	59.10
LLaMA-1B + Full + Q-Former	35.75	48.57	57.12

Pathway prediction proves to be the most challenging task in our evaluation, with subset accuracies reaching only 35-36% compared to 62-68% for cell type annotation and 61-65% for disease prediction. This substantial difficulty arises from the combinatorial complexity of predicting two pathways simultaneously. The LLaMA-3.2-1B model with FSDP training and MLP projector maintains its position as the best-performing configuration with 36.42% subset accuracy and 59.10% weighted F1. However, the improvements over the **strongest baseline** (**scGPT**) are notably more modest (**0.29** percentage points in subset accuracy, **0.75** points in weighted F1) compared to the entity recognition tasks. Particularly striking is the underperformance of the 3B + LoRA + MLP configuration, which falls **6.11** percentage points below the **scGPT** baseline in subset accuracy (30.02% vs. **36.13%**), representing the most severe performance degradation observed.

The gap between subset accuracy (36.42%) and Jaccard similarity (50.67%) for the best model reveals important insights into error patterns: models frequently predict one correct pathway while missing the other. This partial success rate of approximately 50% suggests that the models capture pathway-level information and biological relevance reasonably well but struggle with the precise ranking and threshold determination. The weighted F1 scores (57-59% for successful configurations) being substantially higher than subset accuracy indicates that models perform better on more frequent pathway combinations. The consistently superior performance of the 1B + FSDP + MLP configuration across all three metrics (subset accuracy, Jaccard similarity, weighted F1), combined with its dominance in cell type, disease, and tissue prediction, reinforces its effectiveness as the most robust and generalizable architectural choice for the Cell2Text framework.

Chapter 7

Conclusion & Future Work

7.1 Conclusion

In this thesis, we presented Cell2Text, a multimodal generative framework that bridges single-cell transcriptomics and natural language by generating interpretable textual descriptions from gene expression data. By integrating pretrained single-cell foundation models with instruction-tuned large language models through learnable projection modules, our approach transforms the traditional cell annotation paradigm from discriminative classification to generative description, enabling both interpretable outputs and improved prediction performance.

Our work demonstrates that end-to-end training of multimodal systems for generating natural language descriptions creates richer biological representations than traditional classification approaches. Rather than treating annotation as isolated prediction tasks, the text generation objective encourages models to learn integrated biological semantics, the relationships between gene expression, cellular function, tissue context, and disease states. This joint learning paradigm not only improves quantitative performance across multiple prediction tasks but also provides human-readable explanations that make the model’s reasoning transparent and verifiable by domain experts.

The ontology similarity metrics reveal that our models develop biologically coherent understanding of cell type relationships. Even when exact predictions are incorrect, errors typically occur within semantically related cell type families, demonstrating that the model has internalized the hierarchical structure of cellular taxonomy rather than memorizing surface-level label associations. This ontological coherence suggests that end-to-end multimodal training naturally captures the compositional structure of biological knowledge.

More broadly, Cell2Text demonstrates that the integration of domain-specific pretrained foundation models with large language models offers a general and scalable strategy for building interpretable AI systems in computational biology. By leveraging the complementary strengths of specialized biological encoders and

general-purpose language models, our framework provides a template for multimodal approaches that extend beyond single-cell annotation. This paradigm can be applied to diverse challenges in computational biology where complex scientific data must be transformed into human-interpretable insights, from spatial transcriptomics to drug response prediction to automated literature synthesis.

7.2 Future Work

The Cell2Text framework establishes a robust foundation for integrating single-cell genomics and natural language processing. Future work can be organized along three primary axes: (1) reversing the generative direction to enable *in silico* hypothesis generation, (2) enhancing the model’s reasoning capabilities through advanced training strategies, and (3) expanding the model’s input modalities for a more holistic cellular view.

Text-to-Cell Generation for Hypothesis Discovery Perhaps the most exciting, yet challenging, future direction is reversing the generative process: a **"Text-to-Cell" framework**. Such a model would be conditioned on a natural language prompt—such as a cell type description, a disease state, or a hypothetical perturbation—and would **generate a corresponding *in silico* gene expression profile** or other gene-level information. This would transform the model from a descriptive tool into a predictive and hypothesis-generating engine. The potential for *in silico* discovery is substantial. For example, recent work on generative cell-to-text models, such as Google’s **cell2sentence**, has already demonstrated that analyzing the learned latent space can uncover novel cancer-related pathways. A Text-to-Cell model would take this a step further, allowing researchers to directly query for novel biology (e.g., "Generate the expression profile of a T-cell resistant to immunotherapy"). The model’s generated output could then be experimentally validated, opening new frontiers in computational biology and drug discovery.

Ontology-Based Curriculum Learning A key challenge in single-cell classification is handling the extreme label imbalance and hierarchical complexity of cell types. Future work could address this by implementing an **ontology-based curriculum learning** strategy. This approach would leverage the Cell Ontology (CL) graph to define a training curriculum from "easy" to "hard." The model would first be trained on **easy, high-level distinctions** (e.g., distinguishing "immune cell" from "neural cell," which are on distant branches of the ontology). Once it masters this, the curriculum would gradually introduce **harder, fine-grained distinctions** (e.g., "CD4+ T cell" vs. "CD8+ T cell," which are closely related leaf nodes). The hypothesis is that this strategy would allow the model to first learn a robust, general representation of the major biological "big pictures" before attempting to resolve the subtle differences required for fine-grained annotation,

potentially improving accuracy and generalization for rare cell types.

Expanding Input Modalities for Holistic Cell Descriptions The current model operates solely on single-cell transcriptomics. A major avenue for future work is to enhance the richness of the input "cell" modality by integrating other forms of single-cell data into a unified, multimodal encoder. A promising direction is the integration of **CITE-seq** (Cellular Indexing of Transcriptomes and Epitopes by Sequencing) data, which would allow the model to generate descriptions that capture both RNA expression and surface protein levels. This could be further extended to incorporate **spatial transcriptomics**, enabling the model to describe cells within their native tissue microenvironment—answering not just "what" a cell is, but "where" it is and "who" its neighbors are. Finally, incorporating **temporal data** from perturbation or developmental time-course experiments would unlock the ability to generate dynamic narratives of cellular processes, state transitions, and differentiation trajectories, providing a truly comprehensive model of the cell.

Interactive Refinement and Human Feedback. Implementing mechanisms for biologists to interactively refine generated descriptions and incorporating their feedback through reinforcement learning from human feedback (RLHF) could align the model more closely with domain expert knowledge and preferences, potentially improving both accuracy and usefulness in practical research settings.

Cross-Species Generalization. Evaluating Cell2Text’s ability to generalize across species (e.g., from human to mouse or other model organisms) would test the universality of the learned cellular representations and could reveal fundamental principles of cell biology that transcend species boundaries.

These future directions highlight the rich potential for advancing multimodal approaches in computational biology, with Cell2Text serving as a foundation for increasingly sophisticated and interpretable AI systems that can assist researchers in understanding the complex language of cellular states and behaviors.

Bibliography

- [1] Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618, 2023.
- [2] Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21, 2024.
- [3] Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B. Tuch, Asim Siddiqui, Kaiqin Lao, and M. Azim Surani. mrna-seq whole-transcriptome analysis of a single cell. *Nature Methods*, 6, 2009.
- [4] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15, 2018.
- [5] Juexin Wang, Anjun Ma, Yuzhou Chang, Jianting Gong, Yuexu Jiang, Ren Qi, Cankun Wang, Hongjun Fu, Qin Ma, and Dong Xu. scgcn is a novel graph neural network framework for single-cell rna-seq analyses. *Nature Communications*, 12, 2021.
- [6] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 2017-December, 2017.
- [7] Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, 2019.
- [8] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry,

- Amanda Aspell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 2020-December, 2020.
- [9] Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. In *ICLR 2022 - 10th International Conference on Learning Representations*, 2022.
- [10] CZI Cell Science Program, Shibli Abdulla, Brian Aevertmann, Pedro Assis, Seve Badajoz, Sidney M Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Pablo Garcia-Nieto, Nayib Gloria, Mim Hastie, Daniel Hegeman, Jason Hilton, Timmy Huang, Amanda Infeld, Ana-Maria Istrate, Ivana Jelic, Kuni Katsuya, Yang Joon Kim, Karen Liang, Mike Lin, Maximilian Lombardo, Bailey Marshall, Bruce Martin, Fran McDade, Colin Megill, Nikhil Patel, Alexander Predeus, Brian Raymor, Behnam Robatmili, Dave Rogers, Erica Rutherford, Dana Sadgat, Andrew Shin, Corinn Small, Trent Smith, Prathap Sridharan, Alexander Tarashansky, Norbert Tavares, Harley Thomas, Andrew Tolopko, Meghan Urisko, Joyce Yan, Garabet Yeretssian, Jennifer Zamanian, Arathi Mani, Jonah Cool, and Ambrose Carr. Cz cellxgene discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Research*, 53(D1):D886–D900, 11 2024.
- [11] Sara Aibar, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, Jean Christophe Marine, Pierre Geurts, Jan Aerts, Joost Van Den Oord, Zeynep Kalender Atak, Jasper Wouters, and Stein Aerts. Scenic: Single-cell regulatory network inference and clustering. *Nature Methods*, 14, 2017.
- [12] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: A revolutionary tool for transcriptomics, 2009.
- [13] Aviv Regev, Sarah A. Teichmann, Eric S. Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, Hans Clevers, Bart Deplancke, Ian Dunham, James Eberwine, Roland Eils, Wolfgang Enard, Andrew Farmer, Lars Fugger, Berthold Göttgens, Nir Hacohen, Muzlifah Haniffa, Martin Hemberg, Seung Kim, Paul Klenerman, Arnold Kriegstein, Ed Lein, Sten Linnarsson, Emma Lundberg, Joakim Lundberg, Partha Majumder, John C. Marioni, Miriam Merad, Musa Mhlanga, Martijn Nawijn, Mihai Netea, Garry Nolan, Dana Pe’er, Anthony

- Phillipakis, Chris P. Ponting, Stephen Quake, Wolf Reik, Orit Rozenblatt-Rosen, Joshua Sanes, Rahul Satija, Ton N. Schumacher, Alex Shalek, Ehud Shapiro, Padmanee Sharma, Jay W. Shin, Oliver Stegle, Michael Stratton, Michael J.T. Stubbington, Fabian J. Theis, Matthias Uhlen, Alexander Van Oudenaarden, Allon Wagner, Fiona Watt, Jonathan Weissman, Barbara Wold, Ramnik Xavier, and Nir Yosef. The human cell atlas. *eLife*, 6, 2017.
- [14] Simone Picelli, Omid R. Faridani, Åsa K. Björklund, Gösta Winberg, Sven Sagasser, and Rickard Sandberg. Full-length rna-seq from single cells using smart-seq2. *Nature Protocols*, 9, 2014.
- [15] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki, Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161, 2015.
- [16] Grace X.Y. Zheng, Jessica M. Terry, Phillip Belgrader, Paul Ryvkin, Zachary W. Bent, Ryan Wilson, Solongo B. Ziraldo, Tobias D. Wheeler, Geoff P. McDermott, Junjie Zhu, Mark T. Gregory, Joe Shuga, Luz Montesclaros, Jason G. Underwood, Donald A. Masquelier, Stefanie Y. Nishimura, Michael Schnall-Levin, Paul W. Wyatt, Christopher M. Hindson, Rajiv Bharadwaj, Alexander Wong, Kevin D. Ness, Lan W. Beppu, H. Joachim Deeg, Christopher McFarland, Keith R. Loeb, William J. Valente, Nolan G. Ericson, Emily A. Stevens, Jerald P. Radich, Tarjei S. Mikkelsen, Benjamin J. Hindson, and Jason H. Bieleas. Massively parallel digital transcriptional profiling of single cells. *Nature Communications*, 8, 2017.
- [17] Christoph Hafemeister and Rahul Satija. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome Biology*, 20, 2019.
- [18] Samuel L. Wolock, Romain Lopez, and Allon M. Klein. Scrublet: Computational identification of cell doublets in single-cell transcriptomic data. *Cell Systems*, 8, 2019.
- [19] Etienne Becht, Leland McInnes, John Healy, Charles Antoine Dutertre, Immanuel W.H. Kwok, Lai Guan Ng, Florent Gehroux, and Evan W. Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, 37, 2019.
- [20] V. A. Traag, L. Waltman, and N. J. van Eck. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), March 2019.
- [21] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K. Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter

- Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods*, 14, 2017.
- [22] Jason D. Buenrostro, Beijing Wu, Ulrike M. Litzenburger, Dave Ruff, Michael L. Gonzales, Michael P. Snyder, Howard Y. Chang, and William J. Greenleaf. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature*, 523, 2015.
 - [23] Robert R. Stickels, Evan Murray, Pawan Kumar, Jilong Li, Jamie L. Marshall, Daniela J. Di Bella, Paola Arlotta, Evan Z. Macosko, and Fei Chen. Highly sensitive spatial transcriptomics at near-cellular resolution with slide-seq2. *Nature Biotechnology*, 39, 2021.
 - [24] Barry Smith, Michael Ashburner, Cornelius Rosse, Jonathan Bard, William Bug, Werner Ceusters, Louis J. Goldberg, Karen Eilbeck, Amelia Ireland, Christopher J. Mungall, Neocles Leontis, Philippe Rocca-Serra, Alan Ruttenberg, Susanna Assunta Sansone, Richard H. Scheuermann, Nigam Shah, Patricia L. Whetzel, and Suzanna Lewis. The obo foundry: Coordinated evolution of ontologies to support biomedical data integration, 2007.
 - [25] Minoru Kanehisa and Susumu Goto. Kegg: Kyoto encyclopedia of genes and genomes, 2000.
 - [26] Antonio Fabregat, Steven Jupe, Lisa Matthews, Konstantinos Sidiropoulos, Marc Gillespie, Phani Garapati, Robin Haw, Bijay Jassal, Florian Korninger, Bruce May, Marija Milacic, Corina Duenas Roca, Karen Rothfels, Cristoffer Sevilla, Veronica Shamovsky, Solomon Shorser, Thawfeek Varusai, Guilherme Viteri, Joel Weiser, Guanming Wu, Lincoln Stein, Henning Hermjakob, and Peter D'Eustachio. The reactome pathway knowledgebase. *Nucleic Acids Research*, 46, 2018.
 - [27] Arthur Liberzon, Aravind Subramanian, Reid Pinchback, Helga Thorvaldsdóttir, Pablo Tamayo, and Jill P. Mesirov. Molecular signatures database (msigdb) 3.0. *Bioinformatics*, 27, 2011.
 - [28] Yann Lecun, Yoshua Bengio, and Geoffrey Hinton. Deep learning, 2015.
 - [29] Yann Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization, 2014.
 - [30] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22, 1951.
 - [31] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4, 1964.

- [32] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $o(1/k^2)$. *Soviet Mathematics Doklady*, 27, 1983.
- [33] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011.
- [34] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5 - rmsprop: Divide the gradient by a running average of its recent magnitude. Neural Networks for Machine Learning, Coursera Online Class, 2012. Available at https://www.cs.toronto.edu/~tijmen/csc321/slides/lecture_slides_lec6.pdf.
- [35] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.
- [36] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15, 2014.
- [37] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *32nd International Conference on Machine Learning, ICML 2015*, volume 1, 2015.
- [38] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization, 2016.
- [39] Yuxin Wu and Kaiming He. Group normalization. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 11217 LNCS, 2018.
- [40] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts, 2017.
- [41] Leslie N. Smith. Cyclical learning rates for training neural networks, 2017.
- [42] Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima, 2017.
- [43] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2, 1989.
- [44] Vinod Nair and Geoffrey E. Hinton. Rectified linear units improve restricted boltzmann machines. In *ICML 2010 - Proceedings, 27th International Conference on Machine Learning*, 2010.
- [45] Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2023.

- [46] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65, 1958.
- [47] Marvin Minsky and Seymour Papert. *Perceptrons*. MIT Press, Cambridge, MA, 1969.
- [48] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323, 1986.
- [49] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36, 1980.
- [50] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60, 2017.
- [51] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.
- [52] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2016-December, 2016.
- [53] Y. Bengio, P. Simard, and P. Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.
- [54] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, November 1997.
- [55] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.
- [56] Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks, 2017.
- [57] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks, 2018.
- [58] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate, 2016.
- [59] Ruibin Xiong, Yunchang Yang, Di He, Kai Zheng, Shuxin Zheng, Chen Xing, Huishuai Zhang, Yanyan Lan, Liwei Wang, and Tie-Yan Liu. On layer normalization in the transformer architecture, 2020.

- [60] Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are rnns: Fast autoregressive transformers with linear attention, 2020.
- [61] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019.
- [62] Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scbert as a large-scale pretrained deep language model for cell type annotation of single-cell rna-seq data. *Nature Machine Intelligence*, 4, 2022.
- [63] Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang. Cellplm: Pre-training of cell language model beyond single cells. In *12th International Conference on Learning Representations, ICLR 2024*, 2024.
- [64] Suyuan Zhao, Jiahuan Zhang, Yushuai Wu, Yizhen Luo, and Zaiqing Nie. Langcell: Language-cell pre-training for cell identity understanding, 2024.
- [65] Xinyu Yuan, Zhihao Zhan, Zuobai Zhang, Manqi Zhou, Jianan Zhao, Boyu Han, Yue Li, and Jian Tang. Cell ontology guided transcriptome foundation model. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 6323–6366. Curran Associates, Inc., 2024.
- [66] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL HLT 2018 - 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, volume 1, 2018.
- [67] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Glue: A multi-task benchmark and analysis platform for natural language understanding. In *EMNLP 2018 - 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, Proceedings of the 1st Workshop*, 2018.
- [68] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, 2016.
- [69] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. Xlnet: Generalized autoregressive pretraining for language understanding, 2020.

- [70] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21, 2020.
- [71] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kam-badur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chat-terji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale,

Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vitor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkan Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damla, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender

A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabisa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Rutu Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojuan Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024.

- [72] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of Machine Learning Research*, volume 139, 2021.
- [73] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022.

- [74] Hadi Abdine, Michail Chatzianastasis, Costas Bouyioukos, and Michalis Vazirgiannis. Prot2text: multimodal protein’s function generation with gnns and transformers. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI’24/IAAI’24/EAAI’24. AAAI Press, 2024.
- [75] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.