

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

TEK-HÜCRE SEVİYELİ ÇOKLU-OMİK VERİ
BÜTÜNLEŞTİRME

MEHMET BURAK KOCA

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMAN: PROF. DR. FATİH ERDOĞAN SEVİLGİN

ŞUBAT 2025

T.C.
GEBZE TEKNİK ÜNİVERSİTESİ
LİSANSÜSTÜ EĞİTİM ENSTİTÜSÜ

TEK-HÜCRE SEVİYELİ ÇOKLU-OMİK VERİ
BÜTÜNLEŞTİRME

MEHMET BURAK KOCA

DOKTORA TEZİ
BİLGİSAYAR MÜHENDİSLİĞİ ANABİLİM DALI

DANIŞMAN: PROF. DR. FATİH ERDOĞAN SEVİLGİN

ŞUBAT 2025

T.R.
GEBZE TECHNICAL UNIVERSITY
GRADUATE SCHOOL

SINGLE-CELL MULTI-OMICS DATA INTEGRATION



MEHMET BURAK KOCA

A THESIS OF DOCTORATE
DEPARTMENT OF COMPUTER ENGINEERING

ADVISOR: PROF. DR. FATİH ERDOĞAN SEVİLGİN

FEBRUARY 2025

DOKTORA JÜRİ ONAY FORMU

GTÜ Lisansüstü Eğitim Enstitüsü Yönetim Kurulunun 03/02/2025 tarih ve 2025/08 sayılı kararıyla oluşturulan jüri tarafından 06/02/2025 tarihinde tez savunma sınavı yapılan Mehmet Burak KOCA'nın tez çalışması Bilgisayar Mühendisliği Anabilim Dalında DOKTORA tezi olarak kabul edilmiştir.

JÜRİ

ÜYE

(TEZ DANIŞMANI) : Prof. Dr. Fatih Erdoğan SEVİLGİN

ÜYE

: Doç. Dr. Mehmet BAYSAN

ÜYE

: Doç. Dr. Yakup GENÇ

ÜYE

: Doç. Dr. Habil KALKAN

ÜYE

: Dr. Öğr. Üyesi İbrahim Çağrı KURT

ONAY

Gebze Teknik Üniversitesi Lisansüstü Eğitim Enstitüsü Yönetim Kurulunun

...../...../..... tarih ve/..... sayılı kararı.

İMZA/MÜHÜR

ÖZET

Tek-hücre veri dizileme, hücresel heterojenliğin anlaşılmasında devrim yaratmış ve karmaşık biyolojik sistemlerin benzeri görülmemiş bir çözünürlükte incelenmesine olanak tanımıştır. Toplu dizilemenin aksine, tek-hücre seviyeli ölçümler farklı hücrelerdeki belirgin özneliliklerin hücresel çeşitlilik, soy ve hastalık mekanizmaları gibi kritik süreçlere etkisi hakkında değerli bilgiler sunmaktadır. Bu özneliliklerin geniş kapsamlı veri kümeleri kullanılarak karşılaştırmalı analizini kolaylaştırmak için birden fazla deneyden elde edilen tek-hücre verileri bir araya getirilmektedir. Veri kümelerinin doğrudan birlikte kullanılması içerdikleri toplu etki gürültüleri sebebiyle yanlış analizlere neden olmaktadır. Aynı tür omik veri kümelerini birlikte kullanabilmek amacıyla gelişmiş hesaplamalı yatay bütünleştirme yöntemleri kullanılarak ölçümler toplu etkiden arındırılmalıdır. Bunun ötesinde, aynı hücrelerden elde edilen birden fazla modalitenin bir arada kullanılmasıyla da hücresel ayrışma artmakta ve özelleşmiş hücre alt grupları için ayırt edici özneliliklerin tespiti kolaylaşmaktadır. Herhangi bir modaliteye aşırı eğilmeden, birleştirilen farklı omik ölçümlerdeki hücre ayrışmalarını ortaya çıkarılabilmek için yüksek başarımlı dikey ya da çoklu-omik bütünleştirme yöntemlerine ihtiyaç duyulmaktadır.

Yatay ve dikey bütünleştirme için bahsedilen zorlukları ele almak için, bu tez çalışmasında derin öğrenme tabanlı bütünleştirme algoritmaları sunulmaktadır. Önerilen yatay bütünleştirme yöntemi SCPRO-HI, karşılaştırma için kullanılan yöntemlere göre toplu etki gürültüsünü etkili bir şekilde azaltarak siluet skorlarında %75 iyileşme ve biyolojik manada %13 oranında daha iyi korunma sağlamıştır. Dikey bütünleştirme algoritması SCPRO-VI, ARI metriğine göre hücresel heterojenliği %20 artırırken, en yakın komşu hücre eşleştirme performansını %18 oranında iyileştirerek nadir hücre popülasyonlarının tanımlanmasını ve karmaşık çoklu-omik etkileşimlerin anlaşılmasını kolaylaştırmıştır.

Önerilen bütünleştirme algoritmaları, tek-hücre veri kümelerinin yüksek performansla yatay ve dikey bütünleştirilmesi için güçlü bir çerçeve sunmaktadır. Geliştirilen algoritmalar kullanılarak elde edilen verinin ileri analizlerdeki başarımlarını incelemek amacıyla bütünleşik veri üzerinde viral patogenezele ilişkili önemli biyobelirteçler analiz edilmiştir. Hücreler için tanımlanan viral aktivite seviyeleri sayesinde yüksek aktivite gösteren hücre altkümeleri tespit edilerek bu kümelerdeki ayırt edici biyobelirteçlerin öznelilik önem analizi ile belirlenmesi sağlanmıştır. HIV-1 hastalarına ait bütünleştirilmiş veri setlerinde, Granzyme-H gibi bilinen belirteçlerin yanı sıra MIAT geni gibi aday biyobelirteçler de ayırt edici faktörler olarak tespit edilerek viral patogeneze hakkında yeni iç görüler sunulmuştur. Mevcut ve gelecekteki bu bulgular, viral enfeksiyonlara karşı hücresel tepkilerin daha derin anlaşılmasına katkı sağlayarak yeni terapötik stratejilerin geliştirilmesine ve hastalıkların erken tespitine olanak sunmaktadır.

Anahtar Kelimeler: Tek-hücre Analizi, Toplu Etki Düzeltme, Yatay Veri Bütünleştirme, Çoklu-omik Veri Analizi, Çok Modaliteli Veri Bütünleştirme, Viral Patogeneze Analizi

ABSTRACT

Single-cell sequencing has revolutionized the understanding of cellular heterogeneity and enabled the study of complex biological systems at an unprecedented resolution. Unlike bulk sequencing, single-cell measurements provide valuable insights into how distinct cellular features influence critical processes such as cellular diversity, lineage, and disease mechanisms. Single-cell data obtained from multiple experiments are integrated to facilitate the comparative analyses of these features using large-scale datasets. However, the direct use of datasets together can lead to inaccurate analyses due to batch effect noise. Therefore, to enable the joint use of omic datasets of the same type, measurements must be cleaned for batch effects using advanced computational horizontal integration methods. Beyond this, the simultaneous integration of multiple modalities obtained from the same cells enhances cellular separation and facilitates the identification of distinguishing features for specialized cell subgroups. High-performance vertical or multi-omic integration methods are needed to reveal cellular separations in integrated multi-omic measurements without overfitting any single modality.

This thesis presents deep learning-based integration algorithms to address the challenges of horizontal and vertical integration. The proposed horizontal integration method SCPRO-HI effectively reduced batch effect noise, improving silhouette scores by 75% and preserving biological meaning 13% better than the compared methods. The vertical integration algorithm SCPRO-VI increased cellular heterogeneity by 20% according to the ARI metric while improving nearest-neighbor cell matching performance by 18%, facilitating the identification of rare cell populations and the understanding of complex multi-omic interactions.

The proposed integration algorithms provide a robust framework for high-performance horizontal and vertical integration of single-cell datasets. Key biomarkers associated with viral pathogenesis were analyzed to evaluate the effectiveness of the integrated data in downstream analyses. Highly active cell subgroups were identified by viral activity levels defined for cells, and distinguishing biomarkers within these clusters were determined using feature importance analysis. In integrated datasets from HIV-1 patients, in addition to well-known markers such as Granzyme-H, candidate biomarkers such as MIAT gene were also identified as distinctive factors, offering new insights into viral pathogenesis. These current and future findings contribute to a deeper understanding of cellular responses to viral infections, paving the way for the development of novel therapeutic strategies and early disease detection.

Keywords: Single-cell Analysis, Batch Effect Correction, Horizontal Data Integration, Multi-omic Data Analysis, Multimodal Data Integration, Viral Pathogenesis Analysis

TEŞEKKÜR

Lisansüstü eğitim hayatım boyunca ihtiyaç duyduğum ve ihtiyaç olduğunu dahi sonradan kavradığım tüm konularda zaman zaman zorlayıcı olsa da her zaman isabetli yönlendirmeleri ile akademik tüm başarılarımın mimarı olan danışmanım Sayın Prof. Dr. Fatih Erdoğan Sevilgen'e verdiği tüm emekler ve paha biçilmez akademik önderliği için tüm içtenliğimle teşekkür ederim.

Akademik çalışmalar için zinde ve gündelik dertlerden olabildiğince arındırılmış zihinler gerektiğine inanıyorum. Akademik hayatım boyunca beni bir çok gündelik yaşam görevinden azlederek bana alan açan, bu sayede çalışmalarına odaklanmamı sağlayan sevgili eşime ve zaman zaman azalan motivasyonumu geri kazanmam için ihtiyaç duyduğum itiş gücünü sağlayan canım oğluma çok teşekkür ederim.

Bana inanıp bugüne kadar verdiğim kararlara saygı duyarak kendi yolumu çizmeme ve olduğum insana dönüşmeme imkan sağlayan değerli aileme çok teşekkür ederim.

Son olarak tüm arkadaşlarıma ve beraber çalıştığım meslektaşlarıma zaman zaman kabuğuma çekildiğim yoğun dönemlerde bana katlandıkları ve destekleri için teşekkür ederim.

Türkiye Bilimsel ve Teknolojik Araştırma Kurumuna (TÜBİTAK) 122E587 numaralı ve *Tek-Hücre Seviyeli Proteomik Veri Kümelerinin Yatay Bütünleştirilmesi* başlıklı 1002-A projemi destekledikleri için teşekkür ederim.

İÇİNDEKİLER

	<u>Sayfa</u>
ÖZET	v
ABSTRACT	vi
TEŞEKKÜR	vii
İÇİNDEKİLER	viii
SİMGELER ve KISALTMALAR DİZİNİ	xi
ŞEKİLLER DİZİNİ	xii
ÇİZELGELER DİZİNİ	xv
1. TEZ KONUSU ve ARKAPLANI	1
1.1. Toplu Dizileme ve Tek-hücre Dizileme	1
1.2. Tek-hücre Dizileme Teknolojileri	2
1.3. Tek-hücre Omik Veri Bütünleştirme Yöntemleri	3
1.4. Viral Patogenezin Tek-hücre Omik Veri Kümeleri ile Analizi	6
1.5. Tezin Katkısı	9
2. YATAY VERİ BÜTÜNLEŞTİRME	10
2.1. Giriş	10
2.2. Yöntem	14
2.2.1. Hücre Eşleştirme	16
2.2.2. Toplu Etki Düzeltme	19
2.2.3. Yüksek Boyutlu Verilerle Çalışma	21
2.2.4. Karşılaştırmada Kullanılan Algoritmalar	22
2.2.4.1. MARIO	22
2.2.4.2. Harmony	23
2.2.4.3. Scanorama	23
2.2.4.4. FastMNN	23
2.2.4.5. ScVI	24
2.2.5. Metrikler	24
2.2.5.1. Ortalama Siluet Genişliği (ASW)	24
2.2.5.2. Düzeltilmiş Rand İndeksi (ARI)	25
2.2.5.3. iLISI ve cLISI	25
2.2.5.4. Yüksek Değişken Protein (HVP) Skoru	26
2.3. Sonuçlar	26
2.3.1. Sentetik Veri Setlerinin Bütünleştirilmesi	26
2.3.2. Türler Arası Veri Setlerinin Bütünleştirilmesi	27
2.3.3. İnsan PBMC Veri Setlerinin Bütünleştirilmesi	31
2.3.4. Yüksek Boyutlu Veri Setlerinin Bütünleştirilmesi	33
2.4. Sonuç	36

2.5. Veri ve Kod Erişilebilirliği	38
3. ÇOKLU-OMİK VERİ BÜTÜNLEŞTİRME	39
3.1. Giriş	39
3.2. Yöntem	42
3.2.1. SCPRO-VI	42
3.2.1.1. Uzaklık Matrislerinin Hesaplanması	43
3.2.1.2. Modalitelerin Gömülmesi	45
3.2.1.3. Çoklu-omik Veri Bütünleştirme	48
3.2.1.4. Ortam ve Model Kurulumu	50
3.2.2. Karşılaştırmada Kullanılan Algoritmalar	51
3.2.2.1. Seurat v4	51
3.2.2.2. TotalVI	51
3.2.2.3. Mofa+	51
3.2.2.4. Mowgli	51
3.2.3. Metrikler	52
3.2.3.1. Hücre Eşleşme Skoru	52
3.2.3.2. Normalleştirilmiş Karşılıklı Bilgi (NMI)	52
3.2.3.3. Çizge Bağlantısı	53
3.2.3.4. F1 Skoru	53
3.2.4. Veri Analizi	54
3.2.4.1. Ayırt Edici Öznelik Analizi	54
3.2.4.2. Hücre Alt Kümesi Tanımlama	54
3.2.4.3. Gen Ontolojisi Analizi	55
3.3. Sonuçlar	56
3.3.1. Deneylerde Kullanılan Veri Kümeleri	56
3.3.2. SCPRO-VI'nin Dayanıklılığının Doğrulanması	58
3.3.3. HIV Hastalarında Aşılama Sonrası Değişimlerin İncelenmesi	62
3.3.4. Çoklu-Örnek Veri Kümelerinde Dikey Bütünleştirme	65
3.3.5. Yüksek Performanslı Algoritmalar ile Karşılaştırma	68
3.3.6. Özgün Hücre Alt Kümelerinin Analizi	74
3.4. Sonuç	77
3.5. Veri ve Kod Erişilebilirliği	79
4. VİRAL PATOGENEZ ANALİZİ	81
4.1. Giriş	81
4.1.1. Benzer Çalışmalar	82
4.2. Yöntem	83
4.2.1. Hücre Viral Aktivite Seviyelerinin Belirlenmesi	83
4.2.2. Viral Patogenezde Aktif Hücre Altkümelerinin Tespiti	85
4.2.3. Hücre Altkümelerinde Biyobelirteç Analizi	86
4.3. Sonuçlar	87
4.3.1. Deneylerde Kullanılan Veri Kümesi	87
4.3.2. Hücrelerin Viral Aktivite Seviyelerinin Belirlenmesi	87
4.3.3. Viral Aktivite Seviyelerinin Analizi	88
4.3.4. Hedef Proteinler Üzerinde Viral Patogenez Analizi	90
4.3.5. Viral Patogenezde Aktif Hücre Altkümelerinin İncelenmesi	94
4.3.6. Viral Patogenezde Aktif Altkümelerde Biyobelirteç Analizi	95

4.4. Sonuç	97
5. SONUÇ	99
KAYNAKLAR	102
ÖZGEÇMİŞ	119
TEZ ÇALIŞMASI KAPSAMINDA YAPILAN YAYINLAR	120
EKLER	121



SİMGELER ve KISALTMALAR DİZİNİ

ASW	: Average Silhouette Score (Ortalama Siluet Skoru)
ARI	: Adjusted Rand Index (Düzeltilmiş Rand İndeksi)
CCA	: Canonical Correlation Analysis (Kanonik Korelasyon Analizi)
DC	: Dendritic Cell (Dendritik Hücre)
GEO	: Gene Expression Omnibus Veritabanı
GTÜ	: Gebze Teknik Üniversitesi
HCA	: Human Cell Atlas (İnsan Hücresi Atlası)
HVG	: Highly Variable Gene (Yüksek Değişken Gen)
HVP	: Highly Variable Protein (Yüksek Değişken Protein)
KL	: Kullback-Leibler Divergence (Kullback-Leibler Mesafesi)
MLP	: Multi-layer Perceptron (Çok katmanlı algılayıcı)
MNC	: Mononuclear Cell (Mononükleer Hücre)
NGS	: Next Generation Sequencing (Yeni Nesil Dizileme)
NMF	: Negatif Olmayan Matris Faktörizasyonu
NMI	: Normalized Mutual Information (Normalleştirilmiş Karşılıklı Bilgi)
NK	: Natural Killer (Doğal Öldürücü)
PCA	: Principal Component Analysis (Temel Bileşen Analizi)
PBMC	: Periferik Kan Mononükleer Hücre
PPI	: Protein - Protein Interaction (Protein - Protein Etkileşimi)
SCP	: Single-cell Proteomics (Tek-hücre Proteomik)
SCPRO-HI	: Single-cell Proteomics Horizontal Integration
SCPRO-VI	: Single-cell Proteomics Vertical Integration
SS	: Silhouette Score (Siluet Skoru)
VAE	: Variational Autoencoder (Varyasyonel Oto-kodlayıcı)
VGAE	: Variational Graph Autoencoder (Varyasyonel Çizge Oto-kodlayıcı)

ŞEKİLLER DİZİNİ

	Sayfa
Şekil 1.1: Tek-hücre dizileme ve toplu dizileme yöntemlerinin karşılaştırmalı şematiği.	2
Şekil 1.2: Veri bütünleştirme yöntemlerinin şematik gösterimi.	4
Şekil 1.3: Tez adımlarının şematik gösterimi. A Proteomik veri kümelerini yatay bütünleştirme. B Çoklu-omik veri kümelerini dikey bütünleştirme. C Hücrelerin viral aktivite seviyelerinin hesaplanması. D Viral patogeneze etkin biyobelirteç analizi.	7
Şekil 2.1: SCPRO-HI algoritmasının bütünleştirme iş akış şematiği. A Verisetlerinin kümelenmesi. B Farklı veri setlerindeki kümelerin çapalanarak bağlı bileşenleri oluşturması. C Hücre çapalarının oluşturulması. D Özgün VAE modelleri kullanılarak toplu etkinin düzeltilmesi. E Düzeltilmiş bolluk vektörlerinin hesaplanması.	15
Şekil 2.2: Toplu etki düzeltme için kullanılan varyasyonel oto-kodlayıcı model mimarisi.	20
Şekil 2.3: Yüksek boyutlu veri kümeleri için düzeltilmiş ölçümler üretmeyi sağlayan yöntemin iş akış şematiği.	22
Şekil 2.4: Sentetik veri setleri ve bütünleştirme sonuçları. A Veri Seti 1 UMAP gösterimi. B Veri Seti 2 UMAP gösterimi. C Bütünleştirme öncesi her iki veri setinin birleştirilmiş UMAP gösterimi. D Bütünleştirilmiş verinin UMAP gösterimi.	27
Şekil 2.5: Tek-hücre proteomik veri setlerinde algoritmaların bütünleştirme sonuçları. A Algoritmaların bütünleştirilmiş ölçümlerinin UMAP gösterimleri. B Proteinlere ait yüksek değişkenlik benzerliklerini gösteren ısı haritaları.	29
Şekil 2.6: Tek-hücre insan PBMC proteomik veri setlerinin bütünleştirme sonuçları.	32
Şekil 3.1: A Modaliteye özgü uzaklık matrisi hesaplaması ve bütünleştirilmiş uzaklıkların oluşturulmasına dair şematik gösterimi. B Çoklu-görünüm varyasyonel çizge oto-kodlayıcılar kullanılarak bütünleşik hücre temsilleri oluşturma şematiği.	42
Şekil 3.2: Modaliteleri gizli uzaya gömmek için kullanılan varyasyonel çizge oto-kodlayıcı model mimarisi.	46
Şekil 3.3: Modaliteleri bütünleştirmede kullanılan çoklu-görünüm varyasyonel çizge oto-kodlayıcı model mimarisi.	49
Şekil 3.4: Önemli ve önemsiz proteinlerin GO analizi özetleri.	58
Şekil 3.5: Bütünleştirme algoritmalarının karşılaştırma metrikleri performanslarının ısı haritası olarak gösterimi.	60
Şekil 3.6: 8 hastadan 3 farklı zaman diliminde elde edilen ölçümlerin bütünleşik sonuçlarının UMAP gösterimi.	61

Şekil 3.7:	Geliştirilen algoritmanın HIV veri seti üzerindeki etkinlik testi sonuçları. A Hasta 1'in ham ve bütünleştirilmiş ölçümlerinin UMAP gösterimleri. B SCPRO-VI gömmelerinin FCGR3A proteini (üst) ve LYZ geni (alt) ifade seviyelerine göre renklendirilmiş UMAP gösterimleri. C SCPRO-VI gömmeleri (üst) ve ham veriler (alt) kullanılarak tüm HIV veri setindeki hücrelerin gerçek türleri (sol) ile en yakın komşuları (sağ) arasındaki eşleşmelerin Sankey diyagramları. D HIV veri setinden örneklenen veri kümesine ait UMAP gösterimleri (üst) ve nümerik performans sonuçları (alt).	63
Şekil 3.8:	MNC veri seti üzerinde algoritmalara ait performans kıyaslama sonuçları. A Algoritma bütünleştirmelerinin hücre tipine göre renklendirilmiş UMAP gösterimleri. B Algoritmaların nümerik karşılaştırma metriklerindeki performans sonuçları. C Artan komşuluk boyutlarında algoritmaların saflık analizi. D Transkriptomik, proteomik ve SCPRO-VI bütünleştirmeleri ile hücre eşleştirme doğruluk sayılarını gösteren küme diyagramı.	69
Şekil 3.9:	Algoritmaların bütünleşik ölçümlerinin örnek etiketleri ile renklendirilmiş UMAP gösterimleri.	71
Şekil 3.10:	Bütünleştirilmiş MNC verilerinde ortaya çıkan hücre alt kümelerine ait analiz sonuçları. A Algoritmaların kendi optimal kümeleriyle renklendirilmiş bütünleştirilmiş gömmelerinin UMAP gösterimleri. B CLP (üst) ve granülosit (alt) hücre alt kümeleri için önemli proteinlerin (P) ve genlerin (F) değerlerinin Stem diyagramı. C SCPRO-VI'nin bütünleştirilmiş gömmelerinin CLP (üst) ve granülosit (alt) hücre alt kümeleri için en ayırt edici proteinin (sol) ve genin (sağ) ifade seviyelerine göre renklendirilmiş UMAP gösterimleri.	75
Şekil 4.1:	Viral aktivite sınıf dağılım diyagramı.	84
Şekil 4.2:	Proteinlere ait z-skor dağılım grafiği.	88
Şekil 4.3:	Bütünleştirilmiş verinin aktivite sınıfı ile renklendirilmiş UMAP gösterimi.	89
Şekil 4.4:	Hücre tiplerine göre aktivite sınıfı dağılımlarının pasta grafik olarak gösterimi.	90
Şekil 4.5:	Hedef proteinlerin aktivite sınıflarına göre beklenen ifade seviyesi değişimini karşılama oranlarının ısı haritası.	91
Şekil 4.6:	Hedef proteinlerin hücre tiplerine göre beklenen ifade seviyesi değişimini karşılama oranlarının ısı haritası.	92
Şekil 4.7:	Hedef proteinlerin örnek dizileme zamanına göre beklenen ifade seviyesi değişimini karşılama oranlarının ısı haritası.	93
Şekil 4.8:	Bütünleşik verinin gözetimsiz kümeleme etiketleriyle renklendirilmiş UMAP gösterimi.	94
Şekil 4.9:	Bütünleşik verinin sitotoksik CD4 T hücre altkümesini diğer CD4 T hücrelerinden ayırmada en etkili 5 genin ifade seviyeleri ile renklendirilmiş UMAP gösterimleri.	95
Şekil 4.10:	CD4 T hücre altkümesini diğer CD4 T hücrelerinden ayırmada en etkili genler ile gen ontoloji zenginleştirme analiz sonuçları.	96

Şekil 4.11: Bütünleşik verinin sitotoksik CD4 T hücre altkümesini diğer 97 tüm hücrelerden ayırmada en etkili 5 genin ifade seviyeleri ile renklendirilmiş UMAP gösterimleri.



ÇİZELGELER DİZİNİ

	<u>Sayfa</u>
Çizelge 2.1: Türler arası veri setlerinin bütünleştirme performans sonuçları.	30
Çizelge 2.2: PBMC veri setlerinin bütünleştirme performans sonuçları.	33
Çizelge 2.3: Transkriptomik veri setlerinin bütünleştirme performans sonuçları.	35
Çizelge 2.4: Kütle spektrometrisi tabanlı proteomik veri setlerinin bütünleştirme performans sonuçları.	36
Çizelge 3.1: Biyolojik bilginin hücre eşleştirme başarımına etki sonuçları.	44
Çizelge 3.2: HIV veri kümesinin ve örneklenmiş veri kümesinin tiplerine göre hücre sayısı dağılımı.	56
Çizelge 3.3: MNC veri kümesinin ve örneklenmiş veri kümesinin tiplerine göre hücre sayısı dağılımı.	57
Çizelge 3.4: Hücre tipi içi ayrışmaya katkıda bulunan en ayırt edici proteinler ve genler.	64
Çizelge 3.5: HIV verisetinin bütünleşik ölçümleri ile tip bazında hücre eşleştirme başarımları.	66
Çizelge 3.6: HIV verisetinin ham ölçümleri ile tip bazında hücre eşleştirme başarımları.	67
Çizelge 3.7: Modalitelerin ve bütünleşik ölçümlerin hücreleri doğru eşleştirme sayıları.	73

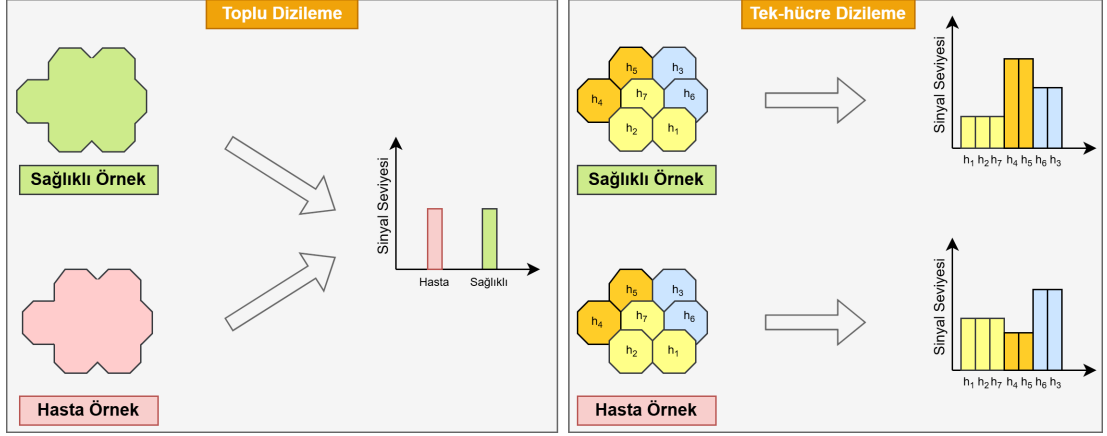
1. TEZ KONUSU ve ARKAPLANI

1.1. Toplu Dizileme ve Tek-hücre Dizileme

Tek hücre çalışmaları mikroskobun icadıyla başlayan köklü bir geçmişe sahip olmasına rağmen, son elli yılda biyolojik sinyallerin toplu biçimde dizilenmesine imkân veren Sanger dizileme [1] ve sonrasında yeni nesil dizileme (ing. Next Generation Sequencing - NGS) teknolojilerinin [2, 3] gelişimi sebebiyle gölgede kalmıştır. Araştırmacılar geçtiğimiz yüzyılın ortalarına kadar dokulardan izole ettikleri tek hücreler üzerinde mikroskop yardımıyla çeşitli deneyler gerçekleştirerek elde ettikleri gözlemleri analiz etmiştir. Doğası gereği bu çalışmalar uzun zaman almaktadır ve hem kullanılan araç gereç hem de insan gücü açısından maliyetlidir [4]. Bu nedenle, bilim dünyası toplu dizileme yöntemlerine yönelmiş ve tek hücre seviyeli analizler ikinci planda kalmıştır.

Toplu dizileme yöntemleri hedef sinyalin (omik verinin) dokuda bulunan tüm hücrelerdeki miktarını toplu olarak ölçmektedir. NGS gibi yöntemler ile elde edilen bu büyük hacimdeki değerli veriler canlılara ait biyolojik süreçleri anlama, analiz etme ve manipüle etmede kullanılmaktadır [5]. Öte yandan son yıllarda geliştirilen tek-hücre dizileme teknolojileri [6, 7] sayesinde hedef dokudaki her bir hücreden biyolojik sinyalleri ayrı ayrı elde etmek mümkün hale gelmiştir. Tek-hücre dizileme ile elde edilen veri, hedef dokuyu oluşturan hücreler arasındaki heterojen yapıyı ortaya çıkarabileceği için oldukça değerlidir. Bununla birlikte tek-hücre dizileme yöntemi ile elde edilen veri geleneksel yöntemlere göre daha pahalı ve daha gürültülü olduğu için homojen yapıdaki dokularda toplu dizileme yöntemi hala kullanılmaktadır [7] .

Dokulardaki heterojen yapı temelde birbirinden ayrılan farklı hücre tipleri, farklı görevler için özelleşmiş hücre altkümeleri veya kansere dönüşen dokular sebebiyle ortaya çıkmaktadır. Yapılan çalışmalar [8, 9] göstermektedir ki toplu dizileme verilerinin analizi ile saptanması mümkün olmayan ve çeşitli hastalıklar ile doğrudan ilişkili hücre alt kümeleri tek-hücre verilerinin analizi ile ortaya çıkabilmektedir. Şekil 1.1’de verilen örnekte rastgele bir *A* genine ait seviyeler toplu dizileme ile incelendiğinde bu genin sağlıklı ve hasta bireylerde benzer ifade seviyelerine sahip olduğu görülmektedir. Öte yandan, aynı gene ait ifade seviyeleri tek-hücre seviyesinde incelendiğinde



Şekil 1.1: Tek-hücre dizileme ve toplu dizileme yöntemlerinin karşılaştırmalı şematığı.

hasta örnekte turuncu renkle gösterilen hücre tipi hariç tüm hücre tiplerinde artış gösterdiği görülmektedir. Bununla beraber, sağlıklı örnekten elde edilen ölçümler A geninin turuncu hücre tiplerinde yüksek ifade seviyesine sahip olması gerektiğini göstermektedir. Heterojen yapının bir başka sebebi ise hücrelerin yaşam döngüsünden kaynaklanmaktadır. Hücreler yaşam döngüleri boyunca farklı görevler üstlenmektedir ve bu sebeple ürettikleri biyolojik sinyallerin miktarı da değişmektedir. Hücre doğası kaynaklı biyolojik sinyal üretimi sırasında ortaya çıkan stokastik gürültüler de heterojen hücre dağılımına sebep olmaktadır [10].

1.2. Tek-hücre Dizileme Teknolojileri

Tek-hücre seviyeli ölçüm yöntemlerinin yüksek başarımla dizileme yapabilmesi için kısıtlı alanda bulunan düşük miktardaki sinyali ölçme hassasiyeti ve farklı teknolojiler ile elde edilen ölçümler için genel veri doğrulama araçlarının bulunmaması gibi problemlerin üstesinden gelmelidir. Bununla beraber son on yılda birçok omik veri tipi için tek-hücre seviyesinde dizileme yöntemi geliştirilmiştir [8]. Toplu dizilemede olduğu gibi en yaygın kullanılan omik veri tek-hücre gen ifade seviyelerini içeren transkriptomik veridir (ing. single-cell RNA sequencing - scRNA-seq). Diğer biyolojik sinyallere göre dizilenmesi daha kolay olan transkriptomik verinin tek-hücre seviyesinde üretimi için birçok yöntem geliştirilmiştir [11] ve önemli bir hacimde veri açık veri tabanlarında sunulmaktadır. Gen ifade seviyelerinin biyolojik neden-sonuç ilişkilerinin gözlemlenmesinde öncül işaretçi olması sebebiyle birçok tek-hücre transkriptomik veri analiz yöntemi geliştirilmiştir ve bu alan gün geçtikçe daha olgun hale gelmektedir.

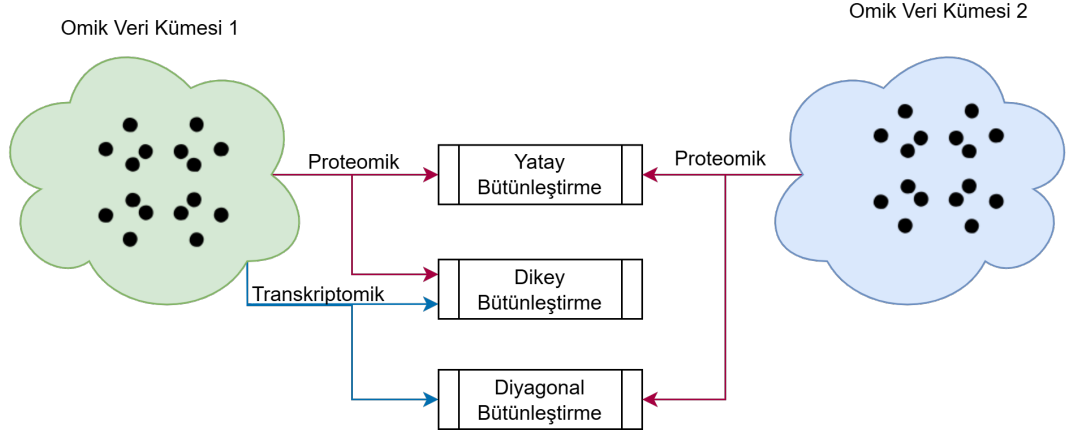
Gen seviyesi haricindeki biyolojik sinyallerin tek-hücre seviyesinde üretilmesi oldukça popüler bir çalışma alanıdır ve her geçen gün yeni yöntemler ortaya çıkmaktadır [8]. Tek-hücre seviyeli Transpozaz-Erişimli Kromatin Dizinlemesi (ing. single-cell Assay for Transposase-Accessible Chromatin - scAtac-seq) [12] yöntemi sayesinde tek-hücre seviyesli DNA metilasyonlarının gen seviyelerine etkisi analiz edilmektedir. Geliştirilen diğer yöntemler sayesinde DNA metilasyon seviyeleri [13], histon değişimleri [14] veya kromatin yapıları da [15] tek-hücre seviyesinde elde edilebilmektedir.

Hücrelerdeki protein mevcudiyetlerini de tek-hücre seviyesinde ölçülebilen çeşitli teknolojiler geliştirilmektedir [16]. Hücre içi ve hücreler arası yaşamsal faaliyetlerde proteinlerin üstlendiği görevler göz önüne alındığında tek-hücre proteomik (ing. single-cell proteomic - SCP) veriler oldukça değer kazanmaktadır. SCP veri üreten bazı teknolojiler antikor tabanlı yeni nesil dizileme yöntemi ile protein ölçümlerini elde etmektedir [17, 18]. Bu yöntem oldukça doğru dizileme sonuçları üretmekle beraber antikor kütüphanelerinin el ile hazırlanması sebebiyle ölçülebilen protein sayısı birkaç yüzü geçmemektedir. Öte yandan kütle spektrometresi kullanarak tek seferde binlerce proteine ait tek-hücre seviyesinde proteomik veri üretebilen teknolojiler mevcuttur [19, 20]. Antikor tabanlı yaklaşıma göre veri doğruluğu daha düşük olsa da binlerce proteine ait verinin aynı anda üretilmesi açısından oldukça değerlidir [21].

Hücrelerden tek seferde birden fazla biyolojik sinyalin dizilenmesi ile çoklu-omik veri kümeleri elde edilmektedir. Bu veri kümeleri genellikle hücrede bulunan gen ifade seviyelerine ek olarak başka bir veya daha fazla tipte omik veriyi içermektedir [22, 23]. Mevcut tek-hücre dizileme teknolojileri ile tek seferde transkriptomik veriye ek olarak epigenomik veri (scMT-seq, SNARE-seq) [24, 25], genomik veri (TARGET-seq) [26], proteomik veri (CITE-seq, ECCITE-seq, REAP-seq) [17, 27, 28] ve dizilenen transkriptomik verinin doku üzerindeki konum bilgisi (MERFISH) [29] üretilmektedir. Hücrelerden aynı anda elde edilen birden fazla omik veri farklı biyolojik sinyaller arasındaki ilişki örüntülerini analiz etme imkanı sunmaktadır [30].

1.3. Tek-hücre Omik Veri Bütünleştirme Yöntemleri

Dokulardaki heterojen yapıları ve bu yapıların sebep olduğu biyolojik neden-sonuç ilişkilerini karşılaştırmalı gözlemlemek amacıyla, genetik veya çevresel faktörlerle



Şekil 1.2: Veri bütünleştirme yöntemlerinin şematik gösterimi.

yapılan değişiklikler ile hücrelerin belirli biyolojik süreçlere tepki vermesini sağlayan hücre pertürbasyon yöntemleri [31] kullanılmaktadır. Diğer yandan doğal süreçler vasıtasıyla veya bir hastalık sebebiyle ortaya çıkan heterojen yapıları ve bunların sebep olduğu biyolojik sonuçları gözlemlemek için farklı kişi veya dokulardan elde edilen tek-hücre verilerine ihtiyaç duyulmaktadır [32–34].

Tek-hücre seviyeli omik veri kümelerinin bütünleştirilmesi problemi Biyoinformatik alanında son yıllarda popülerliği gittikçe artan bir çalışma alanıdır [35]. Birden fazla tek-hücre seviyeli omik veri kümesi bütünleştirilerek hücre tipi sınıflandırması [36], eşsiz hücre gruplarının tespiti [37], hücre gidişlerinin ortaya çıkarılması (ing. trajectory inference) [38], hücresel yaşlanmanın tespiti [39] ve doku üzerindeki uzamsal biyolojik yolların tespiti [40] gibi birçok probleme çözüm aranmaktadır. Yatay, dikey ve diyagonal olmak üzere üç farklı omik veri bütünleştirme türü bulunmaktadır [41]. Bütünleştirilen veri kümelerinin aynı veya farklı tipte bir sinyale ait olması ve bu veri kümelerinin aynı ya da farklı hücrelerden elde edilmesine bağlı olarak bütünleştirme yaklaşımları birbirinden ayrılmaktadır (Şekil 1.2).

Aynı tipte biyolojik sinyalin farklı hücre gruplarından dizilmesi ile elde edilen omik veri kümelerini bütünleştirme işlemine yatay bütünleştirme denmektedir. Birden fazla kişi veya dokudan elde edilen tek-hücre veri kümelerinin doğrudan birlikte kullanılabilmesi biyolojik ve teknik gürültüler sebebi ile mümkün değildir [8, 41]. Teknik gürültü "toplu etki" (ing. "batch effect") olarak isimlendirilmektedir. Toplu etki gürültüleri yaygın olarak hücre nesillerinin üretimi sırasında kullanılan deney

ortamı veya deneyi gerçekleştiren araştırmacıya bağlı olarak stokastik bir biçimde ortaya çıkmaktadır [42]. Bununla beraber kullanılan teknolojilerin ölçümleri farklı derinliklerde gerçekleştiriyor olması da toplu etkiye sebep olmaktadır [43]. Biyolojik gürültüler ise birden fazla dokudan ya da kişiden elde edilen veriler arasında ortaya çıkmaktadır [8]. Bununla birlikte dizilen omik verinin hücrenin tamamından ya da sadece hücre çekirdeğinden elde edilmesi de biyolojik gürültüye sebep olmaktadır. Biyolojik gürültünün miktarı sırasıyla dokular arası farklılığa, dizilemenin yapıldığı hücre bölgesine ve farklı kişilerden veri elde edilmesine bağlı olarak artmaktadır [44]. Yatay bütünleştirmenin temel amacı farklı deneylerden elde edilen aynı tipte omik veri kümelerini hizalayarak bu veri kümeleri üzerindeki toplu etkiyi ortadan kaldırmaktır.

Hücreler birbiri ile etkileşimli birçok biyolojik sinyali barındıran karmaşık sistemlerdir. Aynı hücreden elde edilen birden fazla omik verinin bütünleştirilmesi dikey bütünleştirme veya çoklu-omik bütünleştirme olarak isimlendirilmektedir [41]. Dikey bütünleştirme, yatay bütünleştirmeden farklı olarak çeşitli biyolojik sinyallerin birlikte kullanılması ile hücreler arasındaki heterojen yapının belirgin hale getirilmesini hedeflemektedir. Böylece, hücre farklılaşma süreçlerinin takibi, hücresel alt popülasyonların daha hassas sınıflandırılması ve biyolojik sinyaller arasındaki ilişkilerin modellenmesi gibi analizlerin doğruluğu artırılmaktadır.

Diyagonal veri bütünleştirme yaklaşımında farklı hücre gruplarından elde edilen ve aynı tipte olmayan omik veri tipleri bütünleştirilerek kapsamlı analizler gerçekleştirilmektedir. Bu yöntemler farklı tipteki omik veriler arasındaki ortak değişimleri veya aralarındaki neden-sonuç ilişkilerini ortaya çıkarmayı hedeflemektedir. Diyagonal bütünleştirme verilerin elde edildiği hücre gruplarının aynı ya da benzer tipte olduğu varsayımına dayandığı için bütünleştirilen omik veri setlerinin elde edildiği dokulara dikkat edilmelidir. Bahsedilen bütünleştirme yöntemleri sayesinde biyolojik süreçleri tek-hücre seviyesinde daha doğru ve kapsamlı incelemek mümkün olmakla beraber bütünleştirme yöntemlerinin kendine has biyolojik ve hesaplamalı zorlukları bulunmaktadır [45].

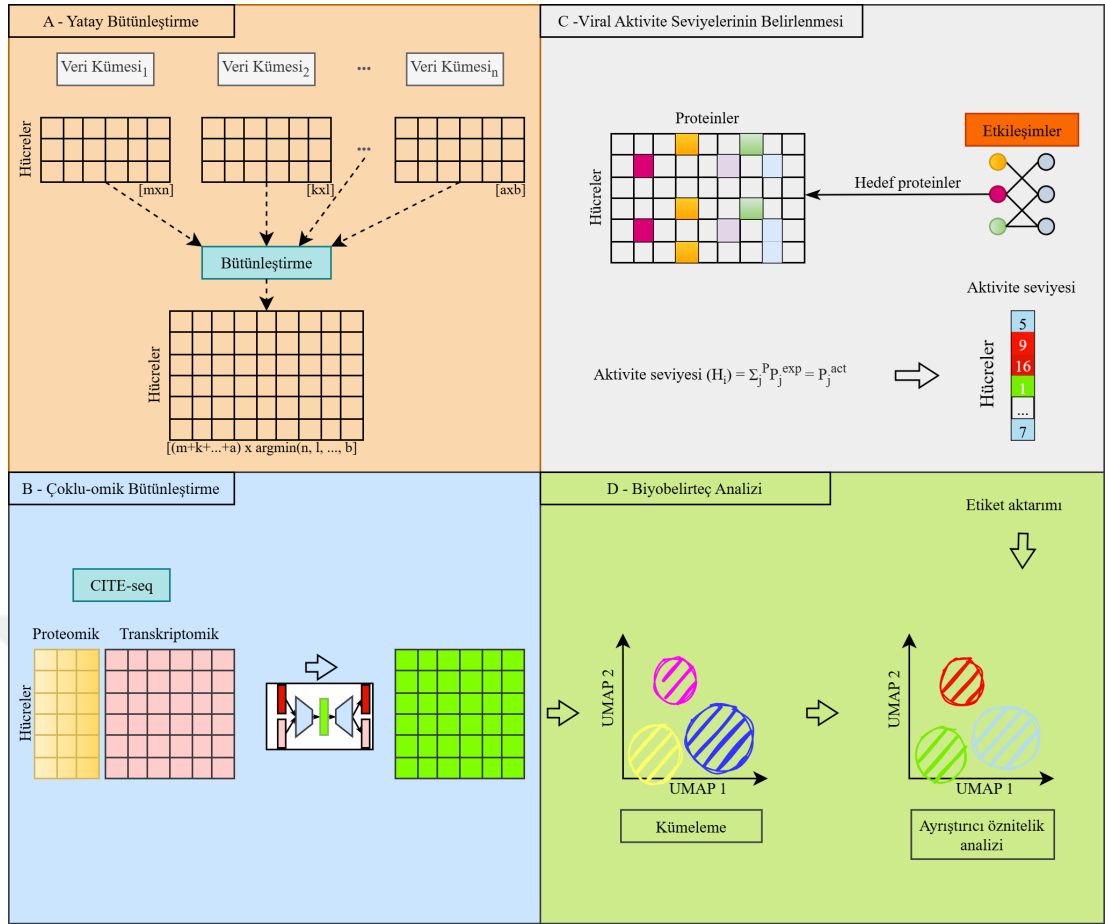
1.4. Viral Patogenezin Tek-hücre Omik Veri Kümeleri ile Analizi

Veri bütünleştirme algoritmalarının yalnızca nümerik performans metrikleri ile değil ileri biyoinformatik analizler ile test edilmesi, bu algoritmaların kullanılabilirliklerini kanıtlamak açısından önemlidir. Geliştirilen yöntemleri test etmek amacıyla tek-hücre çoklu-omik veri kümelerinin bütünleştirilmesi ile elde edilen bütünleşik veride viral patogeneze için önemli hücre altkümeleri tespit edilerek bu altkümeler için ayrıştırıcı biyobelirteç analizi gerçekleştirilmiştir (Şekil 1.3).

Tek-hücre omik veri kümelerini bütünleşik biçimde kullanarak heterojen yapıdaki dokularda patojenlerin hastalık yapıcı mekanizmalarının analizi popüler bir biyoinformatik problemidir [46–48]. Yapılan çalışmalar ile farklı hücre gruplarının patojenlere neden farklı seviyede tepkiler gösterdiği ve hatta bazı hücre gruplarının hiç etkilendiğine cevap aranmaktadır [47]. Patojenlerin konak canlı hücreleri üzerindeki değişken etkilerinin analizi ile hastalıkların önlenmesi ve tedavisine değerli katkılar sağlamak mümkündür.

Bu alanda geliştirilen çalışmalar çoğunlukla enfekte olmuş ve sağlıklı dokular üzerinden elde edilen tek-hücre gen ifade seviyelerinin karşılaştırmalı analizlerine dayanmaktadır [49]. Tek-hücre seviyeli transkriptomik verinin yaygın olarak kullanılmasının temel sebebi diğer omik veri tiplerine göre mevcut teknolojilerle erişilebilirliğinin çok daha yüksek olmasıdır. Dahası son yıllarda geliştirilen teknolojiler [50, 51] sayesinde patojenlerden de tek-hücre gen ifade seviyelerini dizileme mümkün hale gelmiştir. Bu sayede hem konak canlı hem de patojen tarafındaki heterojen yapıların bulaşı ve hastalığa etkisini incelemek mümkün hale gelmektedir.

Konak canlı ile patojen arasındaki etkileşimleri farklı biyolojik boyutlarda incelemek amacıyla transkriptomik veri harici tek-hücre seviyeli omik veri kümelerini kullanan çalışmalar da mevcuttur. Örnek bir çalışmada, tek-hücre seviyeli epigenomik veri kümeleri kullanılarak DNA üzerindeki başkalaşımların patojenlerin bulaşma mekanizmalarına etkisi incelenmektedir [52]. Mevcut bir diğer çalışmada tek-hücre seviyeli proteomik ve transkriptomik veri barındıran ECCITE-seq [27] teknolojisinden faydalananak HIV-1 virüsünün bulaştığı hücrelerdeki protein seviye değişimleri incelenmektedir [53]. Bir başka çalışmada SARS-COV-2 ile enfekte kişilerin başkalaşan



Şekil 1.3: Tez adımlarının şematik gösterimi. **A** Proteomik veri kümelerini yatay bütünleştirme. **B** Çoklu-omik veri kümelerini dikey bütünleştirme. **C** Hücrelerin viral aktivite seviyelerinin hesaplanması. **D** Viral patogeneze etkin biyobelirteç analizi.

hücre gruplarındaki protein seviyeleri analiz edilmektedir [54]. Bu analizlerle enfeksiyonun heterojen yapıdaki hücre grupları üzerindeki etkisi incelenmektedir. Elde edilen bilgiler sayesinde viral patogenezin hücresel boyuttaki etkileri çok sayıda hücreden elde edilen ölçümler kullanılarak aydınlatılmaktadır.

Yöntemlerinin etkinliğini test etmek amacıyla tez çalışmasında önerilen analiz için öncelikle çoklu-omik veri kümelerindeki proteomik ölçümler üzerindeki toplu etki geliştirilen yatay bütünleştirme algoritması kullanılarak giderilmiştir (Şekil 1.3A). Bütünleştirme sonrası veri kümelerinde ortak bulunan proteinlere ait temizlenmiş ölçümler tüm hücreler için elde edilmektedir. Proteomik ve transkriptomik ölçüm içeren çoklu-omik veri kümeleri daha sonrasında dikey bütünleştirilerek hücresel heterojen yapıyı daha iyi ortaya çıkaran bütünleşik hücre gömmeleri (ing. embeddings) elde

edilmiştir (Şekil 1.3B).

Bütünleşik veri üzerinde viral patogeneze aktif rol oynayan hücrelerin tespit edilebilmesi için toplu dizileme patojen ve konak arasındaki protein – protein etkileşim ağlarındaki hedef proteinlerden faydalanılarak her bir hücre için viral aktivite seviyesi hesaplanmaktadır (Şekil 1.3C). Hesaplanan aktivite seviyeleri daha sonra viral patogeneze etkin özelleşmiş hücre alt kümelerinin tespitinde kullanılmaktadır. Son adımda Şekil 1.3B’de elde edilen bütünleşik ölçümler üzerinde kümeleme işlemi gerçekleştirilerek özelleşmiş hücre alt kümeleri tespit edilmektedir (Şekil 1.3D). Tespit edilen altkümelerindeki hücrelerin viral aktivite seviyelerinden faydalanılarak patogeneze aktif hücre altkümeleri belirlenmektedir. Son olarak tespit edilen aktif hücre altkümelerindeki hücrelerin ham proteomik ve transkriptomik ölçümleri üzerinde öznitelik önem analizi gerçekleştirilerek bu hücreleri diğer hücrelerden ayıran biyobelirteçler belirlenmektedir.

Yukarıda önerilen analiz yöntemiyle tanımlanan biyobelirteçler viral patogeneze mekanizmalarının anlaşılmasına imkân sağlamaktadır. Farklılaşan hücresel alt kümelere özgü moleküler belirteçlerin ortaya çıkarılmasıyla, virüslerin konak sistemleri nasıl manipüle ettiği ve farklı hastalık sonuçlarına neden olduğu daha iyi anlaşılabilir. Bu biyobelirteçler ile enfeksiyon sırasında bozulan önemli yollar ve hücresel işlevler belirlenerek hastalık yapıcı etkilerin keşfedilmesine olanak sağlanmaktadır. Ayrıca, enfekte hücrelerin heterojen yanıtlarının incelenmesi, viral kalıcılık veya temizlenme süreçlerine nasıl katkıda bulunduklarını anlamamıza yardımcı olmaktadır. Bu tür keşifler, hedefe yönelik tedaviler ve erken teşhis için kullanılabilecek biyobelirteçler sayesinde viral hastalıklarla mücadelede yeni yaklaşımlar geliştirilmesine olanak sağlamaktadır.

Tez çalışması kapsamında geliştirilen bütünleştirme yöntemleri ve elde edilen deneysel sonuçlar ile birlikte üç ana bölüm altında verilmektedir. İlk bölümde, proteomik veri kümelerinde toplu etkileri gidermek için geliştirilen yatay bütünleştirme yaklaşımı detaylı bir şekilde açıklanmaktadır. İkinci bölümde, transkriptomik ve proteomik verilerin dikey bütünleştirilmesi ile hücresel heterojenliğin birleşik ölçümler yoluyla daha net bir şekilde ortaya çıkarılması için geliştirilen yöntem ele alınmaktadır. Son bölümde ise geliştirilen yöntemler ile bütünleştirilmiş veriler üzerinde gerçekleştirilen analizler ile viral patogeneze aktif rol oynayan özelleşmiş hücre alt kümelerinin ve bu altkümelerde ayırıştırıcı olan biyobelirteçlerin tespitine yönelik çalışmalar yer almaktadır.

Her bölüm, hesaplamalı, biyolojik ve istatistiksel perspektifleri bir araya getirerek kapsamlı bir analiz sunmaktadır. Her bölümün sonunda elde edilen bulgular, amaçlanan veri bütünleştirme veya patogeneze analizine katkıları tartışılmaktadır.

1.5. Tezin Katkısı

Yapılan bu tez çalışması ile Biyoinformatik alanına sekiz önemli katkı sağlanmaktadır:

- Aynı tip omik veri barındıran tek-hücre veri kümeleri arasında benzer hücrelerin tespiti için özgün bir yöntem geliştirilmiştir.
- Tek-hücre seviyeli proteomik veri kümelerinin biyolojik mana korunarak yüksek performansla yatay bütünleştirilmesini sağlayan SCPRO-HI yöntemi geliştirilmiştir.
- SCPRO-HI yöntemi Python dilinde kodlanarak açık kaynaklı bir analiz aracı olarak araştırmacıların kullanımına sunulmuştur.
- Biyolojik bilginin dahil edilmesine imkân sunan hücreler arası benzerlik metriği önerilmiştir.
- Tek-hücre seviyeli çoklu-omik veri kümelerini açıklanabilir ilişkilerle dikey bütünleştirilmesini sağlayan SCPRO-VI yöntemi geliştirilmiştir.
- SCPRO-VI yöntemi Python dilinde kodlanarak açık kaynaklı bir analiz aracı olarak araştırmacıların kullanımına sunulmuştur.
- Hücrelerin viral aktivite seviyelerinin hesaplanmasını sağlayan özgün bir etiketleme algoritması geliştirilmiştir.
- Viral aktivite etiketlerinden faydalanılarak tek-hücre seviyeli bütünleşik ölçümler üzerinde viral patogeneze etkin özelleşmiş hücre altkümelerinin tespiti ve bu kümeler için ayrıştırıcı biyobelirteçlerin belirlenmesi için bir analiz çerçevesi önerilmiştir.

2. YATAY VERİ BÜTÜNLEŞTİRME

2.1. Giriş

Proteinler, çeşitli biyolojik süreçlerde kritik roller oynayan önemli biyomoleküllerdir. Son yirmi yılda, protein seviyelerinin hızlı ve düşük maliyetli bir şekilde ölçülmesini sağlayan yüksek verimli dizileme araçları geliştirilmiştir ve bu araçlar, çeşitli ileri analizlerde kullanılmaktadır. Ancak, bu yöntemler verileri toplu bir şekilde dizileyerek hücreden hücreye heterojenlik hakkında bilgi sağlamaz. Farklı hücre gruplarının hastalıklar ve metabolik aktiviteler üzerindeki etkisini anlamak için tek-hücre seviyesinde ölçümlere sahip olmak esastır. Tek-hücre veri setleri kullanılarak, bireysel hücrelerin benzersiz özelliklerini ve biyolojik süreçlere katkılarını daha kapsamlı ve doğru bir şekilde anlamak mümkündür. Ancak, farklı bireylerden veya dokulardan elde edilen tek-hücre veri setlerinin biyolojik ve teknik gürültü nedeniyle doğrudan birlikte kullanılması mümkün değildir [41].

Toplu etki genellikle, hücre hattı oluşturulurken kullanılan deney ortamına veya deneyi gerçekleştiren araştırmacıya bağlı olarak rastgele bir şekilde ortaya çıkar [42]. Ayrıca, dizileme için kullanılan farklı teknolojiler de toplu etkiye yol açabilir[43]. Diğer yandan, biyolojik gürültü, birden fazla doku veya bireyden elde edilen verilerde ortaya çıkmaktadır [8]. Bunun yanı sıra, omik verilerin tüm hücreden veya sadece hücre çekirdeğinden dizilenmesi de biyolojik gürültüye neden olmaktadır. Dokular arasındaki farklılıklara, hücre çekirdeği veya tüm hücrenin dizilenmesine ve farklı bireylerden elde edilen verilere bağlı olarak biyolojik gürültü miktarı artar [44]. Toplu etki, ileri analizlerin doğruluğunu artırmak için yüksek değişken proteinleri korumak gibi veri setlerinin biyolojik anlamını koruyarak kaldırılmalıdır.

Farklı deneylerden elde edilen aynı tür omik veri setlerini toplu etkiyi kaldırarak bütünleştirme işlemi yatay bütünleştirme olarak isimlendirilmektedir [41]. Son yıllarda, tek-hücre omik veri setlerinin bütünleştirilmesi, Biyoinformatik alanında popüler bir araştırma konusu haline gelmiştir [36]. Bütünleştirilmiş tek-hücre veri setleri, hücre türü sınıflandırması [35], benzersiz hücre gruplarının tanımlanması [37], hücre gidişizlerinin ortaya çıkarılması [38], hücresel yaşlanmanın tespiti [39] ve dokulardaki uzamsal

yolakların tanımlanması [40] gibi çeşitli problemleri çözmek için kullanılmaktadır.

Tek-hücre transkriptomik verilerinin artan popülaritesi, bu veri setlerini bütünleştirmek için çeşitli yöntemlerin geliştirilmesini sağlamıştır. Bu yöntemlerin birçoğu çapa tabanlı bir yaklaşım kullanmaktadır [55]. Bu yöntemler genellikle benzer hücrelerin tanımlanmasını ve bunların bütünleştirme için bir referans olarak kullanılmasını içermektedir. En popüler çapa tabanlı yöntem olan Ortak Enyakın Komşuluk (MNN) [56], benzer hücrelerin tespitinde yüksek başarımlı yöntemler tarafından yaygın olarak kullanılmaktadır [57, 58]. MNN algoritması, birbirinin en yakın komşusu olan hücre çiftlerini bularak benzer hücreleri tanımlar. Ancak, bu yöntemler veri setlerinin büyük ve karşılıklı en yakın hücre çiftleri az olduğu durumlarda başarısız olabilmektedir [44].

Geleneksel MNN yaklaşımına yönelik bir endişe, bireysel hücrelerin benzerliğini değerlendirirken, hücrelerin ait olduğu gruplar yerine bireysel hücrelerin eşleşmesinden kaynaklanan yanlış eşleşmelerin oluşabilmesidir. Bu nedenle, scMerge [59] ve SMNN [57] gibi yaklaşımlar, önce benzer hücreleri gruplandırır ve ardından bu kümelere ayrı ayrı MNN işlemini uygular. Alternatif olarak, CLAIRE [60] gibi yöntemler, sınırlı sayıda MNN çiftini kullanarak yapay eşleşmeler oluşturmakta ve ardından yanlış eşleşmeleri gidermek için belirlenen gerçek MNN çiftleri ile yapay eşleşmeler iyileştirilmektedir.

MNN'ye alternatif olarak, hücrelerin omik vektörleri arasındaki mesafelere dayalı benzerlik çizgeleri oluşturan çizge tabanlı yöntemler de bulunmaktadır. Bu benzerlik çizgesi, gözetimsiz bir şekilde hücre kümelerinin tanımlanması için kullanılmaktadır [61]. Aynı kümede bulunan hücreler benzer hücreler olarak belirlenir ve toplu etki giderilirken referans olarak kullanılır. Sadece ölçümlere göre sınırlı sayıda benzer hücrenin olduğu veri setlerinde, çizge tabanlı yöntemler çapa tabanlı yöntemlere göre daha yüksek performans göstermektedir. Diğer yandan, bütünleştirilmiş veri setlerinin performans analizleri, çapa tabanlı yöntemlerin biyolojik anlamı korumada çizge tabanlı yöntemlere göre daha başarılı olduğunu göstermiştir [44].

Derin öğrenme tabanlı yaklaşımlar, yatay bütünleştirme için giderek daha popüler hale gelmektedir [62–65]. Bu yöntemlerin motivasyonu, bütünleştirme işlemini gözetimsiz bir şekilde gerçekleştirme yeteneği ve derin öğrenme modellerinin elde edilen verinin biyolojik anlamını korumadaki başarısına dayanmaktadır [44]. Ancak, bütünleştirilmiş veriler genellikle yeni bir uzaya taşınmakta ve bu durum, elde edilen

verilerin ileri analizlerde doğrudan omik verileri olarak kullanılmasını zorlaştırmaktadır. Bu, bütünleştirilmiş verilerdeki biyolojik sinyallerin tüm ölçümlerini gerektiren gelişmiş analiz tekniklerinin kullanımını engelleyebilmektedir. Bununla birlikte, bütünleştirilmiş veri setindeki sinyallerin göreceli ölçümlerinin yeterli olduğu analizlerde derin öğrenme tabanlı algoritmalar yaygın olarak tercih edilmektedir.

Tek-hücre proteomik veri setlerinin bütünleştirilmesine yönelik yöntemlerin eksikliği, büyük ölçüde yüksek çıktılı tek-hücre proteomik verilerinin üretilmesine olanak tanıyan teknolojilerin henüz tam olgunlaşmamış olmasından kaynaklanmaktadır. Ancak, yeni dizileme teknolojilerinin ortaya çıkışıyla birlikte [66], bu tür veri setlerinin açık kaynak veri tabanlarındaki sayısı giderek artmaktadır.

Transkriptomik veri setlerini bütünleştirmek için geliştirilen yöntemler, proteomik veri setlerine de uygulanabilmektedir. Farklı veri setlerindeki benzer hücrelerin tanımlanması ve aşırı öğrenme nedeniyle bütünleştirme sırasında biyolojik anlamın kaybolmasının önlenmesi gibi zorluklar, her iki tür omik veri için de ele alınması gereken sorunlardır. Bununla birlikte, bu yöntemler yüksek boyutlu gen verilerini işlemede etkili olsa da, genellikle sınırlı protein ölçümlerine sahip tek-hücre proteomik veri setlerine uygulandığında performansları düşmektedir.

Sınırlı kapsam zorluğuna ek olarak, proteinlerin düşük bolluk seviyeleri ve tek-hücre proteomik veri setlerindeki seyrek yapı da göz önünde bulundurulmalıdır. RNA molekülleriyle karşılaştırıldığında, proteinler hücrelerde genellikle daha düşük miktarlarda bulunur; bu da doğru tespiti zorlaştırır ve daha yüksek seviyelerde gürültüye neden olur. Ayrıca, tek-hücre proteomik veriler genellikle tüm proteinlerin her hücre için ölçülmediği seyrek verilerdir. Bu nedenle, bu zorlukların üstesinden gelmek ve alandaki mevcut boşluğu doldurmak için tek-hücre proteomik veri seti bütünleştirmeye yönelik yüksek performanslı algoritmalara ihtiyaç duyulmaktadır.

Bildiğimiz kadarıyla, MARIO [67] algoritması, tek-hücre proteomik veri setlerini bütünleştirmek için mevcut olan tek yöntemdir. Bu algoritma, özellikle tek-hücre proteomik veri setlerinin yatay bütünleştirilmesi için tasarlanmıştır ve transkriptomik veri setlerinin bütünleştirilmesine odaklanmamaktadır. Yöntem, iki farklı veri setindeki hücreler arasında, paylaşılan ve tüm öznelikleri kullanarak iki mesafe matrisi hesaplar. Daha sonra, bu mesafe matrislerinin interpolasyonu kullanılarak hücre çapaları

oluşturulur. Eğer çapa kümeleri, karşılıklı bir k-means kümeleme kompozisyonunda eşleşmiyorsa, çapalar filtrelendir. Kalan çapalar, Kanonik Korelasyon Analizi (CCA) algoritması [35] kullanılarak veri setlerini yeni bir uzayda bütünleştirilmektedir.

Bu çalışmada, antikor tabanlı tek-hücre proteomik veri setlerini toplu etkiyi kaldırarak bütünleştirmek için varyasyonel oto-kodlayıcı (VAE) [68] modelleri kullanan özgün bir yöntem (SCPRO-HI) sunulmaktadır. Bu yöntem, seyrek veri setlerinde bile yüksek miktarda ve yüksek doğrulukta hücre eşleşmesine olanak tanımaktadır. Ortaya çıkan hücre grupları, bazı tekil hücrelerde belirli proteinler için bolluk verisi olmasa bile her protein için bir bolluk profili sunmaktadır.

Benzer kümeler, kümeler içindeki ayırt edici proteinler kullanılarak tanımlanmaktadır. Bu ayırt edici proteinler, düşük bolluklu protein varlığında bile, kümeler arasındaki heterojenliği sergileyebilmektedir. Ayırt edici bu özneliklerin önceliklendirilmesiyle, yöntemimiz düşük protein bolluğu sorununu etkili bir şekilde ele alarak veri setleri arasında biyolojik olarak ilişkili kümeleri yakalamamızı sağlamaktadır.

Önerilen VAE modeli, CLAIRE [60] tarafından kullanılan karşıt öğrenme yaklaşımına benzer bir eğitim stratejisi izleyerek bütünleştirmede hücreleri, pozitif ve negatif çapa hücreleri ile birlikte kullanır. Her iki yöntemin, eş zamanlı olarak ancak bağımsız çalışmalarda geliştirilmiş olduğu belirtilmelidir. Ancak, karşıt öğrenme yöntemlerinden farklı olarak, yöntemimiz iki Siyam oto-kodlayıcı yerine tek bir varyasyonel oto-kodlayıcı kullanmaktadır.

Önerilen yöntem, eğitim sırasında iki eş zamanlı görev gerçekleştirir: pozitif hücreleri birbirine yaklaştırmak ve onları negatif çapadan ayırmak. Sonuç olarak, model, hücreleri bir gizli alana gömmek yerine, her hücre için düzeltilmiş protein ölçümleri üretir. Böylece, her protein için daha doğru ve daha az seyrek ölçümler elde edilir.

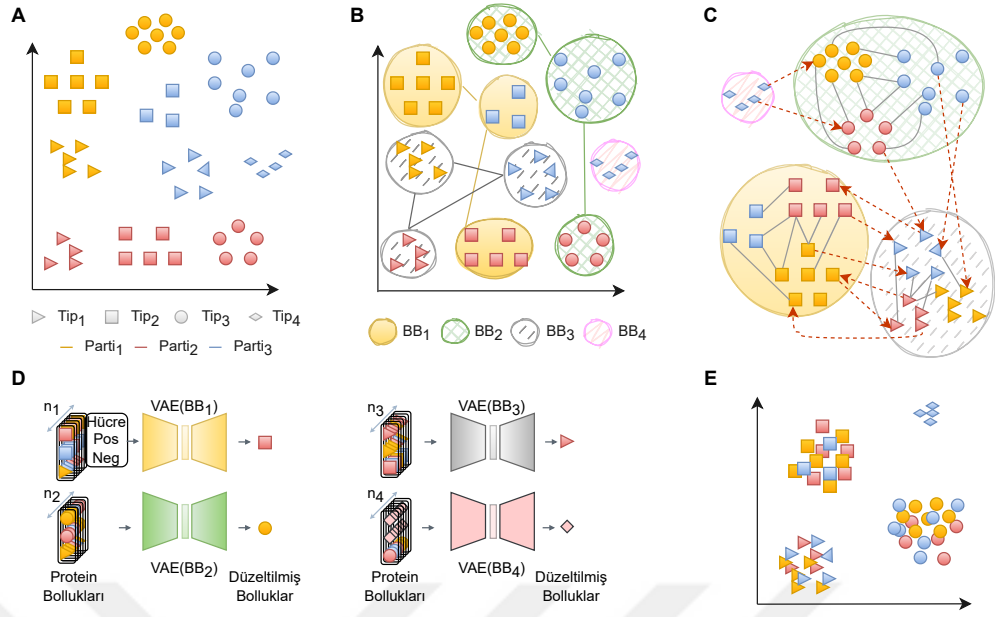
Toplu etki, herhangi bir etiket veya ön biyolojik bilgi kullanılmadan, programatik olarak tespit edilen her hücre türü için bağımsız bir VAE modeli eğitilerek modellenir. Sonuç olarak, hücrelerin protein ölçümlerinin düzeltilmesi ve ortak bir alana bütünleştirilmesi sağlanmakta ve böylece hücrelerdeki proteinlerin düzeltilmiş ölçümleriyle ileri analizlerin doğruluğu artırılmaktadır.

Önerilen yöntem, yatay veri bütünleştirme için geliştirilmiş yüksek başarılı Harmony [37], Scanorama [69], fastMNN [56], scVI[62] ve MARIO[67] algoritmalarıyla

karşılaştırılmıştır. Bu algoritmaların performansı, bütünleştirme için sentetik ve gerçek dünya tek-hücre proteomik veri setleri kullanılarak hesaplamalı ve biyolojik olarak değerlendirilmiştir. Ek olarak, önerilen modelin diğer omik veri türlerine genellenebilirliğini değerlendirmek için dört açık kaynak tek-hücre transkriptomik (scRNA-seq) veri kümesinden derlenen bir veri kümesi kullanılmıştır. Deneysel sonuçlar, SCPRO-HI algoritmasının, antikör tabanlı tek-hücre proteomik veri setlerini bütünleştirmede rakiplerinden üstün olduğunu silüet skoru metriğinde %75’lik önemli bir iyileşme göstererek kanıtlamıştır. Dahası yöntemin HVP’lerin korumada %13 daha iyi sonuçlar elde etmesi biyolojik anlamı koruduğunu göstermektedir. Ayrıca, scRNA-seq ve kütle spektrometresine dayalı proteomik veri setlerindeki performans sonuçları, algoritmanın çeşitli yüksek boyutlu omik veri setlerinin yatay bütünleştirilmesi için uygulanabilirliğini doğrulamaktadır.

2.2. Yöntem

SCPRO-HI yöntemi, Şekil 2.1’de gösterildiği gibi iki ana adımdan oluşmaktadır: hücre eşleştirme (A’dan C’ye) ve toplu etki düzeltme (D ve E). Hücre eşleştirme adımında, hücreler kümelenir ve aralarında çapalar oluşturulur. Bu kümeler ve çapalar, diğer yatay bütünleştirme algoritmalarının işlem hattında hücre eşleşmeleri olarak veya ileri analiz görevlerinde kullanılabilir. Partilerin bütünleştirilmesi, ölçümleri düşük boyutlu bir veri uzayına haritalamadan, varyasyonel oto-kodlayıcılar kullanılarak düzeltilmesiyle gerçekleştirilir. VAE’ler, hücre eşleştirme adımında oluşturulan çapaları kullanarak farklı partiler arasındaki benzer hücreler için karşılık gelen düzeltilmiş ölçümler üretir. Önerilen çapa oluşturma algoritması, scMerge [59] veya SMNN [57] yöntemlerine benzer şekilde her bir veri setini ayrı ayrı kümeler ve farklı veri setlerindeki benzer kümeler arasında çapalar oluşturur. Ancak, yüksek değişken genler/proteinler veya sabit ifade edilen genler (scSEGs) gibi küme içi özelliklere dayanmak yerine, veri setleri arasındaki benzer kümeleri tespit etmek için kümeleri diğer kümelerden ayırt edici proteinler kullanılmaktadır. Her küme için, bu kümenin kendi veri setindeki diğer kümelerden ayrılmasında her proteinin önemi hesaplanmaktadır. Farklı veri setlerindeki kümelerin, geri kalanından benzer proteinlerle ayrıldığı durumlarda, bu kümelerin aynı türde hücreler içerdiği varsayılmaktadır.



Şekil 2.1: SCPRO-HI algoritmasının bütünleştirme iş akış şematığı. **A** Verisetlerinin kümelenmesi. **B** Farklı veri setlerindeki kümelerin çapalanarak bağlı bileşenleri oluşturması. **C** Hücre çapalarının oluşturulması. **D** Özgün VAE modelleri kullanılarak toplu etkinin düzeltilmesi. **E** Düzeltilmiş bolluk vektörlerinin hesaplanması.

Kümeler ve çapaları, bağlı bileşenleri belirlemek için bir çizge olarak temsil edilir; düğümler kümeleri ve kenarlar kümelerin benzerliğini yani çapaları ifade eder. Her bağlı bileşen, belirli bir hücre türünü temsil eder ve bir hücrenin pozitif çapaları, kendi kümesiyle aynı bağlı bileşen içinde bulunan kümelerden seçilir. Haghverdi'nin MNN yöntemini[56] kullanarak aynı bağlı bileşende fakat farklı kümelerde bulunan hücreler arasında çapalar oluşturulur. Ayrıca, her çapalı hücre çifti için, ilgili bağlı bileşenden dışında kalan hücrelerden rastgele bir negatif çapa seçilir.

Benzer hücrelerin öncelikle kümelenmesi, MNN algoritmasının etkinliğini önemli ölçüde artırır. Bu durum, özellikle düşük bolluklu proteinlerin neden olduğu zorlukları azaltmada değerlidir. Biyolojik özellikleri benzer olan hücreleri gruplandırarak, yaklaşımımız her bir hücrenin k-en yakın komşuluğunda birbiriyle örtüşebilecek farklı hücre türlerini ayırmayı mümkün kılar. Bu iki adımlı süreç, daha kesin ve fazla miktarda hücre eşleşmelerine katkıda bulunur.

Adım D'de, varyasyonel oto-Kodlayıcı modelleri, toplu etki düzeltilmiş protein bolluk vektörleri üretmek için eğitilir. Bu modeller, bir hücrenin pozitif ve negatif çapalarının

ölçümleriyle birlikte, hücrenin kendi ölçümlerini kullanarak toplu etkilerini kaldıracak şekilde eğitilir. Bu, modelin hem hücreye hem de pozitif çapaya benzeyen, ancak negatif çapaya benzerlik göstermeyen yeni protein bollukları üretmesini sağlar. Her bağlı bileşen için VAE modeli, o bağlı bileşendeki hücreler için oluşturulan hücre üçlülerini (hücre, pozitif çapalı hücre, negatif çapalı hücre) kullanarak gözetimsiz bir şekilde eğitilir. Modeller eğitildikten sonra, hücrelerin kendileri pozitif çapalar olarak kullanılarak düzeltilmiş protein bollukları üretilir.

Diğer VAE tabanlı yöntemlerden [62, 70] farklı olarak, yöntemimiz ölçümleri yeni bir uzaya haritalamak yerine kendi uzayında düzeltir ve böylece protein bolluk ölçümlerindeki toplu etkiyi temizlemiş olur. Bu yaklaşım, ileri analiz sırasında veri setlerindeki her bir proteinin bireysel olarak değerlendirilmesine olanak tanır. Aşağıdaki bölümlerde, hücre eşleştirme ve toplu etkinin kaldırılması adımlarının her biri için kullanılan yöntemler ve teknikler dahil olmak üzere detaylı açıklamalar sağlanmaktadır.

2.2.1. Hücre Eşleştirme

SCPRO-HI'yi yatay bütünleştirme için kullanmak amacıyla, tüm veri setlerinin aynı proteinler için ölçümler içermesi gerekmektedir. Bu nedenle, tüm veri setleri, ortak proteinler kullanılarak filtrelenir. Daha sonra, dizileme sırasında kullanılan teknoloji nedeniyle veri setlerinin alanı değişebileceğinden, eşleştirmeden önce veri setleri z-skor normalizasyonu kullanılarak normalize edilir:

$$z_i = \frac{x_i - \mu}{\sigma} \quad (2.1)$$

Formül 2.1'de, x_i hücre i 'nin protein bolluk ölçümlerini, z_i normalleştirilmiş bolluğu, μ ve σ ise sırasıyla veri setinin ortalamasını ve standart sapmasını ifade eder.

Normalize edilmiş ölçümler, toplu etkiyi hafifletmek ve hesaplama kaynakları gereksinimlerini azaltmak için iki boyutlu bir uzaya gömülür. UMAP algoritması [71] hücre gömmeleri oluşturmak için kullanılır. Bu yöntem, yüksek başarımlı algoritmalarda yaygın olarak kullanılan manifold öğrenme tabanlı bir yöntemdir [72]. UMAP'ın optimizasyon stratejileri, yüksek boyutlu verileri daha düşük boyutlu bir alana haritalarken komşuluk yapısını korumayı önceliklendirir. Bu yaklaşım, ölçümlerdeki gürültüyü etkili bir şekilde azaltırken hücreler arasındaki yerel ilişkileri

korur. Sonuç olarak, aynı tür hücreleri bu iki boyutlu uzayda kümelemeyi daha kolay hale getirmektedir.

Yoğunluk tabanlı DBSCAN algoritması [73], UMAP gömmeleri için uygun olduğundan, kümeleme için kullanılmaktadırlar. DBSCAN algoritması, UMAP algoritmasıyla benzer olarak değerlendirilen komşuların sayısına ve bir komşuluk yarıçapına dayalı olarak kümeleri belirler (Şekil 2.1A). Yarıçap veya epsilon parametresi, hücrelerin aynı kümenin bir parçası olarak kabul edilebilmesi için birbirine ne kadar yakın olması gerektiğini belirtir. Bu parametre, kümeleme üzerinde güçlü bir etkiye sahip olduğundan dikkatlice ayarlanmalıdır. Rahmah'ın stratejisi [74], optimal epsilon değerini belirlemek için kullanılır. Bu strateji, önce herhangi bir nokta çiftinin arasındaki mesafeyi hesaplar ve ardından nadir bulunan mesafelerin minimumunu epsilon değeri olarak seçer. Bu, genellikle aynı mesafeye sahip hücrelerin bir arada tutulmasını sağlar. Minimum komşu sayısı 3 olarak ayarlanmıştır; bu değer, yerel olarak alınabilir ve aktarılabilir veri sağladığından yaygın olarak kullanılmaktadır.

Veri setleri bireysel olarak kümelendir. Farklı veri setlerindeki benzer kümeler çapalanır çünkü bütünleştirme işleminin aynı tür hücre gruplarını bir araya getirmesi beklenir (Şekil 2.1B). Benzerlik metriği, kümelerdeki proteinlerin öznelik önem değerine dayanır ve bu değer, bir proteinin kendi kümesini aynı veri setindeki diğer kümelerden ayırmadaki önem düzeyini temsil eder.

Protein önem değerleri karar ağaçları kullanılarak hesaplanmaktadır. Her küme için, protein bollukları kullanılarak ikili sınıflandırıcı bir karar ağacı eğitilir. Karar ağaçları, gözetimsiz kümeleme ile belirlenen küme etiketleri kullanılarak denetimli bir şekilde eğitilir. Eğitimde mevcut kümeye ait hücrelere 1 etiketi, veri setindeki diğer hücrelere ise 0 etiketi atanır. Protein önem değerleri, eğitilmiş modellerden elde edilir; bu değerler, model eğitimi sırasında ilgili özneliğin (proteinin) kullanılmasıyla elde edilen bilginin yüzdesini temsil eder ve toplamaları her zaman 1'e eşittir.

Farklı veri setlerindeki her küme çifti arasındaki benzerlik, protein önem vektörleri arasındaki kosinüs benzerliği ölçülerek hesaplanır. Kosinüs benzerliği, küme çiftinin ilgili önemli proteinleri arasındaki ikili benzerlik yüksek olduğunda artma eğilimindedir. Sonuç olarak, 0 değeri, eşleştirilmiş kümelerin herhangi bir ortak önemli proteine sahip olmadığını belirtirken, 1 değeri, protein önem değerlerinin aynı olduğunu ifade eder.

Küme çapaları, varsayılan olarak benzerlikleri %60'tan yüksek olan küme çiftleri arasında kurulur. Deneysel olarak belirlenen bu eşik değeri, genelleştirilmiş bütünleştirme ile hücre türü hassasiyetine sahip bütünleştirme arasında denge kurmak için ayarlanabilir. Daha düşük bir benzerlik eşiği, toplu etkiyi birleştirilmiş bir şekilde modelleyen daha genel bir bütünleştirme ile sonuçlanır. Öte yandan, daha yüksek bir eşik değeri, kümeler arasındaki çapaların sayısını azaltır; bu da bütünleştirme kalitesini düşürür ancak hassasiyeti artırır, çünkü her benzersiz hücre kümesi kendi toplu etkisini modeller. Eşik değeri seçilirken, hedeflenen ileri analizlerin özel gereksinimleri göz önünde bulundurulmalıdır.

Çapalar ve kümeler, bağlı bileşenleri belirlemek için bir çizge olarak temsil edilir. Çizgede çapalar kenarları, kümeler ise düğümleri ifade eder. Bağlı bir bileşen, her iki düğümün bir yol ile bağlandığı ve bağlı bileşeni oluşturan düğümlerin dışında kalan diğer düğümlerle bağlantılı olmadığı bir alt çizgedir. Bağlı bir bileşendeki tüm hücre kümeleri, benzer proteinleri paylaşarak veri setlerinden ayrıldığından, farklı veri setlerindeki belirli bir hücre türünün hücrelerinin bir bağlı bileşen oluşturduğu varsayılmaktadır.

Algoritma, eşleşmeyen kümelere bir bağlı bileşene bağlanma şansı daha sunmaktadır. Bu yöntemle, bağlı bileşendeki bazı kümelere oldukça benzeyen kümelerin dahil edilmesini sağlar. Bu işlem, bağlı bileşendeki hücreleri birleşik bir küme olarak ele alarak, küçük farklarla eşleşmeyen kümelerin başarıyla bağlanmasını mümkün kılar. Bağlı bileşenlerin önemli proteinleri, bireysel kümeler için uygulanan aynı prosedürle hesaplanır ve her eşleşmeyen küme ile her bağlı bileşen arasındaki benzerlik hesaplanır. Eşleşmeyen küme benzerlik değeri %60 eşik değerini aşan bağlı bileşenlerden en benzerine çapalanır. Bu işlemden sonra eşleşmeden kalan kümeler, kendi başlarına bağlı bileşen olarak kabul edilir.

Hücrelerin pozitif çapaları, bütünleştirilmiş verilerde aynı hücre türüne ait hücreleri bir araya getirmek amacıyla kendi bağlı bileşenleri içinden seçilir (Şekil 2.1C). Çapalar, bir bağlı bileşendeki her hücre için MNN algoritması kullanılarak bireysel olarak hesaplanır. MNN algoritmasında, iki hücre, farklı hücre kümelerindeyken ve birbirlerinin k-en yakın komşuluğunda bulunduklarında çapa oluşturulur. Algoritma, bir bağlı bileşendeki hücrelerin tamamını MNN'leri belirlerken dikkate alır. Farklı veri kümelerindeki benzer hücrelerin tespit edilerek veri kümelerinin homojen biçimde bütünleştirilebilmesi için hücrenin kendi kümesindeki hücreler hariç tutulmaktadır. Her hücre için oluşturulan

pozitif çapa sayısı 0 ile k arasında değişmektedir. MNN için belirlenen k komşuluk düzeyi genişledikçe çapalanan hücre sayısı artmakta fakat bu durum yerel komşulukların kaybolmasına ve dolayısıyla bütünleştirme başarımının düşmesine sebep olmaktadır. Bu çalışmada k değeri için orijinal çalışmada önerilen 20 değeri belirlenmiştir.

VAE modellerini eğitmek ve toplu etkiyi düzeltmek için hücre üçlüleri (hücre, pozitif çapalı hücre ve negatif çapalı hücre) oluşturulur. Üçlü sayısı, pozitif hücre çapalarının toplam sayısına eşittir; çünkü üçlüler, bir hücrenin her pozitif çapa çiftine bir negatif çapa eklenerek oluşturulur. Üçlüdeki negatif çapa, ilgili hücreyle aynı bağlı bileşende olmayan kümelerden rastgele seçilir. Hücre üçlüleri oluşturulduktan sonra hücre eşleştirme adımı tamamlanır.

2.2.2. Toplu Etki Düzeltme

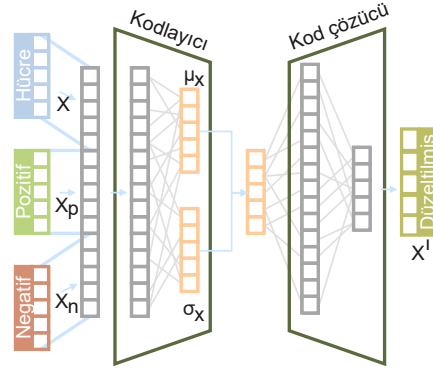
Şekil 2.2’de gösterilen özgün VAE modeli, toplu etki olmaksızın tam protein bolluklarını oluşturmak için geliştirilmiştir. Model, pozitif çapalanmış hücreler arasındaki benzerliği artırırken, negatif çapalanmış hücrelerle olan benzerliği azaltmayı amaçlar. Modeli eğitmek için aşağıdaki kayıp fonksiyonu kullanılmaktadır:

$$L = \exp((X' - X)^2 + (X' - X_p)^2 - (X' - X_n)^2) + KL[N(\mu_x, \sigma_x), N(0, 1)] \quad (2.2)$$

Kayıp fonksiyonu 2.2’de, X' modelin çıktısı olan bir hücrenin düzeltilmiş protein bolluk vektörüdür. X , hücredeki proteinlerin ham ölçümüdür. X_p ve X_n sırasıyla pozitif ve negatif çapa hücrelerinin ham bolluk vektörleridir. $\exp(\cdot)$ fonksiyonu, negatif çapaların benzerliğinden kaynaklanan negatif kayıp değerlerini önlemek için kullanılmıştır.

KL, Kullback-Leibler sapması [75] için bir kısaltmadır ve gizli uzay örnekleme fonksiyonunun μ ve σ değerlerini kullanarak gizli uzay ile Normal dağılım arasındaki mesafeyi hesaplar. Bu tür bir kayıp fonksiyonu, modelin, hem hücreye hem de pozitif çapalarına benzeyen bir çıktı bolluk vektörü üretmesini zorlamaktadır. Böylece, bütünleştirme işlemlerinde tüm veri setleri için partiye özgü gürültüden arındırılmış ve veri setleri arasındaki karşılıklı bilgiyi optimize eden bir veri uzayı elde edilmektedir.

Önerilen yeni VAE modeli, üç protein bolluk vektörünü alır ve bunları uç uca ekleyerek birleştirir (Şekil 2.1D). Kodlayıcı ve kod çözücüde, her biri veri setlerindeki protein sayısının 3 katı boyutunda bir yoğun katman bulunur. Kodlayıcıdaki yoğun katmanın



Şekil 2.2: Toplu etki düzeltme için kullanılan varyasyonel oto-kodlayıcı model mimarisi.

çıktısı, örnekleme için kullanılan σ ve μ katmanlarına bağlanır ve bu katmanlar gizli gömmeleri oluşturur. Gizli gömme X_s , kod çözücündeki yoğun katmana iletilir. İlk yoğun katman daha sonra düzeltilmiş ölçümleri öğrenen X boyutunda bir yoğun katmana bağlanır. Son yoğun katmanın çıktısı, modelin çıktısı olarak hizmet etmektedir.

Önerilen algoritma, hücre türüne özgü protein bolluğu zirve değerlerini yumuşatmadan, toplu etkileri her hücre türü için ayrı ayrı modellemeyi amaçlar. Bu nedenle, her bağlı bileşen için bağımsız bir VAE modeli eğitilir. VAE modelleri, kendi bağlı bileşenlerindeki hücrelerden tanımlanan hücre üçlülere kullanılarak eğitilir. Eğitim süreci tamamlandıktan sonra modeller, yalnızca eğitimde kullanılan hücreler için değil, aynı zamanda veri setlerindeki tüm hücreler için düzeltilmiş protein bolluk vektörleri üretmeye hazır hale gelmektedir. Toplu etki, bağlı bileşendeki tüm kümeler arasındaki olası hücre çifti çapaları kullanılarak modellenir. Bu nedenle, modeller, eğitiminde yer almayan hücrelerdeki toplu etkileri de düzeltme potansiyeline sahiptir.

Düzeltilmiş bolluklar, eğitilmiş modellerden üretilirken hücrelere kendileri pozitif çapa olarak verilir. Bu yaklaşımın mantığı, farklı partilerde bulunan ancak aynı toplu etkiye sahip iki hücreyi kullanarak düzeltilmiş ölçümler oluşturmaktır. Negatif çapalar, eğitim sürecinde olduğu gibi, rastgele seçilerek toplu etki temizleme aşamasında da modele girdi olarak verilir. Modelin eğitim sürecine dahil edilmeyen hücrelerin negatif çapaları, daha önce açıklandığı gibi rastgele seçilir. Tüm hücreler kendi bağlı bileşenleri için eğitilmiş VAE'lerden geçirilerek düzeltilmiş protein bollukları elde edilir (Şekil 2.1E).

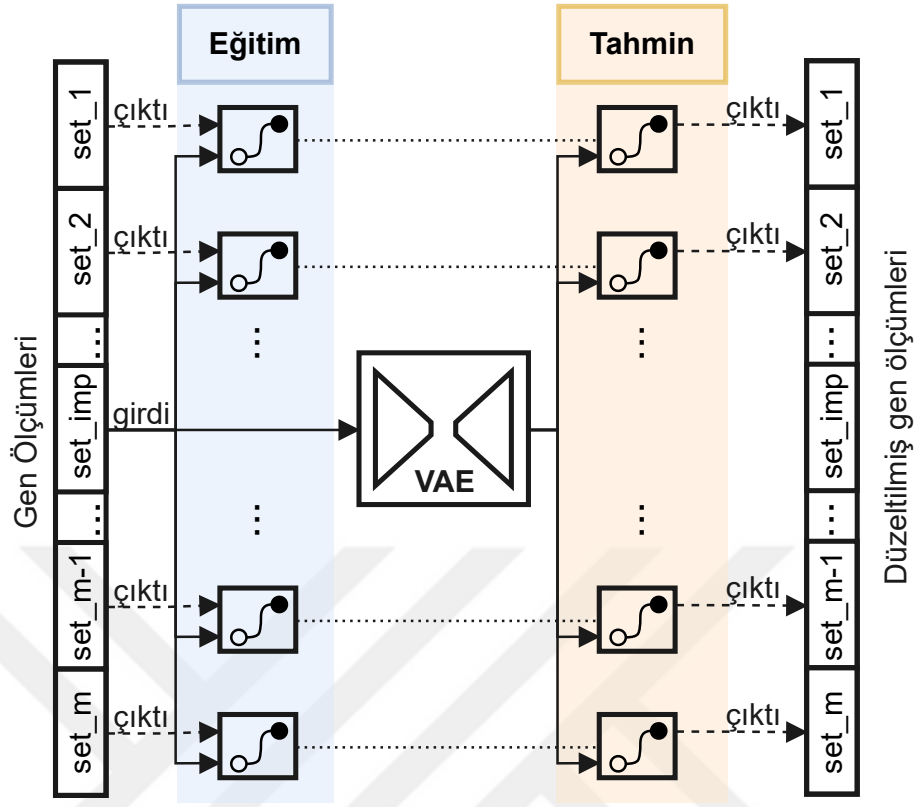
2.2.3. Yüksek Boyutlu Verilerle Çalışma

SCPRO-HI algoritması, tipik olarak onlarca veya yüzlerce protein içeren tek-hücre proteomik veri setlerinin bütünleştirilmesi için geliştirilmiştir. Ancak, önerilen algoritma, on binlerce gen ölçümünü içeren transkriptomik veri setleri gibi daha yüksek boyutlu omik veri setlerinin bütünleştirilmesi için genişletilebilir. Önerilen algoritma, özellik sayısı birkaç yüzü aştığında öğrenme tabanlı bileşenleri hesaplama açısından maliyetli olduğu için verimsiz hale gelebilir.

Bu kısıtı aşmak için algoritma, yüksek boyutlu veri setlerinin bütünleştirilmesi sırasında tüm ölçümleri kullanmak yerine VAE modellerinin eğitiminde n önemli genin kullanılmasına odaklanır. Bir bağlı bileşen için önemli gen listesi kullanıcı tarafından sağlanabilir veya veri setinin geri kalanından ayrılan en önemli n geni belirlemek için bağlı bileşen üzerinde protein önem hesaplama prosedürü uygulanarak tespit edilebilir. Bu anahtar genlerden yararlanarak, VAE modellerinin çıktısı hücre heterojenliğiyle ilgili bilgilerin çoğunu yakalar ve bu da onu çeşitli ileri analizler için uygun hale getirir. Ancak, algoritma tüm genler için düzeltilmiş ölçümleri hesaplamak için bir geliştirme de sunmaktadır. Her bağlı bileşenin önemli genleri hariç tutulur ve geriye kalan genler rastgele n gen içeren m sete ayrılır. Daha sonra, bağlı bileşenin önemli genleri ile setlerdeki genler arasındaki ilişkiyi modellemek için Şekil 2.3’de gösterilen çok katmanlı algılayıcı (MLP) regresyon modelleri kullanılır.

Regresyon modelleri, ilgili bağlı bileşenlerindeki tüm hücreler kullanılarak eğitilir. Her bir gen seti ile bir regresyon modeli ile eğitilir. Modellere girdi olarak hücrelerdeki önemli genlerin ifade seviyeleri verilerek setlerdeki genlerin ifadeleri çıktısı elde edilmektedir. Sonuç olarak, bu regresyon modelleri, önemli genler ile setlerdeki genler arasındaki ifade ilişkisini modellemektedir. Bu sonuç, toplu etki düzeltme için kullanılan VAE modellerinin, yalnızca önemli genleri değil, veri setindeki tüm genleri düzeltebilecek şekilde dönüştürücüler olarak kullanılmasını sağlar.

Regresyon modelleri eğitildikten sonra algoritma, tüm genler için toplu etkileri düzeltme imkânı sunmaktadır. İlk olarak, her bağlı bileşenle ilişkili VAE modeli, her hücredeki önemli genler için toplu etkiyi kaldırmak için kullanılır. Daha sonra, bu düzeltilmiş ifadeler, kendi bağlı bileşenlerine özgü regresyon modellerine girdi olarak verilmektedir. Regresyon modelleri, tüm genler için toplu etki düzeltilmiş ifadeleri üretmektedir. Bu



Şekil 2.3: Yüksek boyutlu veri kümeleri için düzeltilmiş ölçümler üretmeyi sağlayan yöntemin iş akış şematığı.

yaklaşım, tüm genler boyunca homojen bir toplu etki modellemesine neden olabilir ancak algoritmada yer alan çok katmanlı algılayıcıların (MLP'ler) genlere özgü doğrusal olmayan düzeltmeleri yakalayabilecek karmaşıklıkta olduğu varsayılmaktadır.

2.2.4. Karşılaştırmada Kullanılan Algoritmalar

2.2.4.1. MARIO

Karşılaştırmalarımızda, MARIO [67] algoritması kullanılmıştır. MARIO, tek hücre verilerinde farklı modaliteler arasında hücre eşleşmesi yaparak, ortak ve ayrı özellikleri dikkate alan bir bütünleştirme yöntemi sunmaktadır. Bu amaçla, robust optimal transport (ROT) yaklaşımını kullanarak, hücreler arasında biyolojik olarak anlamlı eşleşmeler oluşturmayı hedefler. Kıyaslamalarımızda, MARIO'nun varsayılan parametreleri

kullanılarak çalıştırılmıştır.

2.2.4.2. Harmony

Kıyaslamalarımızda, Harmony [37] algoritması da test edilmiştir. Harmony, çekirdek bileşen analizi (PCA) sonrası hiyerarşik grup regresyonu kullanarak örnekler veya koşullar arasındaki partisyon etkilerini gideren bir toplu etkisi düzeltme yöntemidir. Algoritma, yüksek boyutlu gen ekspresyon verilerini uyumlu hale getirmek için iteratif bir şekilde ağırlıklar güncelleyerek bütünleştirilmiş bir latent alan oluşturur. Harmony, hızlı hesaplama süresi ve esnek uyum mekanizması ile geniş çaplı tek hücre bütünleştirme çalışmalarında sıkça tercih edilen yöntemlerden biridir. Tüm deneylerde, Harmony'nin varsayılan ayarları kullanılarak hesaplamalar gerçekleştirilmiştir.

2.2.4.3. Scanorama

Karşılaştırmalarımızda, Scanorama [69] algoritması kullanılmıştır. Scanorama, L2-normalize edilmiş gen ekspresyon vektörleri üzerinde çalışan uzayda en yakın komşuları tespit ederek toplu etkilerini düzelten bir bütünleştirme algoritmasıdır. Özellikle büyük ölçekli veri setlerinde hızlı çalışabilmesi ve partisyon etkilerini başarılı bir şekilde giderebilmesi nedeniyle sıkça tercih edilmektedir. Algoritma, tüm veri setleri arasında ortak hücresel yapıları belirleyerek bunları birleştiren iteratif bir eşleştirme stratejisi kullanmaktadır. Kıyaslamalarımızda, Scanorama varsayılan parametrelerle çalıştırılmıştır.

2.2.4.4. FastMNN

Çalışmamızda, fastMNN [56] algoritması ile de kıyaslamalar gerçekleştirilmiştir. fastMNN, farklı veri setlerinden gelen hücreler arasında karşılıklı en yakın komşu eşleşmelerini belirleyerek toplu etkilerini gideren bir bütünleştirme yöntemidir. Geleneksel MNN yaklaşımının hesaplama yükünü azaltmak için hızlı PCA ve kademeli bütünleştirme stratejileri kullanmaktadır. fastMNN, özellikle farklı koşullardan gelen veri setlerini bütünleştirme açısından güçlü bir yöntem olarak kabul edilmektedir. Kıyaslamalarımızda varsayılan parametreler kullanılarak analizler gerçekleştirilmiştir.

2.2.4.5. ScVI

Karşılaştırmalarımızda, scVI [62] algoritması da kullanılmıştır. scVI, varyasyonel otokodlayıcılar (VAE) temelinde çalışan derin öğrenme tabanlı bir olasılıksal modeldir ve yüksek boyutlu tek hücre gen ekspresyon verilerini düşük boyutlu bir gizli uzaya gömmeyi amaçlamaktadır. Algoritma, hücreler arasındaki örüntüleri yakalayıp gürültüyü filtreler ve biyolojik olarak anlamlı gömüler üretir. scVI, özellikle eksik verilerin modellenmesi, gen düşüklüğünün düzeltilmesi ve toplu etkilerin giderilmesi gibi alanlarda güçlü bir yaklaşım sunmaktadır. Karşılaştırmalarda, scVI'nin varsayılan parametreleri kullanılmıştır.

2.2.5. Metrikler

Ortalama siluet genişliği (ASW), ayarlanmış rand indeksi (ARI) ve HVP Scib paketi [44] kullanılarak varsayılan parametrelerle hesaplanmıştır. iLISI, cLISI metrikleri Harmony paketinde [37] sunulduğu hali ile kullanılmıştır. HVP skoru Luecken tarafından önerilen HVG koruma stratejisi [44] ile hesaplanmıştır.

2.2.5.1. Ortalama Siluet Genişliği (ASW)

Ortalama siluet genişliği metrik, aynı anda hem küme içi benzerliği hem de kümeler arası heterojenliği yansıtır. Metrik, her bir hücre c için siluet katsayısı $(b - a) / \max(a, b)$ formülü kullanılarak hesaplanır. Burada, a , c ile aynı türdeki hücrelerin küme merkezine olan ortalama uzaklığını, b ise diğer hücre türlerinin tüm küme merkezlerine olan ortalama uzaklığını ifade eder. ASW değerleri 1 (en iyi) ile -1 arasında değişmektedir.

Scib paketi, Scikit-learn paketi [76] ile aynı gerçeklemeyi takip etmektedir. Scib uygulamasında değerler $(ASW + 1) / 2$ formülü ile normalize edilerek 0 ile 1 arasında ölçeklenir. ASW değerlerinin 0.5 civarında olması, hücre türlerine göre kümelerin toplandığını ancak kümeler arasında ciddi bir örtüşme olduğunu gösterir. ASW değerlerinin 0'a yakın olması, aynı türdeki hücreler arasında belirgin bir kümelenme olmadığını ve verinin tamamen karışık olduğunu ifade eder.

2.2.5.2. Düzeltilmiş Rand İndeksi (ARI)

Düzeltilmiş rand indeksi metrik, NMI gibi iki kümeleme arasındaki hizalanmayı, hizalanan hücre kümelerinde bir arada bulunan hücre çiftlerini sayarak hesaplar. Ancak, ARI ayrıca farklı kümelerdeki hücreler arasındaki ortak çiftleri de dikkate alır. Bu nedenle, ARI farklı türdeki hücre kümelerinin ayrımını da yansıtmaktadır. Bu durum NMI metriğinde mevcut değildir.

ARI metriği, Scib *Metrics* modülündeki *ari* fonksiyonu ile hesaplanır. Yöntem, Scikit-learn'den gelen *adjusted_rand_score* fonksiyonunu kullanır; bu fonksiyon, gerçek etiketler ve tahmin edilen etiketleri girdiler olarak alır ve ARI skorunu hesaplar. NMI metriğinde açıklandığı gibi, hücre türleri gerçek etiketler ve Leiden algoritmasının tahmin edilen kümeleme etiketleri kullanılarak hesaplanmıştır. Kümelemeler arasında eşleşme olmadığında metrik 0'a yakın değerler alır. Bu durum yapılan deneylerde aynı türdeki hücreler arasında benzerlik olmaksızın rastgele gömmelerle sonuçlanan bir bütünleştirme anlamına gelmektedir. Performans sonuçlarının 1'e yakın olması, aynı tür hücreler arasında yüksek düzeyde kümelenme ve farklı tür hücre kümeleri arasında güçlü ayrım olduğunu gösterir.

2.2.5.3. iLISI ve cLISI

iLISI (ing. integration LISI), LISI skorunu toplu etiketlerine dayalı olarak hesaplayarak partilerin karışım derecesini ölçerek bütünleştirme başarısını değerlendirir. Yüksek iLISI skoru, farklı partiler arasındaki karışımın daha iyi olduğunu ve toplu etkilerinin başarılı bir şekilde giderildiğini gösterir.

cLISI (ing. cell-type LISI), hücre tipi etiketleri kullanılarak hesaplanır ve benzer hücrelerin ne derece saflıkta bir araya getirilebildiğini ölçer. Bu metrik, bütünleştirme sonrası biyolojik olarak anlamlı kümelerin korunmasına bir ölçüt sağlar. cLISI skoru iLISI'den farklı olarak, partilerdeki hücre popülasyonlarının ne kadar iyi korunduğunu gösterir; böylece aşırı bütünleştirmenin biyolojik farklılıkları ortadan kaldırıp kaldırmadığını analiz etmeyi sağlar. Her iki metrikte Harmony paketinde [37] sunulduğu hali ile varsayılan parametreler kullanılarak hesaplanmıştır.

2.2.5.4. Yüksek Değişken Protein (HVP) Skoru

HVP (yüksek değişkenlik gösteren proteinler) skoru, biyolojik sinyalin bütünleştirme sonrası değişkenliğinin ne kadar korunduğunu değerlendirmektedir. Protein sayısı oldukça az olduğu için HVP hesaplanırken tüm proteinler yüksek değişken olarak kabul edilmiştir. HVP skoru, her parti için hesaplanan HVP örtüşme katsayılarının ortalaması alınarak elde edilmiştir. Bu skor, veri bütünleştirmenin biyolojik çeşitliliği nasıl etkilediğini değerlendirmek için kullanılmıştır.

2.3. Sonuçlar

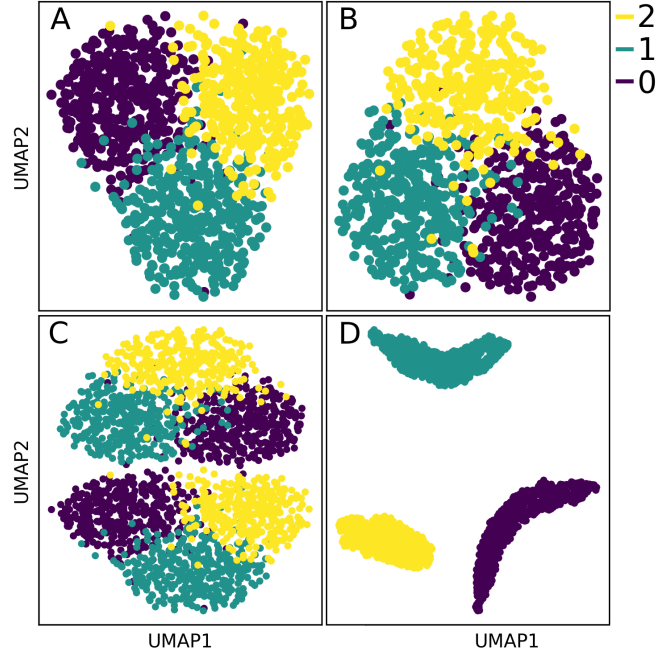
2.3.1. Sentetik Veri Setlerinin Bütünleştirilmesi

SCPRO-HI algoritması, önerilen hücre eşleştirme ve toplu etki giderme yöntemlerinin işlevselliğini doğrulamak için kontrollü bir ortamda test edilmiştir. Bu amaçla, her biri 1000 hücreden ve 100 protein ölçümünden oluşan iki sentetik veri seti oluşturulmuştur. Protein bollukları, her bir bileşenin farklı bir hücre tipini temsil ettiği üç bileşenli bir Gauss karışım modeli kullanılarak simüle edilmiştir. Sonuç olarak, her iki veri setinde de üç farklı hücre grubu bulunmaktadır (Şekil 2.4A ve 2.4B).

Sentetik veri setlerinde, her hücre grubunda rastgele belirlenen üç proteine farklı değerler atanmış ve diğer tüm ölçümler hücreler arasında ortak bir ortalama (μ) değer kullanılarak hesaplanmıştır. Veri setleri arasında hücre gruplarını ayırt etmek için her rastgele eşleştirilmiş hücre kümesi için üç farklı protein seçilmiştir. Belirlenen bu proteinlere ait ölçümler ilgili hücre kümesini kendi veri kümesi içinde diğer hücrelerden ayırmaktadır.

Gerçek veri setleriyle uyumlu olması için ayırt edici proteinlerin sayısı üç olarak belirlenmiştir. Ayırt edici proteinlerin bollukları, tüm hücreler için yeni bir μ değeri kullanılarak oluşturulmuştur. Bu sayede, aynı proteinlerin bollukları ile diğer hücrelerden ayrılan farklı hücre gruplarına sahip iki veri seti elde edilmiştir. Böylece, önerilen algoritmanın etkinliğini değerlendirmek için gerekli yapı sağlanmıştır.

Her veri seti için rastgele oluşturulan bir Gauss gürültüsü, toplu etkiyi simüle etmek için protein bolluk değerlerine eklenmiştir. Sonuç olarak, her biri üç aynı hücre tipini içeren iki veri seti elde edilmiştir (Şekil 2.4C). Aynı hücre tipindeki



Şekil 2.4: Sentetik veri setleri ve bütünleştirme sonuçları. **A** Veri Seti 1 UMAP gösterimi. **B** Veri Seti 2 UMAP gösterimi. **C** Bütünleştirme öncesi her iki veri setinin birleştirilmiş UMAP gösterimi. **D** Bütünleştirilmiş verinin UMAP gösterimi.

hücre gruplarının farklı merkezleri bulunmaktadır ancak aynı ayırt edici proteinleri paylaşmaktadırlar. Bütünleştirme sonuçları, SCPRO-HI algoritmasının farklı veri setlerinde aynı ayırt edici proteinlere sahip hücre gruplarını tespit etmedeki etkinliğini göstermiştir. Ayrıca, eşleştirilen hücre gruplarının bütünleştirilmiş veri setinde birlikte kümelendiği gözlemlenmiştir. Bu durum düzeltilmiş ölçüm uzayında benzer hücrelerin başarıyla yakınlştırılarak bütünleştirildiğini göstermektedir (Şekil 2.4D).

2.3.2. Türler Arası Veri Setlerinin Bütünleştirilmesi

Bu çalışmada, SCPRO-HI algoritmasının performansı, dört yüksek performanslı tek-hücre transkriptomik veri bütünleştirme aracıyla karşılaştırılarak değerlendirilmiştir: Harmony, Scanorama, fastMNN ve scVI. Ayrıca, bildiğimiz kadarıyla tek-hücre proteomik veri setlerini bütünleştirmek için özel olarak tasarlanmış tek algoritma olan MARIO algoritması da karşılaştırmaya dahil edilmiştir. Değerlendirme için, her biri 120.000 kan hücresinden ve yedi hücre tipinden oluşan dört açık kaynak veri seti kullanılmıştır: (i) H1N1 virüsüyle karşılaşan 35 sağlıklı insandan alınan [77], (ii) IFN γ ile uyarılan insanlar, (iii) IFN γ ile uyarılan rhesus makakları (Macaca mulatta), ve (iv)

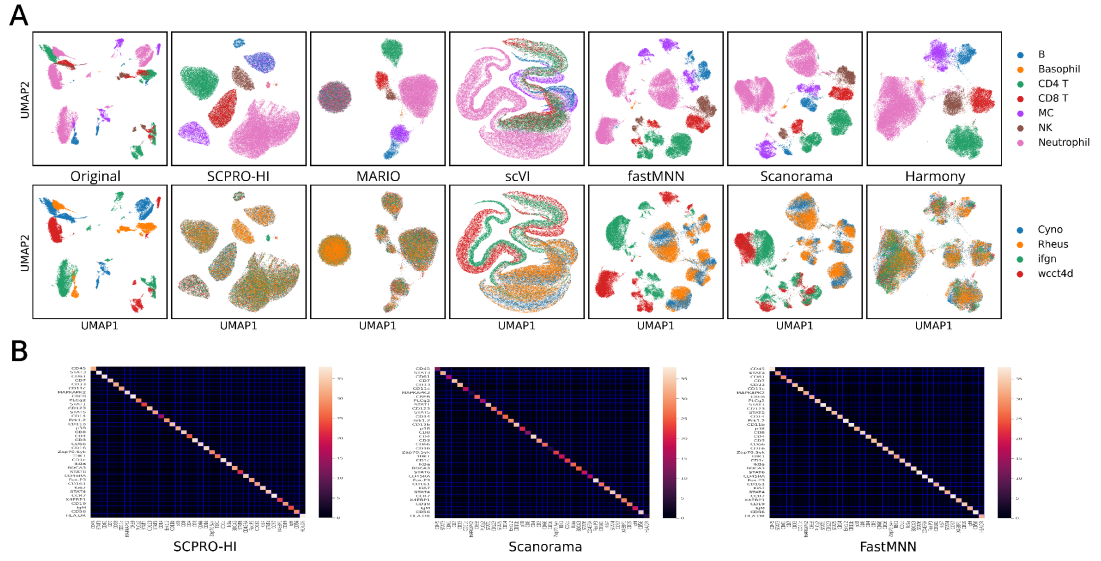
IFN γ ile uyarılan cynomolgus maymunları (*Macaca fascicularis*) [78].

Birinci veri setinde, H1N1 virüsüyle enfekte olan 35 bireyden alınan bağışıklık hücre alt kümelerindeki 40 proteinin kütle sitometri profili oluşturulmuştur. Yazarlar, bu veri setindeki örnek grupları arasında dizileme plakasına özel toplu etkileri bildirmiştir. Çoklu-tür veri setleri, 39 protein dizilenerek üretilmiştir. Ölçümler, 86 sağlıklı insan, 32 rhesus makak (*Macaca mulatta*) ve 32 cynomolgus makak (*Macaca fascicularis*) bireylerinden alınan tam kanın fosfo-akış bağışıklık sinyal profili yoluyla elde edilmiştir. Bu veri setleri, türler arasındaki biyolojik varyasyonu korurken toplu etkileri kaldırma zorluğu nedeniyle bütünleştirme algoritmaları için zorlu bir test imkânı sunmaktadır.

Şekil 2.5A'da (orijinal) yer alan iki boyutlu projeksiyonlar hem veri seti içi hem de veri setleri arasındaki biyolojik varyasyonları göstermektedir. Özellikle, H1N1 veri seti diğerlerinden belirgin bir şekilde ayrılmaktadır. Bu durum diğer veri setlerindeki insan örnekleri dikkate alındığında dahi görülmektedir. Ayrıca, insan hücrelerinin biyolojik varyasyon nedeniyle ayrıştığını, ancak maymun ve makaklardan elde edilen veri örneklerinin birbirine daha fazla karıştığını belirtmek önemlidir.

Değerlendirmelerin tamamı, dört veri seti arasında ortak olan tüm proteinler kullanılarak gerçekleştirilmiştir. MARIO algoritması için, bu yöntem özellikle parti-özel proteinlerden yararlandığı için her veri setinde bulunan tüm proteinler kullanılmıştır. Tüm veri setlerinde ortak olan 38 protein, Çizelge A.1'de listelenmiştir. CD235a proteini, rhesus makaklarda ölçülmediği için bütünleştirme sürecinden çıkarılmıştır. Benzer şekilde, CD27 proteini yalnızca H1N1 virüsüne maruz kalan insanlarda ölçüldüğü için kullanılmamıştır. Tüm algoritmalar, yazarları tarafından açıklandığı şekilde, açık erişim Python paketleri kullanılarak test edilmiştir.

SCPRO-HI algoritması, veri setlerini toplamda 32 hücre kümesine ayırmıştır. Her bir küme için, kümedeki hücrelerin çoğunluğunun gerçek hücre türü etiketine dayalı bir hücre tipi atanmıştır. Bu etiketler yalnızca küme eşleştirme performansını değerlendirmek amacıyla kullanılmıştır; entegrasyon için yalnızca protein bollukları kullanılmış, hücreler hakkında önceden elde edilmiş herhangi bir bilgi kullanılmamıştır. Değerlendirme sonuçları, küme eşleştirme algoritmamızın verimliliğini göstermektedir. Sonuçlara göre 43 küme çapa bağlantısından 37'si aynı hücre tipine sahip kümeler arasında oluşturulmuştur.



Şekil 2.5: Tek-hücre proteomik veri setlerinde algoritmaların bütünleştirme sonuçları. **A** Algoritmaların bütünleştirilmiş ölçümlerinin UMAP gösterimleri. **B** Proteinlere ait yüksek değişkenlik benzerliklerini gösteren ısı haritaları.

Kümeler, 12 bağlı bileşen oluşturmıştır. Bu bileşenlerden yedisi iki veya daha fazla küme içerirken, beşi herhangi bir bağlı bileşendeki küme ile yeterli benzerlik göstermeyen izole kümelerden oluşmaktadır. İkinci bir şans verilmesine rağmen, bu kümeler bağlanamamıştır. Bu izole kümeler, veri setlerindeki nispeten farklı hücre tiplerini içerdikleri için daha ayrıntılı bir şekilde analiz edilebilir.

Algoritmaların performans sonuçları, beş yaygın kullanılan metrikle değerlendirilmiştir: Silhouette Skoru (SS), Düzeltilmiş Rand İndeksi (ARI), iLISI skoru, cLISI skoru ve yüksek değişken protein (HVP) koruma skoru. Deneylere ait performans sonuçları Çizelge 2.1’de sunulmuş, bütünleştirme sonuçları ise Şekil 2.5A’da gösterilmiştir. Altta ki renklendirme parti etiketleri, üstteki renklendirme hücre tiplerine göre yapılmıştır. Sonuçlar, SCPRO-HI algoritmasının farklı yığınlardan gelen aynı tip hücreleri bir araya getirerek hücreleri bütünleştirmedeki etkinliğini vurgulamaktadır. MARIO, Harmony ve Scanorama algoritmaları da parti bütünleştirme ve benzer hücrelerin yakın kümelenmesi açısından iyi performans sergilemektedir. Buna karşılık, fastMNN algoritması veri setlerini düzgün şekilde bütünleştirememiş ve farklı kümelerdeki aynı tip hücreleri bir araya getirememiştir. scVI algoritması da parti karışımında zorlansa da dönüştürülmüş uzayda aynı tip hücrelerin yakın kümelenmesi açısından potansiyel bir başarıyı göstermektedir.

Silhouette skoru iki farklı şekilde hesaplanmıştır. İlk olarak, küme merkezlerini belirlemek için gerçek hücre tipleri kullanılmıştır. İkinci olarak, varsayılan yaklaşım olarak hesaplamalı olarak tespit edilen küme merkezleri kullanılmıştır. Gerçek etiketler dikkate alındığında, algoritmamızın silhouette skoru, ikinci en iyi yöntemin performansının neredeyse iki katı kadar yüksektir. Algoritmamızın silhouette skoru, aynı hücre tipine sahip hücrelerin düzeltilmiş uzayda benzer olduğunu göstermektedir. Ayrıca, SCPRO-HI algoritması, hücre kümesi etiketleri kullanılarak hesaplanan silhouette skorlarında üstün bir performans göstermektedir. Küme etiketleri, K-ortalamlar kümeleme algoritması kullanılarak üretilmiştir. Küme sayısı k , partilerdeki hücre tiplerinin yedi olması sebebiyle 7 olarak belirlenmiştir. K-ortalamlar kümeleme algoritması kullanılarak üretilen küme etiketleri ile hesaplanan silhouette skorları, hücre kümelerinin düzeltilmiş uzayda iyi ayrıldığını göstermektedir.

Düzeltilmiş Rand İndeksi (ARI), gerçek hücre etiketleri ile hesaplanan hücre tipleri arasındaki uyumu değerlendirir. Her hücrenin entegre uzaydaki en yakın üç komşusunun gerçek hücre etiketleri kullanılarak, hücreye komşularının çoğunluk etiketine göre yeni bir etiket atanır. Scanorama, Harmony ve SCPRO-HI algoritmalarının sonuçlarının yakın olması benzer hücreleri gruplamada hepsinin başarılı olduğunu göstermektedir.

iLISI ve cLISI metrikleri sırasıyla bütünleştirilen veri setlerindeki parti karışımlarının ve hücre tiplerinin küme kompozisyonunun kalitesini ölçer. Geliştirilen algoritmanın farklı veri setlerindeki hücreleri karıştırabilme yeteneği en iyi iLISI skorunun elde edilmesini

Çizelge 2.1: Türler arası veri setlerinin bütünleştirme performans sonuçları.

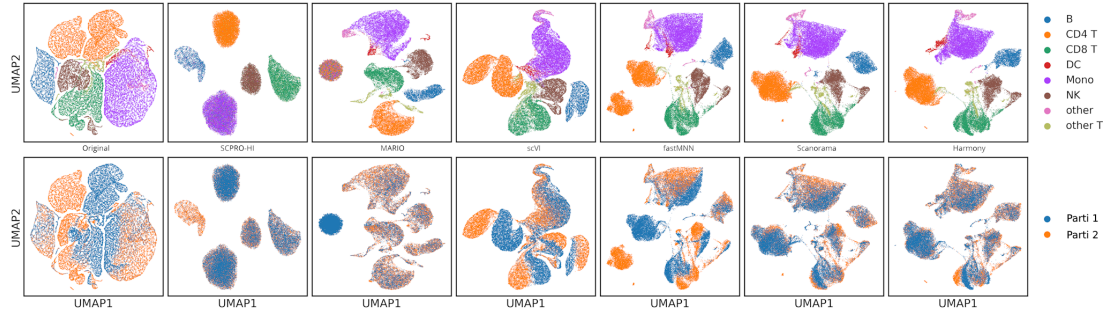
Yöntem	SS(hücre tipi)	SS(küme)	ARI	iLISI	cLISI	HVP
Ham veri	0.19	0.20	0.91	0.01	0.99	-
fastMNN	0.19	0.22	0.92	0.01	0.99	0.79
Harmony	0.26	0.19	0.96	0.40	0.77	-
Scanorama	0.23	0.23	0.97	0.22	0.96	0.64
scVI	0.11	0.54	0.88	0.16	0.94	-
MARIO	0.15	0.41	0.63	0.61	0.53	-
SCPRO-HI	0.46	0.63	0.96	0.67	0.89	0.72

sağlamıştır. Ancak, bütünleştirilen verideki kümelerde, çoğunluk hücre tiplerine kıyasla az sayıda da olsa farklı hücre tipi bulunması cLISI skorunu nispeten düşürmektedir. Bu az sayıdaki yanlış eşleşmiş hücrelerin kümülatif etkisi büyüyerek, farklı partileri karıştırmakta güçlük çeken yöntemlere göre daha düşük cLISI skorlarına yol açmaktadır. Biyolojik anlamın korunması, HVP metriği ile değerlendirilmiştir. Luecken tarafından önerilen HVG koruma stratejisi [44] aynen uygulanmıştır. Algoritmamız, proteinlerin değişkenliğini Scanorama'ya göre yaklaşık %13 daha iyi korumuştur. Şekil 2.5B'de normalleştirilmiş varyanslarına göre sıralanan protein değişkenlik derecelerinin, gürültülü veriler ve düzeltilmiş ölçümler arasındaki benzerliklerini gösteren ısı haritaları verilmiştir. Sadece proteinlerin öz-ilişkileri, ham ve bütünleştirilmiş verilerdeki önem sıraları arasındaki benzerlikler renklendirilmiştir. Daha açık hücreler, protein varyansının daha iyi korunduğunu ifade etmektedir. Tüm algoritmalar tatmin edici performans sergilerken, fastMNN algoritması neredeyse doğrusal düzeltme stratejisi nedeniyle protein varyansında en az farklılığı göstermiştir. SCPRO-HI'a ait sonuçlar incelendiğinde, algoritmanın proteinlerin değişkenliğini büyük ölçüde koruduğunu göstermektedir. FastMNN algoritması, neredeyse doğrusal düzeltme stratejisi nedeniyle HVG koruma açısından büyük bir başarı göstermektedir. Öte yandan, scVI, Harmony ve MARIO algoritmaları protein bolluklarını yeni düşük boyutlu bir uzaya taşıdıkları için bu metrik bu yöntemler için hesaplanamamıştır.

2.3.3. İnsan PBMC Veri Setlerinin Bütünleştirilmesi

Karşılaştırma algoritmaları, Hao ve arkadaşlarının [72] çalışmasında sunulan periferel kan mononükleer hücre (PBMC) veri setleri kullanılarak test edilmiştir. Toplu etkinin büyüklüğünün, aynı türden elde edilen veri setlerine kıyasla, türler arası veri setlerinde daha yüksek olma eğilimindedir. Bu deney ile algoritmamızın farklı toplu etki seviyelerine karşı genelleme yeteneği test edilmiştir.

PBMC veri setleri, HIV ile enfekte olmuş sekiz donörden alınan örneklerden elde edilmiştir. Örnekler iki ayrı parti olarak bölünmüştür. Toplamda, veri setleri 161,764 hücre içermekte ve 228 proteine ait ölçümleri kapsamaktadır. Veri setlerinde yedi hücre tipi bulunmaktadır. PBMC veri setlerinin entegrasyonu sırasında, türler arası veri setlerinin entegrasyonu için kullanılan metodolojinin aynısı uygulanarak tüm



Şekil 2.6: Tek-hücre insan PBMC proteomik veri setlerinin bütünleştirme sonuçları.

karşılaştırma yöntemlerinin performansı değerlendirilmiştir.

Deneylere ait nümerik performans sonuçları Çizelge 2.2’de ve bütünleşik verinin UMAP görselleri Şekil 2.6’da sunulmuştur. Geliştirilen algoritma ve scVI, insan PBMC veri setlerini entegre ederken siluet skoru açısından eşit derecede başarılıdır. Etkileyici bir şekilde, her iki algoritma da ikinci en iyi yöntemle göre siluet skorunda %20’lik bir iyileşme sağlamıştır. Küme sayısı k , partilerde bulunan yedi farklı hücre tipi temel alınarak 7 olarak belirlenmiştir. Ayrıca, sonuçlar algoritmamızın aynı türden hücreleri bir araya getirirken, düzeltilmiş protein ölçümleri temelinde onları diğer hücre türlerinden etkili bir şekilde ayırabildiğini göstermektedir. Bu başarı, tahmin edilen hücre tipleri dikkate alındığında siluet skorunda %15’lik bir iyileşme sağlamaktadır.

Önerilen algoritma, düzeltilmiş rand indeksi (ARI) skoru açısından rekabetçi bir performans sergilemekle beraber rakiplerine göre daha düşük bir performans göstermektedir. Bu fark, küçük hücre gruplarının yanlış kümelenerek farklı türden daha büyük hücre grupları ile bir arada bütünleştirilmesinden kaynaklanmaktadır. Kümeleme hatası, bu küçük grupların gürültülü protein ölçüm alanında diğer tür hücre gruplarına yakın olduğu durumlarda hücrelerin yanlış etiketlenmesinden kaynaklanmaktadır.

Algoritmamızın iLISI metriğindeki kayda değer başarısı, PBMC veri setleri üzerinde gerçekleştirilen deneyler boyunca tutarlılığını korumuş ve Harmony algoritmasını %25 oranında geride bırakmıştır. Ancak, cLISI metriğinde algoritmamız, Scanorama ile birlikte ikinci sırayı paylaşmaktadır ve bu sonuçlar önceki deneysel sonuçlardan sapmaktadır. Dikkate değer bir şekilde, fastMNN ve scVI algoritmaları cLISI metriklerinde üstün bir performans sergilemekle birlikte, sırasıyla 0.04 ve 0.03 düşük iLISI skorları ile partileri etkili bir şekilde entegre etmede başarısız olmuştur.

İnsan PBMC veri setlerinin bütünleştirilmesi üzerine yapılan deney, algoritmamızın farklı toplu etki seviyelerine sahip veri setlerinde dayanıklılığını göstermektedir. Bununla birlikte, türler arası veri setlerin bütünleştirilmesinde umut vaat eden bir performans sergileyen MARIO algoritmasının, PBMC veri setinde bu performansını sürdüremediği görülmüştür fakat MARIO'nun genel performansı proteomik veri setlerinin entegrasyonu için umut verici kalmaktadır. Öte yandan, Harmony ve Scanorama yöntemleri her iki veri setinde benzer bir performans sergileyerek dayanıklılıklarını kanıtlamıştır.

FastMNN algoritması, çoğu metrikte artan bir performans gösterirken, karşılaştırma algoritmaları arasında tutarlı bir sıralamaya sahiptir. scVI algoritması bu deneyde önemli bir performans artışı göstermiştir; ancak, düşük iLISI skoru ile kanıtlandığı üzere partileri etkili bir şekilde karıştırmakta zorlanmıştır.

Genel olarak, algoritmamız farklı veri setlerinde tutarlılığını koruyarak, farklı toplu etki seviyelerini ele alabilme yeteneğini kanıtlamıştır. Bazı yöntemler belirli yönlerde iyileşme veya sınırlamalar gösterirken, algoritmamız PBMC veri setlerinin bütünleştirilmesi için güçlü bir çözüm yöntemi olarak öne çıkmıştır.

2.3.4. Yüksek Boyutlu Veri Setlerinin Bütünleştirilmesi

Harmony ve Scanorama algoritmaları, SCPRO-HI algoritması ile karşılaştırmalı olarak scRNA-seq veri setleri üzerinde test edilmiştir. Bu karşılaştırma, algoritmamızın

Çizelge 2.2: PBMC veri setlerinin bütünleştirme performans sonuçları.

Yöntem	SS(hücre tipi)	SS(küme)	ARI	iLISI	cLISI	HVP
Ham veri	0.32	0.52	0.90	0.03	0.99	-
fastMNN	0.43	0.52	0.96	0.04	0.99	0.90
Harmony	0.45	0.54	0.97	0.70	0.64	-
Scanorama	0.45	0.53	0.97	0.65	0.66	0.53
scVI	0.55	0.62	0.90	0.03	0.99	-
MARIO	0.17	0.35	0.54	0.62	0.63	-
SCPRO-HI	0.55	0.65	0.92	0.88	0.65	0.65

yüksek boyutlu veri setlerindeki başarımını değerlendirmek için yapılmıştır. Harmony ve Scanorama algoritmaları, birçok çalışmada scRNA-seq veri setlerini bütünleştirmede sergiledikleri başarımlar nedeniyle seçilmiştir. MARIO algoritması, yalnızca tek-hücre proteomik veri setlerine yönelik olarak tasarlandığı ve yüksek boyutlu transkriptomik veri setleri için herhangi bir geliştirmesi bulunmadığı için bu testlere dahil edilmemiştir. Ayrıca, Harmony ve Scanorama algoritmalarının diğer çalışmalarda daha üstün performans sergilediği kanıtlandığından, fastMNN algoritması testlerimize dahil edilmemiştir. 16 çekirdekli CPU ve 64 GB RAM içeren test ortamımızda scVI algoritması test edilemediği için deneylere dahil edilememiştir.

Karşılaştırmada kullanılan dört veri seti, 10 farklı hücre tipinde 56399 fare beyin hücresinden elde edilen 2000 gen ifadesini içermektedir [79]. Testlerimizde, önemli gen sayısı 100 olarak belirlenmiştir. Bu sayı, antikor tabanlı tek-hücre proteomik veri setlerindeki protein ölçümlerine uygun olacak şekilde seçilmiştir. Sonuç olarak, her bir bağlı bileşen için 19 regresyon modeli bulunmaktadır. Beklendiği üzere, küme eşleştirme kalitesi azalmış ve 39 küme çapa bağlantısının 25'i aynı türdeki kümeler arasında kurulmuştur. Ancak, veri setlerinin farklı türlerden ortak hücre kümeleri içermesi nedeniyle eşleştirme algoritması hala ümit vericidir. Algoritmaların performans sonuçları Çizelge 2.3'te sunulmaktadır.

Genişletme algoritmamız genel olarak umut verici performans sonuçları sergilemiştir ve iki metrikte en yüksek puanları elde etmiştir. Harmony algoritması, gerçek hücre tiplerine dayalı siluet skorunda en iyi performansı göstermiştir, ancak Scanorama algoritması orijinal verilerin siluet skorunda iyileşme sağlayamamıştır. Algoritmamız, iyi performans göstermiş ancak proteomik veri setleri üzerindeki deneylerde görülen belirgin iyileştirmeleri sergileyememiştir. Hesaplamalı olarak üretilen küme etiketlerine dayalı siluet skoru, algoritmamızın iyi kümelenmiş düzeltilmiş ölçümler oluşturduğunu göstermektedir. Tüm üç algoritma ARI skorunu iyileştirmiş ve benzer sonuçlar elde etmiştir. iLISI skoru, SCPRO-HI algoritmasının partileri karıştırmadaki kalitesini yansıtmaktadır. Scanorama algoritması, HVG'leri korumada algoritmamızdan daha iyi performans göstermiştir, çünkü algoritmamız önemli genler ve gen setleri arasındaki ilişkiyi modelleme sırasında bazı bilgileri kaybetmektedir.

Algoritma, seyrek ve yüksek boyutlu tek-hücre proteomik verilerini işleme etkinliğini değerlendirmek için kütle spektrometresi tabanlı tek-hücre proteomik veri setleri ile

de test edilmiştir. Bu testlerde, Leduc ve arkadaşları [80] tarafından pSCoPE yöntemi kullanılarak oluşturulan veri seti kullanılmıştır. Bu veri seti, her biri hem monositler (U937 hücreleri) hem de melanom hücrelerinden (WM989-A6-G3) ölçümler içeren iki parti içermektedir. İlk parti 497 melanom hücresi ve 502 monosit, ikinci parti ise 271 melanom hücresi ve 273 monosit içermektedir. Her iki veri setinde de 2,844 protein için ölçümler elde edilmiştir fakat veri seti önemli ölçüde seyrek.

Önceki transkriptomik veri seti deneylerinde kullanılan deneysel kurulum tekrarlanmıştır. En önemli 100 protein seçilerek 29 farklı alt kümeye ayrılmıştır. Algoritma, her iki veri setinde hücreleri etkili bir şekilde iki gruba kümelemiş ve aynı hücre türlerindeki kümeleri başarıyla eşleştirerek iki bağlı bileşen oluşturmuştur.

Algoritma, daha önce transkriptomik veri setlerinin değerlendirilmesinde açıklandığı üzere Harmony ve Scanorama algoritmaları ile karşılaştırılmıştır. Bütünleştirme performans sonuçları Çizelge 2.4'te sunulmaktadır. Bu sonuçlar, transkriptomik veri setlerinin entegrasyonu bağlamında elde edilen bulgularla tutarlıdır.

Algoritmamız, yüksek boyutlu seyrek veri setleri içinde hücre kümelerini başarılı bir şekilde eşleştirdiğini göstererek dayanıklı bir eşleştirme stratejisi sergilemektedir. Ek olarak, yüksek siluet skorları, algoritmamızın düzeltilmiş uzayda benzer ölçümler üreterek hücreleri bağlı bileşenlere etkili bir şekilde grupladığını göstermektedir.

Ancak sonuçlar, algoritmamızın yüksek değişken proteinleri korumada Scanorama algoritmasına kıyasla daha düşük performans sergilediğini de ortaya koymaktadır. Bununla birlikte, %57'lik koruma oranı, SCPRO-HI algoritmasının birincil odak noktası yüksek boyutlu veri setleri olmadığı için sonuçlar kabul edilebilir düzeydedir.

Performans sonuçları, SCPRO-HI algoritmasının, tek-hücre transkriptomik veri setleri

Çizelge 2.3: Transkriptomik veri setlerinin bütünleştirme performans sonuçları.

Yöntem	SS(hücre tipi)	SS(küme)	ARI	iLISI	cLISI	HVP
Ham veri	0.13	0.22	0.92	0.01	1.0	-
Harmony	0.21	0.39	0.95	0.19	0.96	-
Scanorama	0.13	0.24	0.97	0.05	1.0	0.75
SCPRO-HI	0.16	0.47	0.96	0.31	1.0	0.68

ve kütle spektrometresi tabanlı tek-hücre proteomik veri setleri üzerinde, mevcut yatay entegrasyon algoritmaları ile rekabetçi sonuçlar elde ettiğini göstermektedir. Bununla birlikte, algoritmanın yüksek boyutlu veri setleri üzerindeki performansının, düzeltilmiş önemli genlerin geri kalan genlerle ilişkisini daha iyi modelleyen çalışmalar ile daha da geliştirilebileceğini düşünmekteyiz.

2.4. Sonuç

Bu bölümde, tek-hücre proteomik veri setlerinin yatay bütünleştirilmesi için geliştirilen SCPRO-HI algoritması sunulmuştur. Algoritma, toplu etki düzeltmesi için hiyerarşik hücre çapalama yöntemi ve özgün bir varyasyonel oto-kodlayıcı modelini birleştirmektedir. Algoritmanın performansı hem sentetik hem de gerçek dünya tek-hücre proteomik veri setlerinde değerlendirilmiş ve mevcut yöntemlerle karşılaştırılmıştır.

Sonuçlar, SCPRO-HI algoritmasının nümerik metriklerde ve biyolojik anlamın korunması açısından rakiplerini geride bıraktığını göstermektedir. Algoritma, diğer algoritmalarla kıyasla %75 daha yüksek bir siluet skoru elde etmekte ve biyolojik anlamı %13 oranında daha iyi korumaktadır. Ayrıca, algoritma yüksek boyutlu transkriptomik veri setlerine uygulandığında rekabetçi bir performans göstermiştir ve bu durum diğer omik veri türlerini entegre etme potansiyelini ortaya koymaktadır.

Tek-hücre proteomik veri setlerinin entegrasyonu, kapsamlı ve güvenilir ölçümler elde etmek için kritik öneme sahiptir. Toplu etkileri ortadan kaldırarak algoritma, veri setlerinin entegrasyon kalitesini artırmakta ve hücre tipi tanımlaması, yaşlanma analizi ve trajektori çıkarımı gibi ileri analiz görevlerinde daha doğru sonuçlar elde edilmesini sağlamaktadır. Algoritmada kullanılan hiyerarşik hücre bağlama tekniği, hücreleri

Çizelge 2.4: Kütle spektrometrisi tabanlı proteomik veri setlerinin bütünleştirme performans sonuçları.

Yöntem	SS(hücre tipi)	SS(küme)	ARI	HVP
Ham veri	0.45	0.41	0.98	-
Harmony	0.67	0.66	0.97	-
Scanorama	0.47	0.36	0.97	0.68
SCPRO-HI	0.92	0.87	0.92	0.57

eşleştirmek ve hücre kümeleri oluşturmak için ayırt edici proteinlerin benzerliğini dikkate almakta ve hücre kümelerini ayırmada önemli katkı sağlayan proteinlerin önemini hesaba katmaktadır. Bu yaklaşım, hücre eşleştirme doğruluğunu artırarak daha güvenilir entegrasyon sonuçlarına yol açmaktadır.

Varyasyonel oto-kodlayıcı model algoritmada pozitif ve negatif çapa hücrelerini temel alarak protein bolluklarını toplu etkilerinden arındırılmaktadır. Protein bolluklarını daha düşük boyutlu bir veri uzayına taşıyan diğer yöntemlerin aksine, SCPRO-HI algoritması toplu etkileri doğrudan protein bollukları üzerinden kaldırarak bireysel protein düzeyinde biyolojik analiz gerektiren ileri analizlere olanak tanımaktadır.

Algoritma, hem sentetik hem de gerçek dünya veri setlerinde uygulanabilirliğini göstermiştir. Sentetik veri setleri, önerilen hücre çapalama ve toplu etki kaldırma yöntemlerinin işlevselliğini değerlendirmemizi sağlamış ve algoritmanın farklı veri setlerindeki aynı ayırt edici proteinlere sahip hücre gruplarını başarıyla bütünleştirebildiğini göstermiştir. PBMC veri setleri ve fare beyni veri setleri gibi tek-hücre proteomik veri setleri, algoritmanın gerçek dünya veri setlerindeki performansını değerlendirmek için kullanılmıştır. Sonuçlar, algoritmanın değişen seviyelerde toplu etkilerle başa çıkmadaki dayanıklılığını ve diğer yöntemlere kıyasla rekabetçi performansını ortaya koymaktadır.

SCPRO-HI algoritmasının kullanılabilirliği, yalnızca biyolojik veri setleriyle sınırlı değildir. Temel olarak yatay bütünleştirme prensiplerine dayanan bu yöntem, aynı türdeki fakat farklı kaynaklardan elde edilen veriler arasındaki uyumsuzlukları gidermek için geniş bir uygulama alanına sahiptir. Bu durum, sadece biyolojik çalışmalarda değil, aynı zamanda görüntü işleme, hava durumu analizi gibi disiplinlerde farklı kaynaklardan elde edilen verilerin tutarlı hale getirilmesi için de kullanılabilir. Örneğin, farklı sensörlerden elde edilen görüntü verilerinin hizalanması veya çeşitli anket çalışmalarından gelen verilerin birleştirilmesi gibi alanlarda SCPRO-HI, veri bütünlüğünü artırmak ve anlamlı kümelenmeler oluşturmak için etkili bir araç olarak değerlendirilebilir.

SCPRO-HI algoritması umut verici sonuçlar gösterse de, bazı sınırlamalar ve geliştirme alanlarına sahiptir. Algoritma, genellikle onlarca veya yüzlerce protein içeren veri setlerini bütünleştirmeye odaklanmaktadır. Öte yandan, binlerce gene sahip

yüksek boyutlu veri setleriyle başa çıkabilecek şekilde algoritmanın genişletilmesi, uygulanabilirliğini artıracaktır. Ayrıca, yüksek boyutlu transkriptomik veri setlerindeki yüksek değişken genlerin daha iyi korunmasının önemli genler ve gen setleri arasında daha iyi bir ilişki modellemesi ile mümkün olacağı öngörülmektedir.

Sonuç olarak, sunulan SCPRO-HI algoritması, tek-hücre proteomik veri setlerinin entegrasyonu için etkili bir yaklaşım sağlamaktadır. Yenilikçi hiyerarşik hücre çapalama tekniği ve varyasyonel oto-kodlayıcı modeli, hücre eşleştirme ve toplu etki düzeltme zorluklarının üstesinden başarıyla gelmektedir. Algoritma, hesaplama metrikleri ve biyolojik anlamın korunması açısından mevcut yöntemleri geride bırakmakta ve tek-hücre proteomiklerdeki aşağı akış analizleri iyileştirme geliştirme potansiyelini ortaya koymaktadır.

2.5. Veri ve Kod Erişilebilirliği

Türler arası CYTOF veri setlerine, <https://flowrepository.org> adresinde şu kimlik bilgileri aracılığıyla erişilebilir: “FR-FCM-Z2NZ” [77] ve “FR-FCM-Z2ZY” [78]. PBMC veri seti (GSE164378), Human Cell Atlas veri tabanından ”humanBloodCiteSeq2” kimliği ile temin edilebilir. Transkriptomik veri seti, şu GitHub projesinde mevcuttur: <https://github.com/theislab/scarches-reproducibility>. Kütle spektrometrisine dayalı tek-hücre proteomik veri setine şu bağlantıdan ulaşılabilir: https://sc.slavovlab.net/Leduc_et_al_2022. Bu çalışmada kullanılan tüm kod ve verilere, şu bağlantıdan erişilebilen GitHub projemiz üzerinden ulaşabilirsiniz: <https://github.com/single-cell-proteomic/SCPRO-HI>.

Veri setleri, tek-hücre çalışmalarına oldukça uygun olan açıklamalı veri (AnnData – H5AD) formatında sağlanmıştır. Her veri seti, orijinal çalışmalardan alınan tüm hücre örnekleri için parti kimliği ve hücre tipi bilgilerini içermektedir. PBMC ve türler arası deneyler için birleştirilmiş veri setleri kullanılmış ve bu veri setlerine ek olarak, bunları oluşturan bireysel veri setleri de proje sayfasında mevcuttur.

3. ÇOKLU-OMİK VERİ BÜTÜNLEŞTİRME

3.1. Giriş

Farklı omik katmanlarda bulunan biyolojik moleküller arasındaki düzenleyici dinamikler henüz tam olarak anlaşılamamıştır. Bu karmaşık etkileşimler genellikle doğrusal olmayan bir yapıya sahiptir ve hücre tipleri arasında önemli ölçüde farklılık gösterir [81]. Tek-hücre teknolojilerindeki son gelişmeler, aynı hücrelerde birden fazla omik katmanın eş zamanlı (eşleştirilmiş) olarak dizilenmesini mümkün kılmıştır [82]. Bu tip veriler hücresel ayrışma hakkında daha derin bilgiler sunmaktadır. Çoklu-omik verilerin bütünleştirilmesi ile bu moleküllerin belirli hücre tiplerinde birlikte değişim analizleri kolaylaşabilmektedir. Dahası, bu örüntülerin tersine analizi ile farklı tipte hücreleri daha iyi ayırt etme olanağı sağlanmaktadır [72]. Ayrıca, bu bütünleşik analizler ile aynı hücre tipi içinde özelleşmiş alt kümeleri ortaya çıkarmak ve daha yüksek hücresel heterojenlik yakalamak mümkün hale gelmektedir.

Farklı modalitelerden hücre görünümünün bütünleştirilmesi (dikey), aynı omik türünde farklı veri setlerinin bütünleştirilmesine (yatay) kıyasla farklı zorluklara sahiptir [83]. Yatay bütünleştirme, veri setleri arasındaki teknik gürültüleri en aza indirmeye odaklanırken, dikey veri bütünleştirmede amaç hücrelerin farklı modalite görünümleri arasındaki en iyi uzlaşmayı sağlayabilmektir. Çoklu-omik verilerden bütünleşik bir hücre temsili oluşturmak, her bir modalitenin kendine has hücreler arası ilişkileri sebebiyle zordur. Her hücre için hangi modalitedeki ilişkinin en doğru olduğunu belirlemek karmaşık bir hesaplamalı problemdir [72]. Problemin zorluğu modaliteler arasında hücre eşleşme bilgisi olmadığında daha da artmaktadır [84].

Tek-hücre çoklu-omik veri bütünleştirme probleminin çözümü için çeşitli yöntemler geliştirilmiştir [85]. Bu yaklaşımlar istatistiksel modelleme [86], benzerlik çizgelerinin bütünleştirilmesi [72, 87] ve ortak bir gizli uzaya gömülmesi [81, 84, 88–95] olarak kategorilere ayrılmaktadır.

İstatistiksel modelleme yöntemleri, tekil omik dağılımları arasındaki korelasyonları yansıtan yeni bir dağılım oluşturarak çoklu-omik verileri bütünleştirmektedir. Bu yöntemlerin dağılım varsayımlarına bağımlılıkları, biyolojik sistemlerin karmaşıklığını

ve deęişkenlięini yakalama kapasitelerini sınırlamaktadır.

Benzerlik çizgelerini bütünleştirme yöntemleri, her bir modalite için hücreler arasında ilgili omik ölçümlere dayalı benzerlikleri hesaplar ve ardından bunları bütünleşik bir benzerlik çizgesine bütünleştirir [72, 87]. Bu yaklaşım, omik katmanlar arasındaki uyum ve çatışmaları analiz etmek için deęerlidir. Bütünleşik benzerlikler, modalitelere özgü çizgelerle kıyaslanarak modalitelerin hücre tipine özgü katkıları deęerlendirilebilmektedir. Ancak, benzerlik çizgelerini bütünleştirmek, benzerliklerin ölçęi ve dağılımındaki farklılıklar sebebiyle zor bir problemidir. Yüksek performanslı bir bütünleştirme algoritması modaliteler arasındaki sayısal dengesizlikler nedeniyle herhangi bir modalitenin bütünleştirme işlemini domine etmesini önlemelidir. Ek olarak, bütünleşik hücre gömmelerinin oluşturulurken, bütünleşik benzerliklerin korunması ileri analizlerde önemli bilgi kayıplarının önlenmesi açısından önemlidir.

Modalitelerin ortak bir gizli uzaya gömülmesi için negatif olmayan matris faktörizasyonu (ing. Non-positive Matrix Factorization - NMF) [88–90], kanonik korelasyon analizi (ing. Canonical Correlation Analysis - CCA) [91] ve derin öğrenme yöntemleri [81, 84, 92–95] kullanılmaktadır. NMF teknikleri, tek-hücre çoklu-omik veri setleri arasında paylaşılan ve veri setine özgü faktörleri tanımlayan düşük boyutlu temsiller üretmektedir. NMF, modalite katkılarını yorumlamak için yararlı olsa da, doğrusal yapısı gereęi doğrusal olmayan ilişkileri yakalamada düşük başarımlı gösterebilmektedir. CCA, deęişkenlerin doğrusal kombinasyonlarını tanımlayarak veri setlerini bütünleştirir. Bu kombinasyonlar modaliteler arası korelasyonları ortaya çıkararak veri setleri içindeki ve arasındaki deęişkenlięi açıklamaktadır. Bu durum aynı zamanda CCA'nın bütünleştirme performansını modalite dağılımlarındaki deęişkenliklerle sınırlamaktadır.

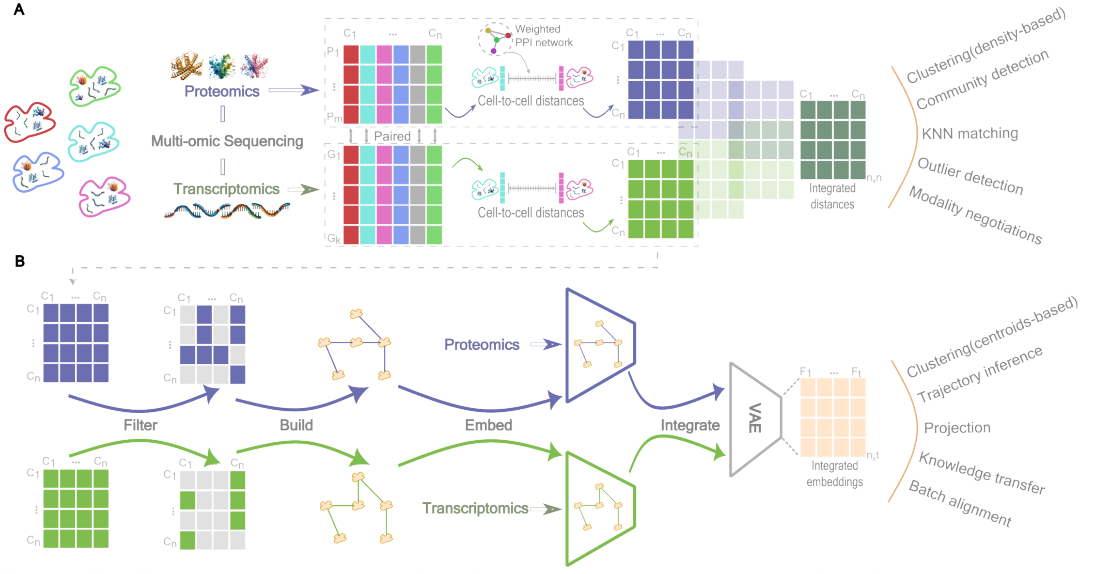
Derin öğrenme yöntemleri, modalitelerin ortak bir gizli uzayda bütünleştirilmesi için güçlü bütünleştirme araçları olarak öne çıkmaktadır [96]. Bu yöntemler, oto-kodlayıcılar [81, 92, 93], evrişimli sinir aęları [94], üretken çekişmeli aęlar [84] ve çizge sinir aęları [95] gibi çeşitli mimarileri kullanır. Derin öğrenme yöntemleri yapıları gereęi modalitelere has öznelilikler arasındaki doğrusal olmayan bağlantıları yakalayarak, doğrusal yöntemlerin sınırlamalarını aşabilmektedir. Doğrusal olmayan ilişkileri yakalama kabiliyeti bu yöntemleri çoklu-omik veri bütünleştirme için uygun çözümler haline getirmektedir. Öte yandan, elde edilen bütünleşik gömmelerin sınırlı açıklanabilirlięi ve yorumlanabilirlięi modaliteler arası ilişkiler hakkında iç görüler

oluşturmayı zorlaştırmaktadır.

Bu bölümde, modalitelerin ortak bir gizli uzaya gömülmesi için çoklu varyasyonel çizge oto-kodlayıcı (VGAE) modeller kullanan bir benzerlik çizgesi bütünleştirme yaklaşımı olan Tek-hücre çoklu-omik dikey bütünleştirme (SCPRO-VI) yöntemi sunulmaktadır. SCPRO-VI, modalitelere özgü benzerlik çizgeleri oluşturmak için destek biyolojik bilgi kullanmaya olanak tanıyan özgün bir uzaklık metriği kullanmaktadır. Bu modalitelere özgü benzerlik çizgeleri, her bir omik türüne özgü ilişkilerin eşit temsiliyi sağlamak amacıyla hücrelerin yerel komşuluklarını modaliteler arasında normalleştirilip dengelenerek bütünleşik bir benzerlik çizgesi üretilmektedir.

Bütünleşik hücre benzerlikleri, topluluk algılama, en yakın k-komşu (k-nn) analizi, aykırı değer tespiti, hücre farklılaşmasına modalite katkılarının değerlendirilmesi ve her hücre tipi için modalite önemini analizi gibi işlemlerde kullanılabilir [72]. Yoğunluk tabanlı kümeleme algoritmaları, yalnızca hücre uzaklıklarını kullanarak hesaplama yapabilmektedir çünkü bu algoritmalar komşuluk ilişkilerine dayanmaktadır [97]. Ancak, uzaklık bilgisini gizli bir uzaya gömmek merkez-tabanlı kümeleme veya bilgi aktarımı gibi vektörel hücre temsilleri gerektiren ileri analizler için gereklidir [81]. Elde edilen bütünleşik çizge, karmaşık ve doğrusal olmayan ilişkileri yakalayıp doğrusal yöntemlerin sınırlamalarını aşan özgün bir çoklu-görünüm VGAE temelli derin öğrenme modeli tarafından bütünleşik hücre gömmeleri oluşturmak amacıyla kullanılmaktadır. Bu gömmeler, çoklu-omik bütünleşik komşulukları ileri analizlere aktarmayı kolaylaştırır [92, 98]. Önerilen yöntem bütünleşik ve modaliteye özgü benzerlik çizgelerindeki hücre ilişkilerinin analizine olanak tanıyarak, bütünleştirme süreci ve modaliteler arası ilişkiler hakkında yorumlanabilir iç görüler sunmaktadır. Bu durum SCPRO-VI'ı derin öğrenmenin gücünü biyolojik yorumlanabilirlik ile birleştiren, hücreler arasındaki açıklanabilir ilişkilerle bütünleşik gizli uzayı şekillendiren güçlü bir veri bütünleştirme aracı olarak öne çıkarmaktadır.

SCPRO-VI'nın verimliliği ve etkinliği çoklu-omik CITE-seq veri seti [72] üzerinde detaylı test edilmiştir. Dahası, geliştirilen yöntem düşük hücre tipi heterojenliğine sahip çoklu-omik CITE-seq veri seti [99] üzerinde gerçekleştirilen kapsamlı deneyler ile yüksek performanslı tek-hücre çoklu-omik veri bütünleştirme araçlarına [72, 88, 90, 92] karşı test edilmiştir. Gerçekleştirilen deneylerde proteomik ve transkriptomik



Şekil 3.1: **A** Modaliteye özgü uzaklık matrisi hesaplaması ve bütünleştirilmiş uzaklıkların oluşturulmasına dair şematik gösterimi. **B** Çoklu-görünüm varyasyonel çizge oto-kodlayıcılar kullanılarak bütünleşik hücre temsilleri oluşturma şematığı.

verilerin bütünleştirilmesine odaklanılmıştır ancak SCPRO-VI nümerik tüm eşleştirilmiş modaliteleri bütünleştirebilme kabiliyetine sahiptir.

Deneysel sonuçlar, SCPRO-VI'nin hücre tipi heterojenliğini diğer yöntemlerden daha etkili bir şekilde artırdığını göstermektedir. Ayrıca, bütünleştirilmiş hücre gömmelerinin derinlemesine analizleri, SCPRO-VI'nin diğer yöntemlerin bütünleşik ölçümlerinde ayırt edilemeyen hücre alt kümelerini tespit edebildiğini ortaya koymaktadır.

3.2. Yöntem

3.2.1. SCPRO-VI

Bütünleştirme yöntemimiz üç ana adımdan oluşmaktadır: her bir modaliteye göre hücreler arasındaki uzaklıkların hesaplanması, modalite gömme modellerinin ön eğitimi ve modalite gömmelerinin bütünleştirilmesi (Şekil 3.1). İlk olarak algoritma, her modalite için hücreler arası uzaklık matrislerini ayrı ayrı hesaplamaktadır (Şekil 3.1A). Daha sonra uzaklık matrisleri bütünleştirilerek bütünleşik uzaklıklar elde edilmektedir. Elde edilen uzaklıkları yansıtan gizli uzayları üretmek için VGAE modelleri kullanılmaktadır. VGAE modellerinin eğitiminde ihtiyaç duyulan komşuluk

matrislerini hesaplamak için modalitelere ait uzaklık matrisleri filtrelendir. VGAE modelleri, her modaliteyi aynı boyutlara sahip ayrık birer gizli veri uzayına gömer. Son olarak, modalite gömmeleri, VGAE modellerinin çıktısına bağlanan bir oto-kodlayıcı model aracılığıyla homojen ara füzyon yaklaşımı [96] ile bütünleştirilir ve bütünleşik hücre gömmeleri çıktı olarak üretilir (Şekil 3.1B). SCPRO-VI algoritmasının tüm adımları aşağıdaki alt bölümlerde detaylandırılmıştır.

3.2.1.1. Uzaklık Matrislerinin Hesaplanması

Hücreler arası uzaklığı ölçmek için özgün bir metrik önerilmiştir. P_1 , $hücre_1$ 'in öznitelik matrisi A_1 'in transpozu A_1^T ile Hadamard çarpımını, P_2 ise $hücre_2$ için aynı çarpımı ifade etmektedir. $Hücre_1$ ve $hücre_2$ arasındaki uzaklık, $|(P_1 - P_2) * W| / (|A_1| * |A_2|)$ formülü ile hesaplanmaktadır. Burada W , tüm öznitelik çiftleri arasındaki önem değerlerinin matrisidir. Elde edilen toplam, hücrelerin L_1 normları $|A_1|$ ve $|A_2|$ 'nin çarpımı ile normalize edilerek ifade seviyesi farkları göz ardı edilmektedir. Metrik aşağıdaki şekilde formüle edilmiştir:

$$uzaklık(h_1, h_2) = \frac{\sum_{i=1}^m \sum_{j=1}^m |(h_1 F_i * h_1 F_j) - (h_2 F_i * h_2 F_j)| * W_{i,j}}{|h_1| * |h_2|}. \quad (3.1)$$

Formül 3.1'de, h_1 ve h_2 sırasıyla $hücre_1$ ve $hücre_2$ 'yi ifade etmektedir. F_i ve F_j , ilgili hücredeki i . ve j . biyolojik özniteliklerdir. $W_{i,j}$, özellik i ile j arasındaki ilişki değerini ifade eden bir ağırlık faktörüdür. Hücre vektörlerinin L_1 normları $|h_1|$ ve $|h_2|$ ile temsil edilmektedir. Bu uzaklık metriği, hücrelerin tüm özelliklerini düğümler ve özelliklerin çarpımını ilişki değerleriyle ağırlıklandırılmış kenarlar olarak modelleyen iki ağırlıklı tam çizge arasındaki izomorfizm skoru olarak da düşünülebilir.

Proteomik veri için uzaklık matrisi hesaplanırken ilişki değeri olarak protein-protein etkileşim değerleri kullanılmıştır. Ağırlıklar için, STRING [100] veritabanından (v12.0) önerilen ilişki değerleri kullanılmıştır. Bu değerler 0 ile 1 arasında olup deneylerde elde edildikleri hali ile kullanılmıştır. Ek olarak, özelliklerin kendisiyle olan çiftleri için ilişki değeri 1 olarak ayarlanmıştır. Bu ilişki değerleri hücre zarında bulunan proteinlerin etkileşim ağlarında iki hücrenin ne kadar uzak olduğunu yansıttığı için aynı hücre tipi arasında uzaklıkların daha düşük olmasını sağlamaktadır [72]. Hücreler

arasında ilgili modalite ölçümlerinin vektörel benzerliğinin ötesinde biyolojik rehberlik kullanılması, farklı tipteki hücreler arasındaki heterojenliği artırmaktadır. Kullanılan ağırlıklar sayesinde hücre tipi farklılaştırması için daha az kritik olan proteinlerin etkisi azaltılmaktadır. Ayrıca, aktif proteinlerin vurgulanması, çeşitli hücre tipleri arasında kümeleme doğruluğunu artırmaktadır (Çizelge 3.1).

Transkriptomik ile hücre uzaklıklarını hesaplamak için özgün uzaklık metriği yerine kosinüs uzaklığı üç nedenden dolayı seçilmiştir. İlk olarak, ham veri yerine PCA gömmeleri kullandığımız için bileşenler arasında herhangi bir biyolojik ilişkili bulunmaması özgün metriği kullanmayı engellemektedir. Ancak, algoritmayı diğer omik türlerinde uygulanabilirliğini görmek için ham ölçümlerle test edilmiştir. İlişki değerini hesaplamak için biyolojik yolaklar kullanılmıştır. Yolaklar, yolak hiyerarşisinin tüm seviyelerini içerecek şekilde Reactome [101] veritabanından alınmıştır. İki gen arasındaki ilişkili değeri, genlerin aynı yolda birlikte bulunma sayısı olarak hesaplanmıştır. İlişki değerleri min-max yöntemi ile normalize edilmiştir ve uzaklık hesaplamasında ağırlık olarak kullanılmıştır.

Sonuçlar, ilişkili değerlerinin, daha az üyeye sahip azınlık hücre türlerinden monositler gibi çoğunluk hücre türlerine olan yanlış eşleşmeleri artırdığını göstermiştir. Daha uzun yolakların, yaygın olarak ifade edilen genler arasında daha yüksek ilişkili değerlerine neden olduğu gözlemlenmiştir. Sonuç olarak, nadir hücre kümelerinin farklılaşmasına neden olan ayırt edici öznelilikler, düşük ilişki değerleri sebebiyle baskılanmıştır. Bunun bir sonucu olarak, genler arasındaki biyolojik ilişkili değerleri için yolakların kullanılmasından vazgeçilmiştir. Özgün uzaklık metriği kullanılarak transkriptomik ile hücre uzaklıklarını hesaplamak için daha ileri çalışmalar yapılmamıştır.

Son olarak, uzaklık metriğinin paralelleştirilebilir algoritmasına rağmen karmaşıklığı dikkate alınması gereken bir unsurdur. Hadamard çarpımları tamamen paralel

Çizelge 3.1: Biyolojik bilginin hücre eşleştirme başarımına etki sonuçları.

Veriseti	PPI ile	PPI'sız Tümü	PPI'sız Diyagonal	Kosinüs Uzaklığı
HIV	92.88	89.58	89.80	90.92
MNC	78.11	77.73	77.54	77.21

olarak gerçekleştirilmiştir çünkü her hücre için ayrı vektörel çarpımlar sonucu hesaplanmaktadır. Ayrıca, çıkarma işlemleri paralel olarak yapılmış ve bellek boyutuna göre partiler halinde hesaplanmıştır. Ancak, algoritmanın karmaşıklığı $O(n * n * m * m)$ olup, burada n hücre sayısını ve m öznitelik sayısını ifade eder. Algoritmanın karmaşıklığı, hem hücre sayısı hem de öznitelik sayısındaki artışla birlikte üstel şekilde artış gösterdiğinden yüksek hesaplama gereksinimlerine sahiptir. Bu durum uzaklık metriğini, 10 binin üzerinde özniteliğe sahip transkriptomik veri için daha az tercih edilir hale getirmektedir.

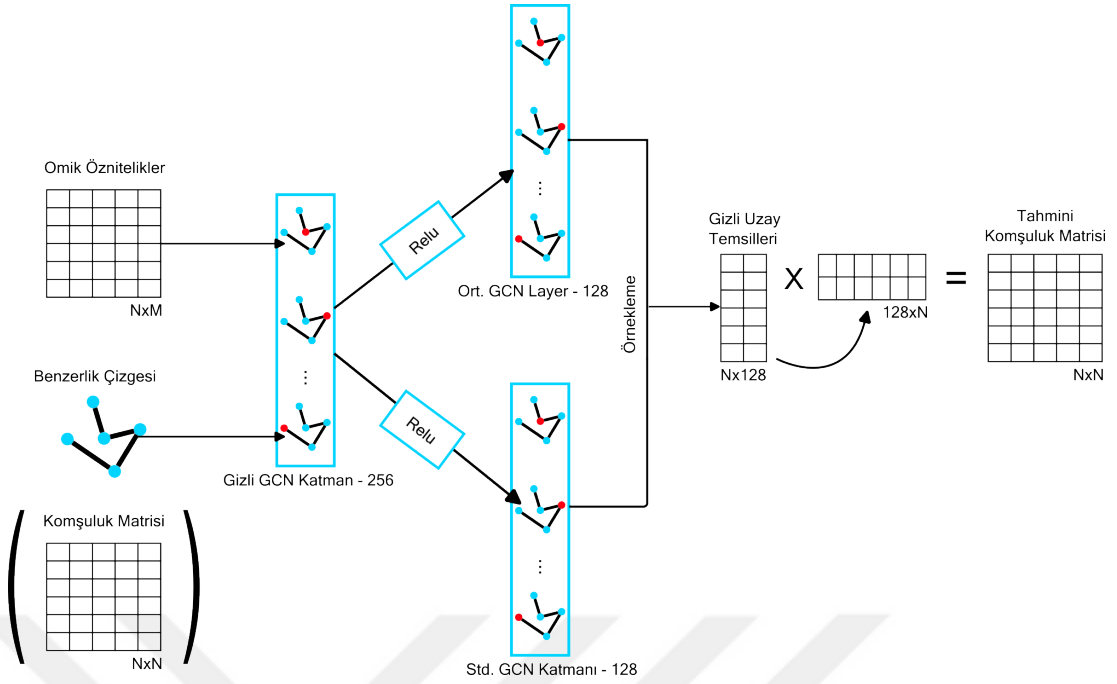
Uzaklık matrisleri normalize edilerek modaliteler arasında ölçek dengesizliğinden kaynaklanan baskınlık sorunu önlenmiştir. Her modalite için her hücre, en yakın 100 komşusunun ortalama uzaklığına bölünerek bireysel olarak normalize edilmiştir. Böylece, hücrelerin en yakın 100 komşusu için modaliteye özgü uzaklık değerleri yerine uzaklık oranları elde edilmiştir.

Normalize edilen komşu sayısı N , hücreler için dikkate alınan küme içi komşuluk boyutunu belirlemektedir. Komşuluk boyutunun artırılması, ölçek farklılıklarının neden olduğu sorunu tekrar artırabilmektedir. Ancak, dar komşuluklar da aynı soruna yol açabilir çünkü normalizasyon yalnızca birkaç küme içi uzaklığı etkileyerek geri kalan uzaklıkları normalize etmeden bırakabilmektedir.

Bu normalize edilmiş uzaklık matrislerine Hadamard çarpımı uygulanarak bütünleştirilmiş uzaklık matrisi D^I elde edilmiştir. Bütünleştirilmiş uzaklıklar daha sonra uçtan uca çoklu-omik veri bütünleştirme modelinin eğitiminde kullanılmıştır.

3.2.1.2. Modalitelerin Gömülmesi

Transkriptomik ve proteomik, uçtan uca çoklu modalite bütünleştirme modelinin eğitimi sırasında daha iyi yakınsama sağlamak için bağımsız VGAE modelleri kullanılarak ayrı ayrı gömülmüştür. Geleneksel oto-kodlayıcılar yerine varyasyonel model kullanılması, benzer dağılımların bütünleştirilmesini kolaylaştırdığı için bütünleştirme tutarlılığını artırmaktadır [62]. Gömme modellerinin ön eğitimi, her modalite içindeki hücre türleri arası heterojenliği artırmıştır. Her model, kendi modalitesine ait uzaklık matrisleri üzerinde eğitilmiştir. Bu önceden eğitilmiş modeller, modaliteye özgü hücre komşuluklarını bütünleştirme modeline aktarmaktadır. Bu yaklaşım, uçtan uca



Şekil 3.2: Modaliteleri gizli uzaya gömmek için kullanılan varyasyonel çizge oto-kodlayıcı model mimarisi.

eğitim sırasında yakınsamayı iyileştirerek modelin bireysel modaliteler içindeki hücre ilişkilerini öğrenmek yerine, öncelikli olarak bütünleştirmeye odaklanmasını sağlamıştır. Her bir VGAE modeli, ilgili modalite için hesaplanan benzerlik çizgesini ve modalitenin ham ölçümlerini girdi olarak hücre gömmesini temsil eden bir vektör çıktısı üretir (Şekil 3.2). Modaliteye özgü benzerlik çizgesi, modalitenin uzaklık matrisinin filtrelenmesiyle elde edilmektedir. Benzerlik çizgesinin komşuluk matrisini üretmek için aşağıdaki filtreleme fonksiyonu kullanılmıştır:

$$A_{i,j} = \begin{cases} 1, & \text{eğer } D_{i,j} < 1 \\ 0, & \text{aksi durumda} \end{cases} \quad (3.2)$$

Formül 3.2’de, A modalitenin hücre uzaklık matrisi D ’nin filtrelenmesi ile elde edilen komşuluk matrisini temsil etmektedir. $A_{i,j}$ ve $D_{i,j}$, sırasıyla hücre i ve j arasındaki komşuluk ve uzaklık ilişkilerini ifade etmektedir. Bu formül, her hücrenin k en yakın komşusu ile (k , 0 ile N arasında değişmektedir) benzerlik çizgesinde arasında ayrıt olmasını sağlamaktadır. Uzaklık matrisleri, her hücrenin N en yakın komşusunun ortalama uzaklığına göre normalize edildiğinden, filtreleme için sabit bir eşik değeri olarak 1 kullanılmıştır. Her hücre için, k komşunun uzaklıkları bu ortalamanın altında

iken, diğer tüm uzaklıklar 1'i aşmaktadır. Veri setleri arasında bu tutarlı eşik değeri, her hücre için uyarlanabilir bir komşuluk boyutu sunar ve hiperparametre optimizasyonuna olan ihtiyacı ortadan kaldırır. Uyarlanabilir komşuluklar, sabit k-ın yaklaşımlarında sıklıkla görülen yanlış eşleşmiş komşuları azaltmaya yardımcı olur ve merkezi hücre kümelerindeki hücreler için komşulukları aşırı derecede sınırlamadan, dışbükey olmayan bölgelerdeki hücreleri hassas bir şekilde ele almayı sağlar.

Varyasyonel çizge oto-kodlayıcı model yalnızca kodlayıcıdan oluşmaktadır. Model, girdileri yeniden oluşturmak yerine komşuluk yapısını öğrenmek için eğitilmektedir. Bu sebeple model çıktı vektörleri kullanılarak tahmini bir komşuluk üretilmekte ve modelin tahmini komşulukları girdi benzerlik çizgesine benzetmesi hedeflenmektedir. Aşağıdaki kayıp fonksiyonu, pozitif ve negatif ilişkileri ayrı ayrı ele alarak, modaliteye ait komşuluk matrisi ile tahmin edilen komşuluk matrisi arasındaki hücre bazlı ortalama Manhattan uzaklığını hesaplamaktadır:

$$\begin{aligned}
 L_{pos} &= \frac{\sum_{i=1}^n \sum_{j=1}^n |A_{i,j} - \tilde{A}_{i,j}| \cdot A_{i,j}}{n^2} \\
 L_{neg} &= \frac{\sum_{i=1}^n \sum_{j=1}^n |A_{i,j} - \tilde{A}_{i,j}| \cdot (1 - A_{i,j})}{n^2} \\
 L &= \frac{L_{pos} + L_{neg}}{2} + KL[N(\mu_z, \sigma_z), N(0, 1)]
 \end{aligned} \tag{3.3}$$

Formül 3.3'de, L kayıp fonksiyonunu, L_{pos} aralarında ayrıt olan hücre çiftleri arasındaki kayıp değerini, L_{neg} aralarında ayrıt olmayan hücre çiftleri arasındaki kayıp değerini belirtmektedir. A belirli bir modalite için komşuluk matrisini ve $\tilde{A}$ model tarafından tahmin edilen komşuluk matrisini temsil etmektedir. KL , örnekleme fonksiyonunun gizli uzaydaki dağılım ile standart Normal dağılım arasındaki tutarsızlığı ölçen Kullback-Leibler (KL) uzaklığını ifade etmektedir [75]. KL uzaklığı, gizli uzayın normal dağılım göstermesini teşvik etmektedir. $\tilde{A}$ aşağıdaki şekilde hesaplanır:

$$\tilde{A} = \frac{Z \times Z^T}{\|Z\|^2} \tag{3.4}$$

Formül 3.4'de, Z , hücrelerin gizli uzay gömmelerini temsil eden matristir. $\tilde{A}$ içindeki değerler 0 ile 1 arasında değişmektedir ve daha yüksek değerler, daha benzer

gömmelere sahip hücre çiftlerini yansıtmaktadır. Böylece, aralarında ayrıt olmayan hücreler arasındaki benzerlikler azaltılırken komşu hücreler arasındaki benzerlikler artırılarak VGAE modellerinin eğitilmesi sağlanır. Pozitif ve negatif örnekleri kayıp hesaplamasında ayrı ayrı ele alan kontrastlı öğrenme yaklaşımı, genellikle benzerlik çizgilerinin seyrekliği nedeniyle negatif örneklerin baskınlığını önlemektedir.

Karşılaştırmalarda kullanılan doğrusal katmanlı model, VGAE modeli ile aynı mimariye sahiptir ancak çizge evrişim katmanları yerine doğrusal katmanlar kullanılarak tasarlanmıştır. Böylece, hücre gömme modeli, standart bir VAE modelinin kodlayıcısına dönüşmüştür. Bu modelin kayıp değeri 3.3'deki kayıp fonksiyonu, Kullback-Leibler uzaklık terimi çıkarılarak hesaplanmıştır.

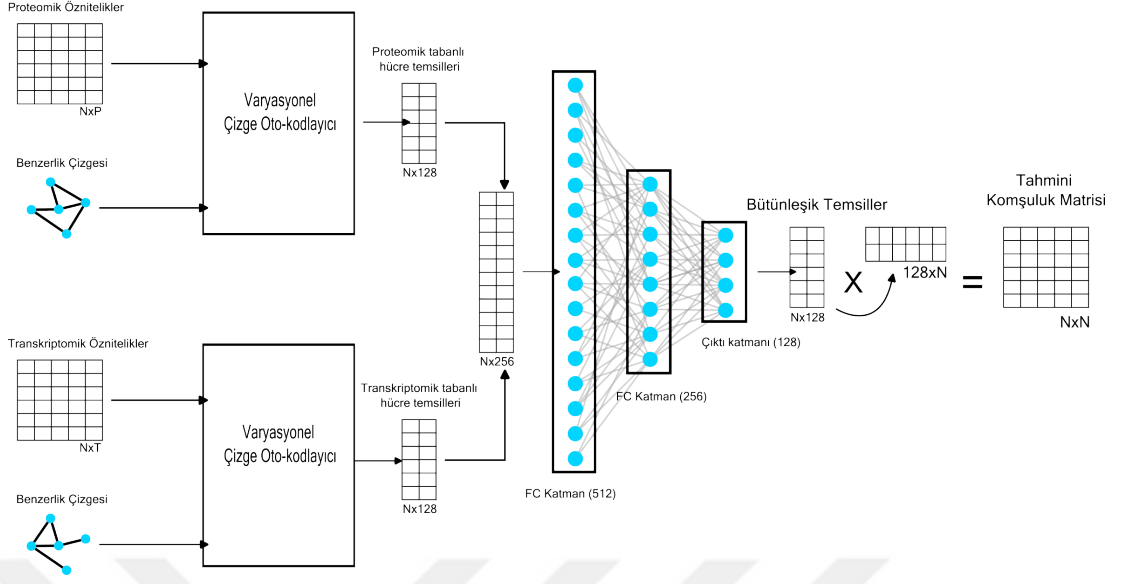
Bütünleştirilmiş uzaklık matrislerinin UMAP gömmeleri karşılaştırma için hesaplanmıştır. UMAP gömmeleri, Scanpy paketindeki *tl.umap* yöntemi ile üretilmiştir [102]. Bütünleştirilmiş uzaklık matrisi, *connectivities* parametresi olarak ayarlanmıştır. *distances* parametresi, her hücrenin yalnızca en yakın 30 komşusunun tutulduğu ve diğer tüm uzaklıkların 0 olarak ayarlandığı filtrelenmiş bütünleştirilmiş uzaklık matrisi olarak kullanılmıştır. Dikkate alınan en yakın komşu sayısı, Seurat paketindeki *RunUMAP* fonksiyonunda önerildiği gibi 30 olarak ayarlanmıştır. Diğer tüm parametreler varsayılan değerlerinde bırakılmıştır.

3.2.1.3. Çoklu-omik Veri Bütünleştirme

Çoklu-omik veri bütünleştirme için geliştirilen model bir çoklu-görünüm varyasyonel çizge oto-kodlayıcı (VGAE) ve VGAE çıktılarını bütünleştiren 2 katmanlı bir kodlayıcıdan oluşmaktadır (Şekil 3.3). Modaliteler, modaliteye özgü önceden eğitilmiş VGAE modelleri kullanılarak bağımsız olarak gömülmüştür. Her VGAE modeli, ham modalite ölçümleri ile modaliteye özgü benzerlik çizgesini alır ve her hücre için gizli uzay gömmelerini üretir. Hücre gömmeleri birleştirilerek kodlayıcıya girdi olarak verilir ve model çıktısı olarak modaliteler bir gizli uzayda bütünleştirilir.

Uçtan uca model, VGAE modellerini eğitmek için kullanılan kayıp fonksiyonuna benzer bir kayıp fonksiyonu ile eğitilmiştir (Formül 3.3). Ancak, her iki modalite dağılımı için de KL uzaklığı dahil edilmiştir; böylece tek bir modaliteye aşırı vurgu yapılmasının neden olduğu gömme bozulmaları önlenmiştir. Bütünleşik uzaklık matrisi D^I , aşağıdaki

A



Şekil 3.3: Modaliteleri bütünleştirmede kullanılan çoklu-görünüm varyasyonel çizge oto-kodlayıcı model mimarisini.

formül ile filtrelenerek gerçek komşuluk matrisi elde edilmiştir:

$$A_{i,j} = \begin{cases} 1 - D_{i,j}^I, & \text{eğer } D_{i,j}^I > th_H \\ 0, & \text{aksi durumda} \end{cases} \quad (3.5)$$

Formül 3.5'te, $A_{i,j}$ ve $D_{i,j}^I$, sırasıyla hücre i ve j arasındaki komşuluk ve uzaklık ilişkisini temsil eder. Eşik th_H , dikkate alınan komşulukta yer almayan hücreler arasındaki ilişkileri filtrelemektedir. Bu eşik, her hücrenin en yakın 100 komşusuna olan ortalama uzaklık olarak hesaplanır. Belirlenen komşuluk içindeki hücre sayısı, uzaklık matrisi hesaplama bölümünde belirtildiği üzere 100 olarak ayarlanmıştır.

Filtrelenen uzaklıklar, Formül 3.5'te olduğu gibi 1'den çıkarılarak ağırlıklı bir komşuluk matrisine dönüştürülür. Tahmin edilen komşuluk matrisi $\tilde{A}$, bütünleştirilmiş gömmeler üzerinde Formül 3.4 uygulanarak hesaplanmaktadır. Sonuç olarak, yalnızca hücreler arasındaki 0-1 ilişkileri değil, aynı zamanda ilişkilerin dereceleri de kayıp fonksiyonu ile optimize edilir. Bu nedenle, bütünleştirme için önemsiz olan ilişkilerin yaklaşık %99'unun budanması, uçtan uca modelin yakınsamasını kolaylaştırmaktadır. Ayrıca, filtreleme ile azalan sıfırdan farklı değer sayısı optimizasyon algoritmasının daha hızlı yakınsamasını sağlamaktadır.

Uçtan uca modelin kayıp fonksiyonunda da kontrastlı öğrenme yaklaşımı kullanılmıştır. Pozitif ve negatif örnekler ayrı ayrı ele alınmış ve Formül 3.3'e benzer şekilde eleman bazında Manhattan uzaklıkları ayrı ayrı hesaplanmıştır. Ancak, maskeleme için ağırlıksız komşuluk matrisi A^U da hesaplanmıştır, çünkü A kenar ağırlıkları yerine 0-1 ilişkileri içerir. A^U , uzaklık matrisi D^I üzerinde Formül 3.2 uygulanarak oluşturulur. Böylece, hesaplanan uzaklıklar A^U ile çarpılarak yalnızca pozitif örnekler arasındaki eleman bazında Manhattan uzaklıkları filtrelenabilmektedir. Negatif kayıp, Formül 3.3'te olduğu gibi uzaklıkların $1 - A^U$ ile çarpılmasıyla hesaplanır.

3.2.1.4. Ortam ve Model Kurulumu

Gömüleme ve bütünleştirme yöntemleri, PyTorch ekosistemi (sürüm 2.4.0) kullanılarak gerçekleştirilmiştir. VGAE modelinde, çizge evrişim katmanları olarak GraphSAGE [103] katmanları kullanılmıştır. İlk evrişim işlemi için komşu sayısı 20, ikinci evrişim işlemi için ise 10 olarak ayarlanmıştır ve bu ayarların orijinal çalışmayla tutarlı olması sağlanmıştır. İlk evrişim katmanının boyutu 256 olup, sonraki μ ve σ evrişim katmanları, 128 boyutunda gizli uzay gömme boyutlarıyla aynı boyuttadır. Gizli uzayın boyutu, her modalitenin giriş boyutları dikkate alınarak ayarlanmıştır. Kodlayıcı, 512 boyutunda bir katmandan ve bunu takip eden yarı boyutta başka bir katmandan oluşmaktadır. Bütünleştirilmiş gömmeler de 128 boyutundadır.

Tüm modellerin optimizasyonunda Adam optimizasyon algoritması ($lr = 0.001$ varsayılan) kullanılmıştır. Modeller, deneylerde varsayılan olarak 500 iterasyon boyunca eğitilmiştir. VGAE modellerinin genellikle 200 iterasyondan sonra yakınsadığını, ancak kayıp değerinin 1000 iterasyon boyunca sürekli azalmaya devam ettiği gözlemlenmiştir. Ayrıca, uçtan uca modelin yalnızca 50 iterasyon bile hızlı bir şekilde yakınsadığını, ancak iyileşmenin 2000 iterasyon kadar devam ettiği gözlemlenmiştir. Deneylerimizde herhangi bir ek hiperparametre optimizasyonu gerçekleştirilmemiştir.

3.2.2. Karşılaştırmada Kullanılan Algoritmalar

3.2.2.1. Seurat v4

Karşılaştırmalarımızda Seurat v4 [72] kullanılmıştır. Testlerde Seurat R paketi kullanılmıştır. k -NN hücre eşleşme skorlarını hesaplarken ağırlıklı en yakın komşu uzaklıkları kullanılmıştır. Tüm diğer metrikler, varsayılan parametrelerle *RunUMAP* yöntemi kullanılarak elde edilen UMAP gömmeleri üzerinden hesaplanmıştır.

3.2.2.2. TotalVI

MNC veri seti üzerinde kıyaslamalarımızda TotalVI [92] derin öğrenme algoritması kullanılmıştır. Ancak, algoritma yalnızca normalize edilmemiş protein sayımları ile çalıştığından ve bu veri HIV veri seti için mevcut olmadığı için HIV veri seti deneylerinde test edilememiştir. Algoritmanın scVI [62] paketi içindeki gerçekleştirilmesi varsayılan parametrelerle kullanılmıştır.

3.2.2.3. Mofa+

Kıyaslamalarımızda, Mofa+ [88] algoritması kullanılmıştır çünkü bu algoritma çoklu-omik veri bütünleştirme için en yaygın karşılaştırma algoritmalarından biridir [104]. Algoritma, mofapy2 paketinin bir sarmalayıcısı olan muon paketi içindeki gerçekleştirilmesi kullanılarak test edilmiştir. Muon uygulamasında tüm parametreler varsayılan olarak kullanılmıştır.

3.2.2.4. Mowgli

Sonuçlarımızı, optimal taşıma ile optimize edilmiş başka bir matris faktörizasyon algoritması olan Mowgli [90] algoritmasının sonuçları ile de karşılaştırdık. Bu algoritma, bütünleştirme için yüksek değişken öz niteliklere (HVG) ihtiyaç duymaktadır. Yüksek değişken genleri, Seurat v3 [35] paketi kullanılarak hesaplanmıştır. HVG sayısı, Seurat v3 makalesinde olduğu gibi 2000 olarak belirlenmiştir. Ayrıca, Mowgli yazarları, yaklaşık 35 bin gen için 2500 HVG ve yaklaşık 8 bin gen için 1500 HVG kullanmıştır. Bu nedenle, filtrelenmiş kıyaslama veri setimizdeki yaklaşık 12 bin gen için 2000 HVG

uygun bir deęerdir. Tm proteinler, orijinal alıřmada olduęu gibi yksek deęiřken olarak belirlenmiřtir.

3.2.3. Metrikler

Ortalama siluet geniřlięi (ASW) ve ayarlanmıř rand indeksi (ARI) Scib paketi [44] kullanılarak varsayılan parametrelerle yatay btnleřtirme blmnde aıklandığı hali ile hesaplanmıřtır. K-en yakın komřu hcre eřleřme skoru ve F1 skoru bu alıřma kapsamında gereklenmiřtir. Normalleřtirilmiř karřılıklı bilgi (NMI) ve izge baęlantısı metrikleri Scib paketinde sunulan fonksiyonlar ile hesaplanmıřtır.

3.2.3.1. Hcre Eřleřme Skoru

Hcre eřleřme skoru veya k-nn eřleřme skoru, hcrelerin yerel komřuluklarının saflık deęerini gerek hcre tiplerini kullanarak lmektedir. Her hcrenin doęru eřleřen komřularının sayısı ařağıdaki forml ile hesaplanır ve ortalaması alınır:

$$Hcre\ eřleřme\ skoru = \frac{1}{n \cdot k} \sum_{i=1}^n \sum_{j=1}^k \begin{cases} 1, & \text{eęer } H_{tip}(C_i) = H_{tip}(C_{i,j}) \\ 0, & \text{aksi durumda} \end{cases} \quad (3.6)$$

Forml 3.6'da, C_i i. hcreyi, $C_{i,j}$ i. hcrenin j. en yakın komřusunu ve $H_{tip}()$ ise gerek etiketleme fonksiyonunu temsil eder. Bylece, tm hcreler k-en yakın komřularıyla doęru eřleřtięinde metrik 1 olur; eęer hibir hcre iin aynı trde komřu bulunmazsa metrik 0 deęerini alır.

3.2.3.2. Normalleřtirilmiř Karřılıklı Bilgi (NMI)

Normalize edilmiř karřılıklı bilgi, bir gerek kmeleme řeması ile verinin hesaplanmıř kmeleme řeması arasındaki hizalanmayı lmektedir. NMI metrik deęerleri 0 ile 1 arasında deęiřir ve 1, gerek hcre kmeleri ile btnleřtirilmiř verideki kmelenme arasında mkemmел bir eřleřmeyi ifade eder. Alakasız kmelenmeler dřk NMI skorları ile sonulanır. Bu nedenle, yksek NMI skorları, btnleřtirilmiř verideki aynı tr hcrelerin bir araya toplandıęını gsterir.

Deneylerde, Scib paketinin $nmi()$ fonksiyonu varsayılan parametrelerle kullanılmıřtır.

Bu yöntem, gerçek kümeleme etiketlerini alır ve Leiden algoritmasını [105] kullanarak bir kümeleme oluşturur. Gerçek hücre türleri küme etiketleri olarak kullanılarak kümeleme hesaplanmıştır. Scib, NMI skorunu hesaplamak için Scikit-learn paketinin *normalized_mutual_info_score()* fonksiyonunu aritmetik ortalama yöntemi parametresi kullanarak hesaplamaktadır.

3.2.3.3. Çizge Bağlantısı

Çizge bağlantısı skoru, aynı tür düğümlerin ne kadarının aynı bağlı bileşende olduğunu ölçmektedir. Scib çizge bağlantısı metriği aşağıdaki formül ile hesaplanmaktadır:

$$\text{Çizge bağlantısı} = \frac{1}{|C|} \sum_{c \in C} \frac{LCC(G(V_c; E_c))}{|V_c|} \quad (3.7)$$

Formül 3.7’de, C benzersiz hücre tiplerinin kümesini, c her bir benzersiz hücre tipini ve V_c c hücre tipindeki düğüm kümesini ifade eder. LCC , bir çizgeyi girdi olarak alıp, çizgedeki en büyük bağlı bileşenin düğüm sayısını döndüren fonksiyondur. $G(V_c; E_c)$, V_c ’nin c tipindeki hücreleri barındıran düğüm kümesi ve E_c ’nin bu düğümler arasındaki kenar kümesi olduğu bir çizgedir. Dolayısıyla, $LCC(G(V_c; E_c))$, c hücre türündeki hücrelerin en büyük bağlı bileşenindeki düğüm sayısını ifade eder. Bu formül, bağlı bileşenlerdeki hücrelerin oranını her hücre türü için hesaplar ve her hücre türü için elde edilen maksimum oranların ortalamasını alır.

Çizge bağlantısı metriği, aynı tipteki hücrelerin bütünleştirilmiş veride ne kadar bir arada olduğunu değerlendirmek için anlamlıdır. Ayrıca, bu metrik, daha düşük bağlantı skorlarına yol açan hücre tipi içi parçalanmalara duyarlıdır. Öte yandan, metrik, her bir hücre tipi için bireysel hesaplamalar yapar. Bu nedenle, herhangi bir hücreler arası heterojenlik olmaksızın güçlü hücre türü içi benzerlik ile yapılan bir bütünleştirme, yüksek performans sonuçlarına yol açabilir. Dolayısıyla, genel bütünleştirme performansı diğer metrikler sonuçları ile birlikte değerlendirilmelidir.

3.2.3.4. F1 Skoru

F1 skoru, deneylerimizde bir sorgu kümesinin hesaplanan kümelenme ile hizalanmasını hesaplamak için kullanılmıştır. F1 skoru aşağıdaki formül ile hesaplanır:

$$\begin{aligned}
Kesinlik &= \frac{|C_Q \cap C_i|}{|C_i|} \\
Duyarlilik &= \frac{|C_Q \cap C_i|}{|C_Q|} \\
F1(C_Q, C_i) &= \frac{2 * Kesinlik * Duyarlilik}{Kesinlik + Duyarlilik} \quad (3.8)
\end{aligned}$$

Formül 3.8’de, C_Q , sorgu kümesindeki hücrelerin kümesini ve C_i , bir kümelenmedeki i . kümede bulunan hücrelerin kümesini ifade eder. Bir kümelenmenin sıralama değeri $MAX(\sum_{C_i \in C} F1(C_Q, C_i))$ olarak hesaplanmıştır. Formül, sorgu hücre kümesi ile kümelenmedeki hücre kümeleri arasındaki maksimum F1 skoruna göre kümelenmeleri sıralar. Böylece, analiz edilen alt kümeleri ortaya çıkarma potansiyeline sahip algoritmaların en iyi kümelenmeleri belirlenmiştir.

3.2.4. Veri Analizi

3.2.4.1. Ayırt Edici Öznitelik Analizi

Ayırt edici proteinler/genler, Scikit-learn paketinin *f_classif* yöntemi kullanılarak belirlenmiştir. Bu yöntem, öznitelikleri ve hedef etiketleri giriş olarak alır ve her öznitelige F-testi uygulayarak etiketlerin sınıflandırılması için öznitelik skorlarını hesaplar. Bir hücre türünün ayırt edici öznitelikleri hesaplanırken, aynı türdeki hücrelerin etiketleri 1 ve diğer hücrelerin etiketleri 0 olarak ayarlanmıştır. Öte yandan, özgün hücre alt kümelerinin ayırt edici öznitelikleri, yalnızca alt küme ile aynı türdeki hücreler kullanılarak hesaplanmıştır. Bu durumda alt kümedeki hücrelerin etiketleri 1 ve diğer hücrelerin etiketleri 0 olarak ayarlanmıştır. Ayırt edici öznitelikler, F-testi skorlarına göre sıralanmış ve en yüksek k öznitelik ayırt edici olarak belirlenmiştir.

3.2.4.2. Hücre Alt Kümesi Tanımlama

Hücre alt kümesi analizinde, tek-hücre veri setlerinin genellikle dışbükey olmayan hücre kümeleri barındırması nedeniyle, yoğunluk tabanlı DBSCAN kümeleme algoritması [73] kullanılmıştır. DBSCAN algoritmasının Scikit-learn gerçekleştirilmesi kullanılmıştır. Minimum örnek sayısı, her hücre türündeki minimum hücre sayısı ile uyumlu olan 5 olarak varsayılan şekilde ayarlanmıştır. MNC veri seti üzerindeki testlerde, her

algoritma için en iyi eps değerini seçmek üzere sistematik bir arama yapılmıştır. İlk olarak, hücrelerin en yakın 5. komşularına olan uzaklıkları sıralanmıştır, çünkü dikkate alınan minimum komşu sayısı 5 olarak belirlenmiştir. Minimum ve maksimum uzaklıklar tanımlanmış ve bunlar arasında 20 eşit aralıklı uzaklık belirlenmiştir. Her bir uzaklık bireysel bir kümeleme için eps değeri olarak kullanılmıştır. Eps değerlerinin sayısı, %5 değişim oranı çözünürlüğü sağlamak için 20 olarak ayarlanmıştır. Daha karmaşık veri setleri için daha fazla bölüme ayırarak eps değerlerinin sayısını arttırmak mümkündür. Sistematik kümeleme, özgün hücre alt kümelerini belirlemek için öncelikle SCPRO-VI algoritması üzerinde gerçekleştirilmiştir. 50'den fazla hücre içeren ve tamamı aynı türden hücrelerden oluşan hücre kümeleri filtrelenmiştir. CLP ve granülosit hücre alt kümeleri, $eps = 0.005$ değerine sahip aynı kümeleme içinde tanımlanmıştır. Daha sonra bu işlem her algoritma için uygulanmış ve her biri için 20 farklı kümeleme elde edilmiştir. CLP ve granülosit hücre alt kümeleri için kümelemeler bireysel F1 skorlarıyla sıralanmıştır. Her alt küme için en yüksek sırayı alan kümeleme veya kümeleme çifti, her algoritma için optimal kümeleme olarak belirlenmiştir.

HIV veri setinde hasta 1'in 7. gün örneklerinde ortaya çıkan hedeflenen alt kümedeki monosit hücrelerinin tanımlanmasında benzer bir yöntem izlenmiştir. Minimum örnek sayısı 5 olarak ayarlanmış ve 20 farklı eps değeri ile sistematik bir arama yapılmıştır. Ortaya çıkan kümelemeler, monosit kümesinin hedeflenen bölgesindeki ortaya çıkarılan monosit alt kümelerinin içerdiği hücre sayısı sayılarak değerlendirilmiştir. Bütünleşik gömmelerimizdeki en büyük monosit alt kümesi, $eps = 0.03$ değeri ile tanımlanmıştır.

3.2.4.3. Gen Ontolojisi Analizi

Gen ontolojisi (GO) analizi, her veri setindeki en önemli ve en önemsiz proteinler için ayrı ayrı yapılmıştır. Bir veri setindeki en önemli ve en önemsiz proteinler, protein-protein etkileşim (PPI) skorlarına göre belirlenmiştir. PPI çiftleri önce veri setindeki proteinler için filtrelenmiş, ardından etkileşim skorlarına göre sıralanmıştır. İlk 50 etkileşimde yer alan proteinler en önemli proteinler, son 50 etkileşimde yer alan proteinler ise en önemsiz proteinler olarak tanımlanmıştır.

En önemli ve en önemsiz proteinlerin analizi için g:Profiler [106] kullanılmıştır. Analiz iki ayrı turda ve varsayılan parametreler kullanılarak gerçekleştirilmiştir. Sonuçlar, her

bir gen ontoloji terim tipi (moleküler fonksiyon, hücresel bileşen ve biyolojik süreç) için detaylı olarak listelenmiştir.

3.3. Sonuçlar

3.3.1. Deneylerde Kullanılan Veri Kümeleri

Bu çalışmada, iki tek-hücre çoklu-omik CITE-seq veri seti kullanılmıştır [72, 99]. Her iki veri seti de Human Cell Atlas (HCA) veritabanından [107] annData formatında elde edilmiştir. Veri setleri, çoklu-omik veri setlerini işlemek için geliştirilmiş annData'nın genişletilmiş bir versiyonu olan MuData formatına dönüştürülerek kullanılmıştır [108]. Biyolojik öznitelikler, HCA veritabanındaki meta dosyalar referans alınarak transkriptomik ve proteomik olarak ayrılmıştır çünkü orijinal annData dosyaları öznitelik türü bilgisi içermemektedir.

HIV veri seti [72] (*Gene Expression Omnibus erişim ID: GSE164378*), 8 HIV hastasından alınan ve üç zaman aralığında örneklenen 24 veri setine ayrılmış 161764 hücreden oluşmaktadır: aşılama öncesi, aşılama sonrası 3. gün ve 7. gün. Veri setinde 188 protein ve 20380 gen için ölçümler bulunmaktadır.

Çizelge 3.2: HIV veri kümesinin ve örneklenmiş veri kümesinin tiplerine göre hücre sayısı dağılımı.

Hücre Tipi	Örneklenmiş Veri	Ham Veri
CD4 T	2003	41001
Mono	1729	49010
CD8 T	956	25469
NK	511	18664
B	207	13800
Diğer	155	3442
DC	142	3589
Diğer T	61	6789
Toplam	5764	161764

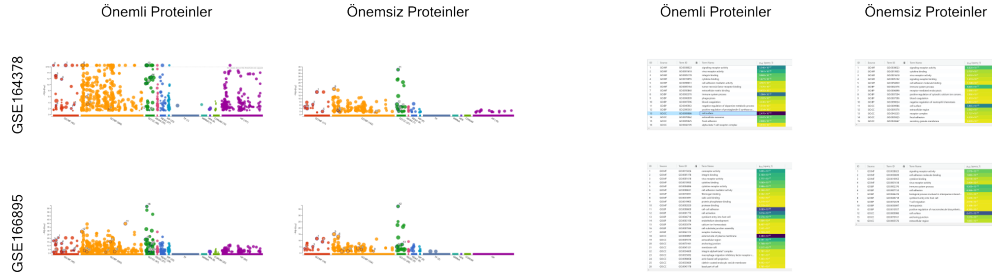
Geliştirilen yöntemin performansı, 24 veri setindeki hücrelerin tamamı arasından rastgele örneklenen bir çoklu-örnek veri setin üzerinde de test edilmiştir. Bireysel veri setlerindeki hücre sayıları ile uyumlu olacak şekilde toplamda 5764 hücre seçilmiştir ve hastalardan alınan hücre sayısı orijinal veri seti ile orantılıdır (Çizelge 3.2).

MNC veri seti, 27998 hücreden elde edilen 197 protein ve 31770 genin ölçümünü içermektedir. TotalVI ve Mowgli, yüksek hesaplama gücü gereksinimleri nedeniyle deneysel kurulumumuzda tam veri seti üzerinde test edilememiştir. Bu nedenle, herhangi bir veri setine has özellik dikkate alınmaksızın rastgele 10 bin hücre veri setinden örneklenmiştir. Çizelge 3.3’de verildiği üzere, örneklenen veri seti, orijinal veri setindeki hücre tipine göre dağılım oranları ile benzer oranda hücre dağılımına sahiptir.

HIV veri setine hali hazırda log-dönüşümü uygulanmış olması ve MNC veri seti tüm özellikler için pozitif ifade değerleriyle normalize edilmiş olması sebebiyle herhangi bir ek normalizasyon yapılmamıştır. Hücreler arası heterojenliği artırırken çelişkiye neden olan gürültülü ölçümleri bastırmak için transkriptomik verilere Temel Bileşenler Analizi

Çizelge 3.3: MNC veri kümesinin ve örneklenmiş veri kümesinin tiplerine göre hücre sayısı dağılımı.

Hücre Tipi	Örneklenmiş Veri	Ham Veri
Common dendritic progenitor	301	796
Common lymphoid progenitor	692	1919
Erythroblast	251	725
Erythroid progenitor	1925	5428
Granulocyte	270	760
Hematopoietic multipotent progenitor	2306	6422
Lymphoid lineage restricted progenitor	968	2778
Megakaryocyte-erythroid progenitor	629	1666
Myeloid lineage restricted progenitor	660	1733
Precursor B	519	1537
Pro-B	922	2607
Proerythroblast	557	1627
Toplam	10000	27998



Şekil 3.4: Önemli ve önemsiz proteinlerin GO analizi özetleri.

(PCA) uygulanmıştır. Bileşen sayısı, proteomik boyutlarla uyumlu biçimde 100 olarak ayarlanmıştır ve her iki veri seti için de verinin %90'ından fazlasını açıklamaktadır.

Transkriptomik verilerdeki genler, ölçüldükleri hücre sayısına göre filtrelendi. Deneylerimizde varsayılan eşik olarak %5 belirlenmiştir. Eşik değeri güvenilirlik ile kapsamlı bir öznitelik kümesi arasında denge sağlayacak şekilde ayarlanabilmektedir. İşlem sonucunda HIV veri setinin alt kümesi için gen sayısı 8417'ye düşmüştür. Aynı filtreleme prosedürü, MNC veri setinde dahil edilen gen sayısını 11361'e indirmiştir. Öte yandan, dizilen protein sayısının sınırlı olması nedeniyle tüm proteinler kullanılmıştır.

Test veri setlerinde bulunan proteinleri analiz etmek için gen ontolojisi (GO) analizi gerçekleştirilmiştir. Her iki veri seti için de en önemli ve en az önemli protein setleri belirlenerek her biri için ayrı ayrı GO analizi uygulanmıştır. Analiz sonuçları, daha yüksek etkileşim skorlarına sahip proteinlerin, zayıf etkileşimlerde görülen proteinlere kıyasla ayırt edici biyolojik süreçlerde daha aktif olduğunu göstermiştir (Şekil 3.4).

HIV veri setindeki [72] önemli proteinlerin ($P = 1.084 \times 10^{-32}$) daha az önemli proteinlere kıyasla ($P = 8.445 \times 10^{-17}$) bağışıklık sisteminde daha aktif bir rol oynadığı gözlemlenmiştir. Ayrıca önemli proteinler, GO:MF virüs reseptörü terimi ile daha yüksek bir skorla ($P = 7.861 \times 10^{-20}$) ilişkilendirilirken daha az önemli proteinler ($P = 8.303 \times 10^{-6}$) daha düşük bir skorla ilişkilendirilmiştir. Bu sonuçlar, hücreler arasındaki uzaklıklar hesaplanırken proteinlerin ağırlıklandırılmasının, aynı biyolojik süreçleri yürüten aynı hücre tiplerini eşleştirme performansını artırabileceğini düşündürmektedir.

3.3.2. SCPRO-VI'nin Dayanıklılığının Doğrulanması

SCPRO-VI'nin performansı, farklı veri setlerinde yeniden kullanılabilirliğini değerlendirmek için 24 bireysel deneyde dizilenmiş hücrelerin birleştirilmesiyle

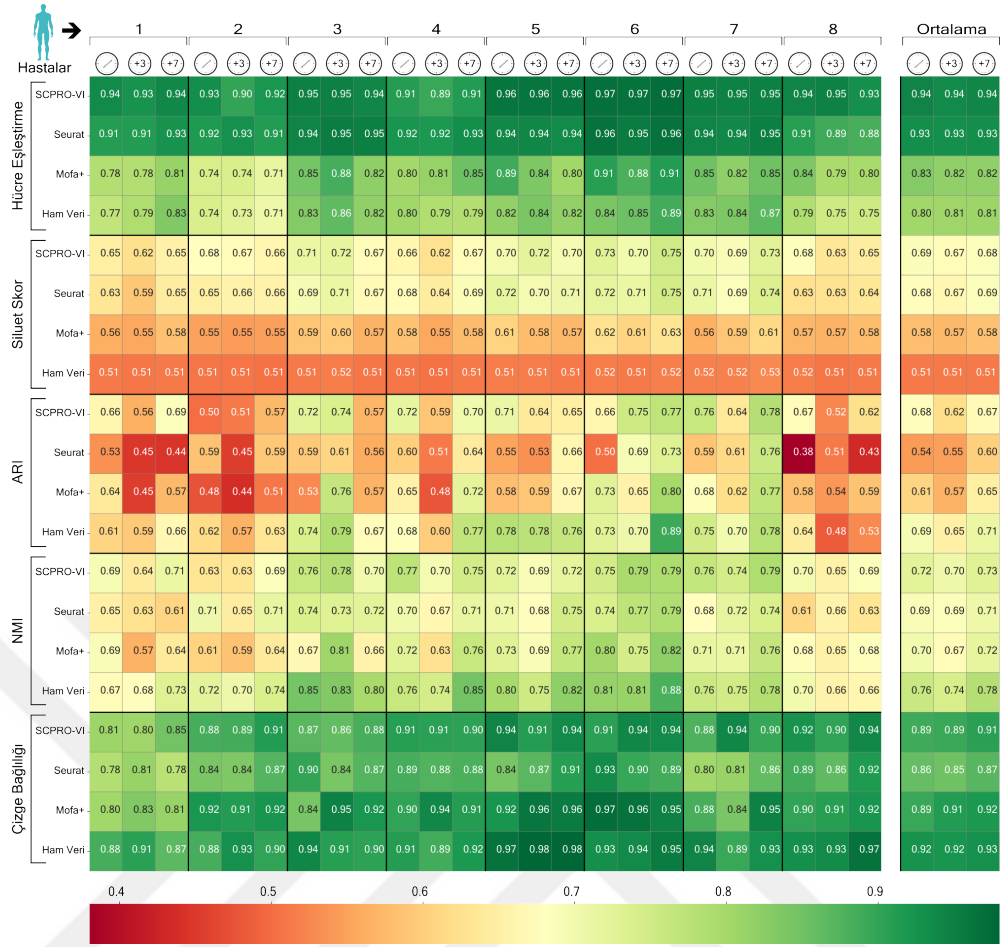
oluşturulan bir veri seti [72] üzerinde test edilmiştir. Veri seti, HIV enfeksiyonu olan 8 hastadan üç farklı zamanda elde edilen hücrelerden oluşmaktadır: aşılama öncesi, aşılama sonrası üçüncü gün ve yedinci gün. Örnekler dondurulmuş ve toplu etkiyi mümkün olduğunca önlemek için aynı anda dizilenmiştir. Dizileme zamanları ve hastalar arasındaki olası farklılıklar toplu etki gürültüsünün etkileri olmadan algoritmanın biyolojik varyasyonlara karşı dayanıklılığını değerlendirmeyi sağlamaktadır.

Bütünleştirilmiş gömmelerin performansı, veri bütünleştirme çalışmalarında yaygın olarak kullanılan beş farklı metrikle ölçülmüştür: 1-NN hücre eşleştirme, ortalama siluet skoru (ASW), düzeltilmiş rand indeksi (ARI), normalize karşılıklı bilgi (NMI) ve çizge bağlantısı. Performans sonuçlarımızı karşılaştırmalı olarak değerlendirmek için Seurat v4 [72] ve Mofa+ [88] algoritmaları deneylere dahil edilmiştir. Performans metrikleri ayrıca ham transkriptomik ve proteomik özniteliklerin uç uca eklenmiş hali üzerinde uygulanarak metrikler için bir temel oluşturulmuştur. TotalVI algoritması, bu veri setinde mevcut olmayan normalize edilmemiş protein sayımlarını gerektirdiği için testlere dahil edilememiştir. Başka bir kıyaslama algoritması olan Mowgli ise uzun çalışma sürelerine rağmen diğer deneylerde umut verici sonuçlar göstermediği için kapsam dışında bırakılmıştır.

Algoritma test sonuçları, zaman noktaları ve örnekler arasında dayanıklılığı takip etmeyi kolaylaştırmak için bir ısı haritası olarak sunulmuştur (Şekil 3.5). Her sütun bir örneğe ait sonuçları içerir. Ardışık üç sütun bir hastaya ait sonuçları temsil eder ve son üç sütun, tüm hastalar arasında ortalama performansları gösterir.

Hastaların örnekleri, aşılama öncesi, aşılama öncesi 3 gün sonra ve aşılama öncesi 7 gün sonra sırasıyla sıralanmıştır. Satırlar algoritmaları temsil eder ve metriklere göre gruplandırılmıştır. SCPRO-VI'a ait sonuçlar çoğunlukla yüksek performans işaret eden yeşil tonlarında gölgelenmiştir. Seurat algoritması da ARI metriği hariç tüm metriklerde yeşil tonlarında sonuçlar elde etmiştir. Ancak, Mofa+ algoritması çizge bağlantısı dışında tüm metriklerde zayıf bir performans sergilemiştir.

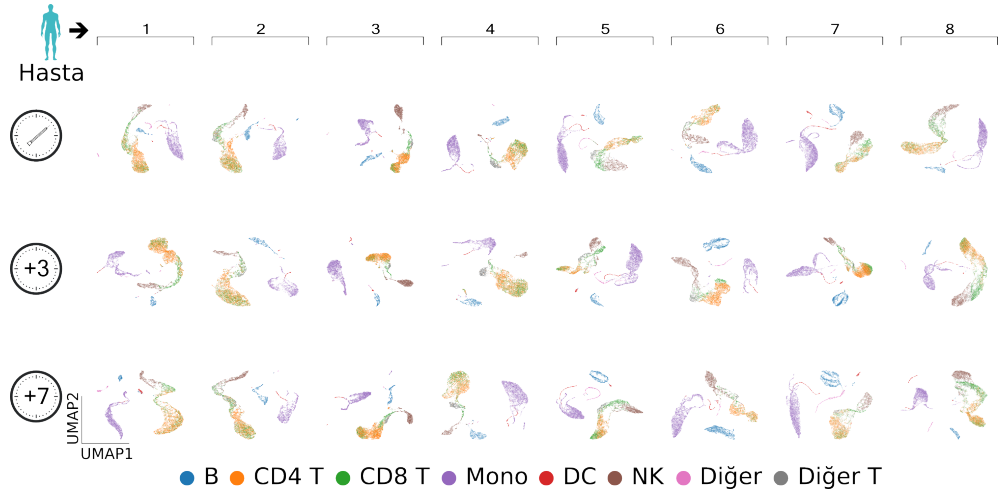
Tüm algoritmalar, 1-NN hücre eşleştirme ve siluet skoru metriklerinde temel performansı iyileştirmiştir. Hücre eşleştirme skoru, önerilen yöntemin aynı tür hücreleri tutarlı bir şekilde gruplama başarısını göstermektedir. Özellikle, eşleştirme performansı, bir hastadan alınan üç örnek arasında en fazla %0.3 varyasyon göstermiştir. SCPRO-VI,



Şekil 3.5: Bütünleştirme algoritmalarının karşılaştırma metrikleri performanslarının ısı haritası olarak gösterimi.

temel hücre eşleştirme performansını önemli ölçüde artırarak %9 ila %21 arasında bir performans artışı sağlamıştır. Seurat da aynı hücreleri bir araya getirme konusunda benzer bir performans sergilemiştir, ancak bizim algoritmamız ortalama performansta Seurat’tan %1 daha iyi sonuçlar elde etmiş ve bu performans artışı hasta 8 için %5’e kadar çıkmıştır. Mofa+, temel performansını %7’ye kadar artırsa da çoğunlukla ham veriyle benzer performans göstermiştir.

Siluet skorları, SCPRO-VI’nin tüm hastalar arasında kümeler arası heterojenliği artırdığını göstermektedir. Temel performans üzerindeki bu önemli iyileştirmeler %23’e kadar çıkmaktadır. Seurat algoritması, küme içi parçalanmalara daha az önem vererek aynı tip hücreleri birden fazla alt kümelere ayırdığı için 1-NN hücre eşleştirme metriklerindeki sonuçlarla benzer bir performans göstermiştir. Mofa+, ham veri siluet skorlarını iyileştirmiş olsa da SCPRO-VI ve Seurat’ın gerisinde kalmıştır. Üç algoritma da tekil hasta örnekleri arasında %5 varyasyon ile aynı dayanıklılığı göstermiştir.



Şekil 3.6: 8 hastadan 3 farklı zaman diliminde elde edilen ölçümlerin bütünleşik sonuçlarının UMAP gösterimi.

ARI ve NMI metrikleri, ham veriyle karşılaştırıldığında hiçbir algoritmanın gerçek hücre kümelemesiyle daha uyumlu bir kümeleme elde edemediğini göstermektedir. Bununla birlikte, SCPRO-VI, benzer performans gösteren Seurat ve Mofa+ ile karşılaştırıldığında %2 daha iyi NMI skorları elde etmiştir. Öte yandan, SCPRO-VI, ARI metriğinde, aynı kümede yer almaması gereken hücreleri ayırmada her iki algoritmadan da daha üstün bir performans sergilemiştir. Algoritmamız, Seurat'tan ortalama %9 ve Mofa+'tan %5 daha iyi ARI skorları elde etmiştir. Ayrıca, SCPRO-VI tüm 24 örnek arasında daha dayanıklı bir ARI performansı sergilemiştir. Algoritmanın ARI sonuçları %50 ile %78 arasında değişirken, Seurat %38 ile %76 arasında ve Mofa+ %44 ile %80 arasında değişim göstermektedir.

Çizge bağlantısı metriği, farklı hücre tiplerinin ortak kümelemelerine duyarlı olmadığı için diğer performans metrikleri dikkate alınarak değerlendirilmelidir. Sonuçlar, ham veride aynı tür hücrelerin birbirine yakın olduğunu, ancak çoğunlukla farklı türdeki hücre kümeleriyle kesiştiklerini göstermektedir. Öte yandan, bütünleştirme algoritmaları, heterojenliği artırmak için farklı türdeki hücre kümelerini ayırmaya çalışmaktadır. Bu nedenle, SCPRO-VI ve Mofa+'ın çizge bağlantısı skorları, ham veriden %2 daha düşük olmuştur. Seurat algoritması ise bir hücre tipi için birden fazla alt küme oluşturma eğiliminde olduğu için %6 daha düşük skorla en kötü sonuçları elde etmiştir.

Algoritmamız, tüm performans metriklerinde dayanıklı ve göreceli olarak daha yüksek performans sonuçları elde etmiştir. Genel olarak, sonuçlar SCPRO-VI'nin küme

içi parçalanmayı azaltmayı, kümeler arası heterojenliği artırma ile dengelediğini göstermektedir. Ayrıca algoritma, tüm metriklerde sonuçlar arasındaki en düşük standart sapmaya sahip olarak dayanıklılığını ortaya koymuştur. Bulgular, SCPRO-VI'nin, farklı biyolojik koşullardan elde edilen çeşitli örneklerden gelen çoklu-omik veri setlerinin dikey bütünleştirilmesi için dayanıklı ve yüksek performanslı olduğunu göstermiştir.

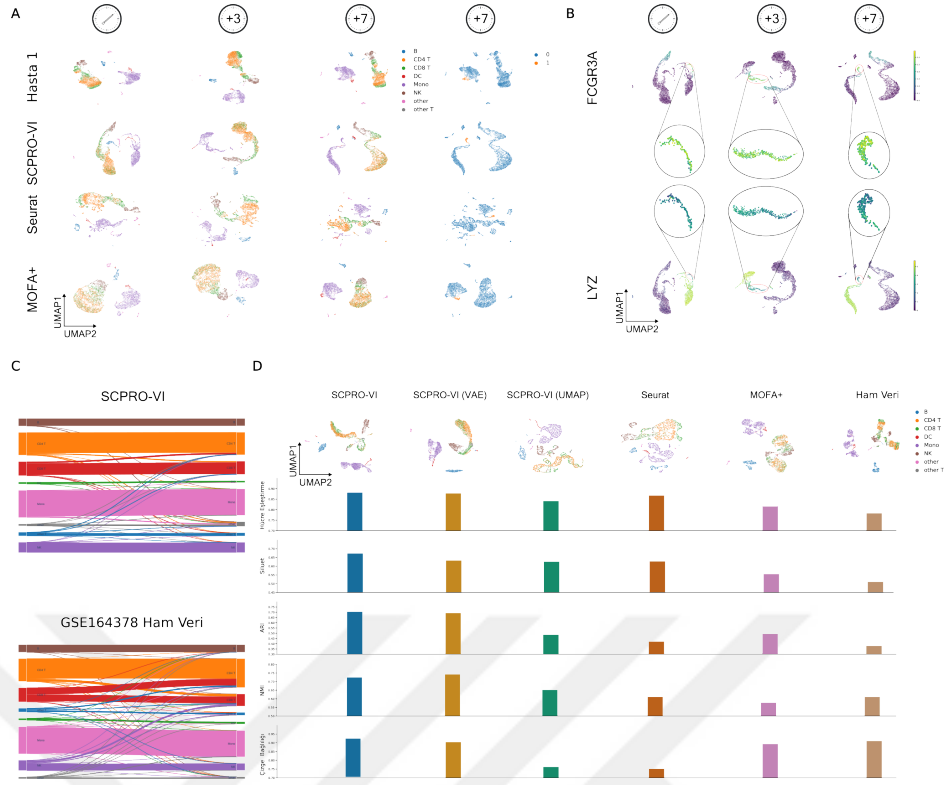
3.3.3. HIV Hastalarında Aşılama Sonrası Değişimlerin İncelenmesi

Aşılamadan sonra transkriptomik ifade ve proteomik mevcudiyet profillerinin değişmesi beklenmektedir [72]. Çalışmalar, HIV hastalarında aşılamadan sonra belirli gen ve proteinlerde yukarı veya aşağı değişimler rapor etmektedir [109]. Bu değişimler, patogeneizde çeşitli görevleri üstlenen hücre alt kümelerini oluşturmaktadır [110].

Bütünleştirme sonuçları, aşı sonrası ortaya çıkan hücre alt kümelerini belirlemek amacıyla incelenmiştir. Yapılan deneyler ile dikey bütünleştirmenin birincil hedeflerinden biri olan yeni hücre alt kümelerini ayırt etmede geliştirilen algoritmanın etkinliği test edilmiştir. UMAP gösterimleri, tüm hastalar için aşılama sonrası birkaç alt kümenin ortaya çıktığını göstermektedir (Şekil 3.6). Hem yedinci günde ortaya çıkan kümeler hem de üçüncü günde ayrılmaya başlayan ve yedinci günde tamamen ortaya çıkan kademeli ayrışan kümeler tespit edilmiştir.

Daha ayrıntılı analiz için kademeli ayrışan bir monosit alt kümesi incelenmiştir (Şekil 3.7A). Analiz edilen monosit hücre alt kümesi, 7. gün veri seti için (en sağ sütun) ham veri ve algoritma gömmelerinde renklendirilmiştir. Mevcut birkaç alt küme arasından analiz edilen bu küme, gözle bile ayırt edilebilecek kadar belirgin olduğu için seçilmiştir. Ek olarak, monositler HIV enfeksiyonunda iyi çalışılmıştır ve bu durum ifade seviyelerindeki potansiyel değişikliklerin yorumlanmasını kolaylaştırmaktadır. İlk olarak, hedef alt küme içindeki hücreleri belirlemek için hücreler kümelendirilmiştir. Daha sonra, hücre tipi içi ayrışmaya katkıda bulunan en ayırt edici proteinler ve genler belirlenmiştir (Çizelge 3.4).

Ayrıştırmacı öznelik analizinde, FCGR3A'nın alt kümenin diğer monositlerden ayrılmasında en önemli protein olduğu tespit edilmiştir. Bu protein, HIV bulaşma ve enfeksiyon süreçlerinde aktif rol oynadığı bilinen bir belirteçtir [111, 112]. Analiz tüm monositler ile diğer tüm hücreler arasında tekrarlandığında, bu protein 174 protein



Şekil 3.7: Geliştirilen algoritmanın HIV veri seti üzerindeki etkinlik testi sonuçları. **A** Hasta 1’in ham ve bütünleştirilmiş ölçümlerinin UMAP gösterimleri. **B** SCPRO-VI gömmelerinin FCGR3A proteini (üst) ve LYZ geni (alt) ifade seviyelerine göre renklendirilmiş UMAP gösterimleri. **C** SCPRO-VI gömmeleri (üst) ve ham veriler (alt) kullanılarak tüm HIV veri setindeki hücrelerin gerçek türleri (sol) ile en yakın komşuları (sağ) arasındaki eşleşmelerin Sankey diyagramları. **D** HIV veri setinden örneklenen veri kümesine ait UMAP gösterimleri (üst) ve nümerik performans sonuçları (alt).

arasında önem sıralamasında 80. sırada yer almıştır. Bu nedenle, ayrıışan hücrelerin FCGR3A ile ilişkili bir fonksiyon için özelleşmiş olabileceği düşünülmektedir.

Benzer analiz genler ile tekrarlanmıştır ve LYZ geninin bu alt küme için en ayırt edici gen olduğu belirlenmiştir. Bu gen, HIV tedavisine olumlu yanıt veren hastaları olumsuz yanıt verenlerden ayırt etmek için kullanılan beş temel belirteçten biri olarak rapor edilmiştir [113]. Ayrıca, LYZ, monositleri diğer hücrelerden ayırt etmede beşinci en önemli gen olarak sıralanmıştır, bu da bu özelleşmiş alt kümeyi yalnızca gen ifade seviyelerine dayanarak tanımlanın zor olduğunu göstermiştir.

FCGR3A proteininin ve LYZ geninin tüm hücreler arasındaki ifade seviyeleri, hücre türlerindeki profillerini analiz etmek için incelenmiştir (Şekil 3.7B). FCGR3A’nın yalnızca hedef alt kümede yukarı regüle edildiği ve tüm doğal öldürücü (NK) hücrelerde

aşağı yönlü değişim gösterdiği gözlemlenmiştir. Buna karşılık, LYZ geni yalnızca monositlerde ve dendritik hücrelerde (DC) aktiftir ve her iki hücre grubunda yukarı yönlü değişim göstermiştir. Ancak, analiz edilen alt kümede aşağı yönlü değişim göstermiştir. Bu sonuçlar, FCGR3A'nın yukarı yönlü değişimi ile LYZ'nin aşağı yönlü değişiminin yalnızca hedef alt kümeye özgü olduğunu göstermektedir.

Hem ayırt edici özellik analizleri hem de ifade seviyesi profili analizleri, transkriptomik ve proteomik verilerin bütünleştirilmesiyle tanımlanan hedef hücre alt kümesinin özelleşmiş bir altküme olduğunu desteklemektedir. Ayrıca, LYZ geni ve FCGR3A proteini, HIV enfeksiyonlu hastalarda bağışıklık hücresi tükenmesi üzerine yapılan çalışmalarda CD16+ monosit alt kümelerinin ortak farklılaşma belirteçleri olarak tanımlanmıştır [110]. Bu bulgular, SCPRO-VI'nin çoklu-omik veri bütünleştirme yoluyla özelleşen hücre alt kümelerini ortaya çıkarmadaki etkinliğini vurgulamaktadır. Hasta 1'in uç uca eklenen ham modalite ölçümlerini ve diğer karşılaştırma algoritmalarının bütünleştirme sonuçları üzerinde hedef altküme oluşumları incelenmiştir (Şekil 3.7A). Ham verilerde, farklı zamanda dizilenen hücrelerin

Çizelge 3.4: Hücre tipi içi ayrışmaya katkıda bulunan en ayırt edici proteinler ve genler.

Protein	F değeri	Gen	F değeri
ENSG00000203747	3919.51	ENSG00000090382	4291.26
ENSG00000007312	920.03	ENSG00000129757	2897.25
ENSG00000170458	693.07	ENSG00000170345	1541.94
ENSG00000197471	333.26	ENSG00000188290	1505.39
ENSG00000135218	304.46	ENSG00000163220	1453.29
ENSG00000182578	203.31	ENSG00000143546	1453.24
ENSG00000185291	199.36	ENSG00000155366	1208.11
ENSG00000168329	168.97	ENSG00000038427	1190.02
ENSG00000115232	162.58	ENSG00000204482	1053.88
ENSG00000169442	159.64	ENSG00000105372	994.71

kümelenmesinde önemli bir değişiklik yoktur. Ancak, analiz edilen alt küme içindeki hücreler 7. gün ölçümlerinde birbirine daha yakın görünmesine rağmen belirgin bir alt küme oluşturmamıştır.

Mofa+, kademeli olarak daralan hücre kümeleri oluşturmakla beraber hedef hücreler, diğer monositlerden tamamen ayrılma gösterememiştir. Seurat, tüm üç veri setinde birden fazla monosit alt kümesi oluşturmuştur. Ancak, hedef alt kümenin UMAP gösterimi, algoritmanın bunu iki farklı kümeye böldüğünü göstermiştir. Bu bulgular, çoklu-omik bütünleştirme yoluyla özel alt kümeleri ortaya çıkarmanın zorlu olduğunu ve hedef altkümenin sadece geliştirilen yöntem ile ortaya çıkarılabildiğini göstermektedir.

3.3.4. Çoklu-Örnek Veri Kümelerinde Dikey Bütünleştirme

Birden fazla örnekten gelen hücrelerin bütünleştirilmesi, karışık hücre popülasyonlarına neden olan biyolojik varyasyonlar nedeniyle düşük performans gösterebilmektedir. Bütünleştirilmiş uzaklıklar kullanılarak her bir hücre tipinde 1-NN hücre eşleşme performansları çoklu örnek veri setinin tamamı kullanılarak değerlendirilmiştir (Şekil 3.7C). SCPRO-VI'nin eşleştirme sonuçları, hücreler arası heterojenlikteki iyileştirmeleri değerlendirmek için uç uca eklenen ham veri ile hesaplanan eşleştirmeler ile karşılaştırılmıştır. Ham veride hücreler arası uzaklıkları hesaplamak için kosinüs uzaklığı kullanılmıştır. Bu metrik yaygın uzaklık metrikleri arasında en iyi sonuçları vermesi sebebiyle tercih edilmiştir (Çizelge A.2).

Sankey diyagramı (Şekil 3.7C), hücrelerin gerçek tiplerini (sol) ve onların 1-NN komşu eşleşmelerini (sağ) göstermektedir. Diyagramın genel görünümü, daha az kesişen çizgiyle gösterilen eşleşme hatalarındaki azalmayı vurgulamaktadır. Ham veriyle yapılan hatalı hücre eşleşmelerinin sayısı, bütünleştirilmiş uzaklıklar dikkate alındığında 8 hücre tipinin 5'inde (CD4 T, CD8 T, Diğer T, NK, Diğer) önemli ölçüde azalmıştır (Çizelge 3.5). Öte yandan, B hücreleri için %0.025, DC hücreleri için %0.05 ve monositler için %0.004 oranında eşleşme doğruluğunda düşüş görülmüştür.

DC ve monosit hücrelerindeki hatalar, küme sınırları gibi belirsiz bölgelerdeki hatalardan kaynaklandığı düşünülmektedir. Ancak, B hücreleri için yapılan 425 hatalı eşleşmenin neredeyse tamamı (413'ü), Diğer T hücreleriyle olmuştur (Çizelge 3.6). Hatalı eşleşen hücrelerin, ana B hücre kümesinden ayrılmış ve bu hücrelerin Diğer T hücrelerine

yakın olduğu gözlemlenmiştir. Geliştirilen algoritma, bu kesişen B hücrelerini Diğer T hücrelerinden ayırt etmekte zorlanmıştır, çünkü bu hücreler B hücresi olarak etiketlenmiş olsalar da Diğer T hücrelerine oldukça benzemektedir.

Ham veriler, 18664 NK hücresinin 6066'sını hatalı eşleştirmiştir. En fazla hatalı eşleşme, 3366 hata ile CD8 T hücre tipinde olmuştur. SCPRO-VI, hatalı eşleşme sayısını 371'e düşürmüştür. Toplam eşleşme performansı, %67'den %98'e önemli ölçüde artarak bu hücreleri T hücrelerinden ayırt etmeyi çok daha kolay hale getirmiştir.

Son olarak, Diğer T hücreleri, ham verilerde CD4 T hücreleriyle %35 ve CD8 T hücreleriyle %23 oranında hatalı eşleşmiştir. Algoritmamız, Diğer T hücrelerinin CD4 T hücreleriyle eşleşme oranını %8'e ve CD8 T hücreleriyle eşleşme oranını %7'ye düşürmeyi başarmıştır.

Algoritmamız, en yüksek iyileştirmeleri Diğer T ve CD8 T hücre tiplerinde sırasıyla %88 ve %70 ile sağlamıştır. CD8 T hücreleri, diğer hücre tipleriyle %52 eşleşme hatası oranına sahiptir. Ham verilerde CD8 T hücreleri, CD4 T hücreleriyle kesişim göstermektedir ve CD8 T hücrelerinin %92'sinin hatalı eşleşmesi CD4 T hücreleriyle olmuştur. Önerilen uzaklık metriği, CD8 ve CD4 T hücreleri arasındaki hatalı eşleşmeleri %88 oranında azaltarak, CD8 T hücrelerinin net bir şekilde ayırt edilmesini sağlamıştır.

Çizelge 3.5: HIV verisetinin bütünlük ölçümleri ile tip bazında hücre eşleştirme başarımları.

Hücre Tipi	B	CD4 T	CD8 T	DC	Mono	NK	Diğer	Diğer T
B	13375	1	0	0	8	3	413	0
CD4 T	3	37024	2112	0	6	15	1594	247
CD8 T	0	2991	21036	0	2	95	928	417
DC	1	2	0	3250	220	0	116	0
Mono	1	5	0	111	46768	0	2125	0
NK	1	18	95	0	3	18293	125	0
Diğer	5	108	29	5	461	22	464	28
Diğer T	0	603	539	0	0	152	0	0

CD8 T hücre eşleşme doğruluğuna yapılan katkıyı değerlendirmek için ek bir deney gerçekleştirilmiştir. Proteomik uzaklıkları önerilen uzaklık metriğiyle hesaplamak yerine her iki modalite için kosinüs uzaklıkları kullanılarak bütünleştirilmiş uzaklıklar hesaplanmıştır. Sonuçlar, CD8 T hücrelerinin hatalı eşleşme oranının %17'den %25'e çıktığını göstermiştir. Bu bulgular, önerilen yeni uzaklık metriğinin 1-NN hücre eşleşme doğruluğunu artırdığını göstermektedir. Ayrıca, modalitelerin bütünleştirilmek yerine ayrı ayrı ele alınmasının bütünleştirme açısından daha etkili olduğunu ortaya koymuştur. Bu sonuçlar, önerdiğimiz yöntemin benzer hücre tiplerini kümelemedeki etkinliğini ve genel 1-NN hücre eşleşme doğruluğunu %80.7'den %91.3'e çıkardığını göstermektedir. SCPRO-VI'nin çoklu örnek veri setleri üzerindeki bütünleştirme performansı, ek performans metrikleri kullanılarak da ölçülmüştür. Ancak, bireysel örneklerle tutarlı bir boyut sağlamak için HIV veri setindeki 161764 hücreden rastgele 6 bin hücre örneklenmiştir. Daha önceki testlerde olduğu gibi, Seurat ve Mofa+ algoritmaları da bu veri setinde karşılaştırmalara dahil edilmiştir (Şekil 3.7D). Ayrıca, aynı mimariyi sahip fakat çizge evrişim katmanları yerine doğrusal katmanlar kullanılan bir varyasyonel oto-kodlayıcı versiyonu da test edilmiştir. Son olarak, bütünleştirilmiş uzaklıklar kullanılarak üretilen UMAP gömmeleri de teste dahil edilmiştir. VAE modeli ve UMAP gömmeleri, çizge oto-kodlayıcı modeli kullanmanın katkısını değerlendirmek için

Çizelge 3.6: HIV verisetinin ham ölçümleri ile tip bazında hücre eşleştirme başarımları.

Hücre Tipi	B	CD4 T	CD8 T	DC	Mono	NK	Diğer	Diğer T
B	13730	34	19	0	3	1	8	5
CD4 T	6	36112	4537	0	7	30	73	236
CD8 T	9	12160	12307	0	3	142	51	797
DC	7	4	6	3420	144	2	6	0
Mono	11	13	13	1006	46997	0	962	8
NK	9	1112	3366	0	16	12598	67	0
Diğer	12	211	61	11	482	9	234	21
Diğer T	7	2380	1601	0	3	94	0	0

karşılaştırmaya dahil edilmiştir.

SCPRO-VI, doğrusal modele kıyasla 1-NN hücre eşleşme doğruluğunda %1 ve UMAP gömmelerine kıyasla %4 artış sağlamıştır. Ayrıca, SCPRO-VI, Seurat algoritmasından daha iyi performans göstererek %88.1 eşleşme doğruluğuna ulaşmış ve Seurat'ın ulaştığı değerden %1.5 daha yüksek sonuçlar elde etmiştir. Mofa+, hücreleri %81.5 doğrulukla eşleştirmeyi başarmış ve ham veri performansını yalnızca %2 artırmıştır.

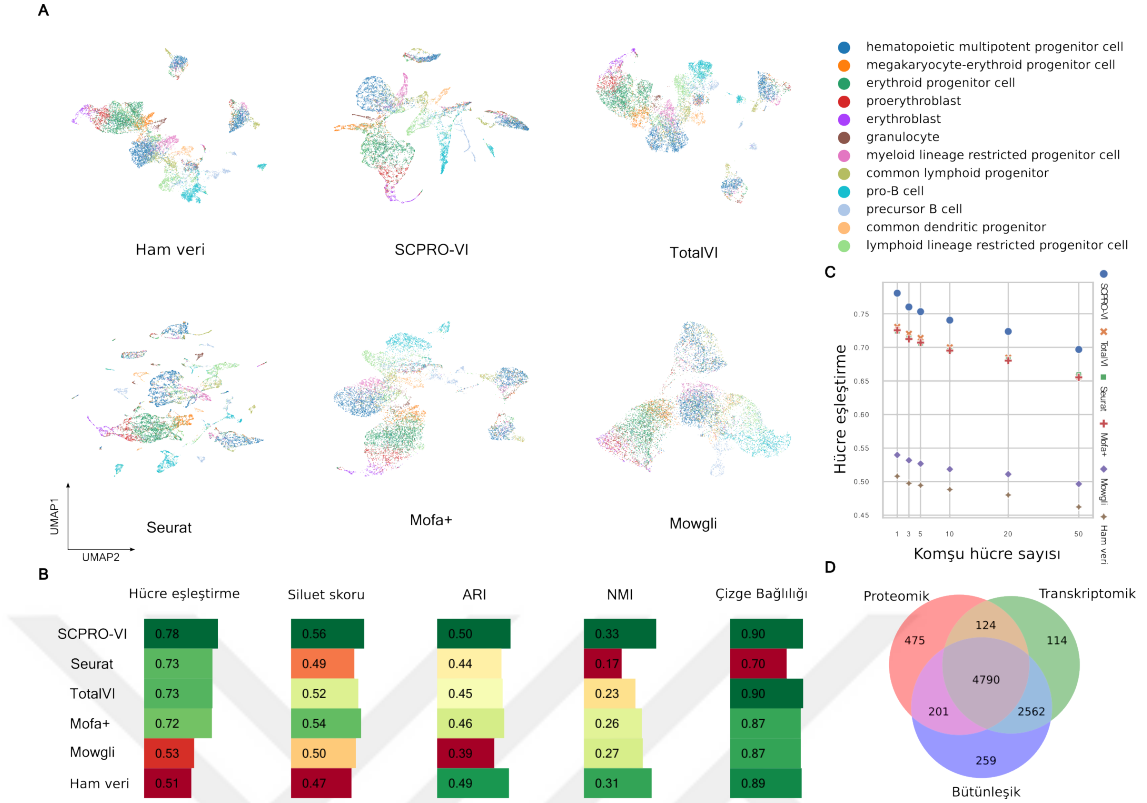
SCPRO-VI, ortalama siluet skoru (ASW) ve Düzeltilmiş Rand İndeksi (ARI) açısından da en iyi performansı sergilemiş ve doğrusal modele kıyasla sırasıyla %5 ve %1 iyileşme sağlamıştır. Ancak, doğrusal model, NMI'de %1.5 ve çizge bağlantısı metriklerinde %2 daha iyi performans göstermiştir. Bu sonuçlar, doğrusal modelin, hem sayısal sonuçlar hem de UMAP gösterimlerinde görüldüğü üzere, küme içi uzaklıkları daha küçük olan gömmeler ürettiğini göstermektedir. Buna karşılık, siluet ve ARI skorları, SCPRO-VI'ye kıyasla doğrusal modelde hücreler arası heterojenliğin azaldığını göstermektedir. Bu gözlem, iki model arasındaki farklılıkla uyumludur; çünkü Varyasyonel Çizge Oto-Kodlayıcı (VGAE) model, hem kayıp fonksiyonu hem de evrişim işleminde kullanılan toplama işlemi aracılığıyla birbiri ile benzer olmayan hücre çiftlerini ayırt etmeyi sağlamaktadır.

Karşılaştırma sonuçları, bütünleştirilmiş uzaklıkları gömmek için derin öğrenme modellerinin UMAP'a üstünlüğünü ortaya koymaktadır. Ayrıca, bu sonuçlar, modelin komşuluk bilgileriyle bilgilendirilmesi durumunda bütünleştirme performansının iyileştiğini ve bu bilginin ilişkili olmayan hücre çiftleri arasındaki benzerlikleri en aza indirmeye yardımcı olduğunu göstermektedir.

3.3.5. Yüksek Performanslı Algoritmalar ile Karşılaştırma

SCPRO-VI algoritmasının çoklu-omik bütünleştirme performansı, [99]'deki karmaşık çoklu-omik veri seti kullanılarak yüksek performanslı algoritmalarla karşılaştırılmıştır. Veri setindeki farklı hücre kümeleri arasındaki girişimler sebebiyle hücreler arası heterojenliğin artırılması zorluğu sebebiyle tercih edilmiştir.

Test veri seti [99], 12 hücre tipine ait 28 bin mononükleer hücreyi (MNC) ve 14 bireyden alınmış örnekleri içermektedir. Kemik iliği örnekleri, farklı gelişim aşamalarında dört yetişkinden ve dört fetüsten alınmıştır. Karaciğer örnekleri altı fetüsten toplanmıştır.



Şekil 3.8: MNC veri seti üzerinde algoritmalara ait performans kıyaslama sonuçları. **A** Algoritma bütünleştirmelerinin hücre tipine göre renklendirilmiş UMAP gösterimleri. **B** Algoritmaların nümerik karşılaştırma metriklerindeki performans sonuçları. **C** Artan komşuluk boyutlarında algoritmaların saflık analizi. **D** Transkriptomik, proteomik ve SCPRO-VI bütünleştirmeleri ile hücre eşleştirme doğruluk sayılarını gösteren küme diyagramı.

Ayrıca, dört örnek yeni doğan göbek kordonu kanından elde edilmiştir. Bu veri setinin çoklu örnek ve çoklu organ yapısı, aynı tür hücrelerin birden fazla kümeye ayrılmasına neden olmuştur. Öte yandan, kümeler önemli ölçüde örtüşmektedir, çünkü tümü mononükleer hücrelerdir. Bu özellikler, veri setini çoklu-omik veri bütünleştirme için uygun hale getirmektedir.

Önerilen algoritma, öncelikle MNC veri setinin bütünleştirilmesinde de test edilen Seurat v4 [72], Mofa+ [88] ve ham verilerin uç uca eklenmesine karşı test edilmiştir. Seurat, çoklu-omik veri bütünleştirme çalışmalarında en yaygın kullanılan kıyaslama algoritmasıdır. Bu yöntem, farklı modaliteler arasında komşulukların bütünleştirilmesine odaklandığı için bütünleştirilmiş uzaklık yaklaşımımızla benzerlik göstermektedir. Bu durum Seurat algoritmasını güçlü bir karşılaştırma adayı yapmaktadır. Mofa+, çoklu-omik tek-hücre veri setlerinin dikey bütünleştirilmesi için güçlü bir diğer

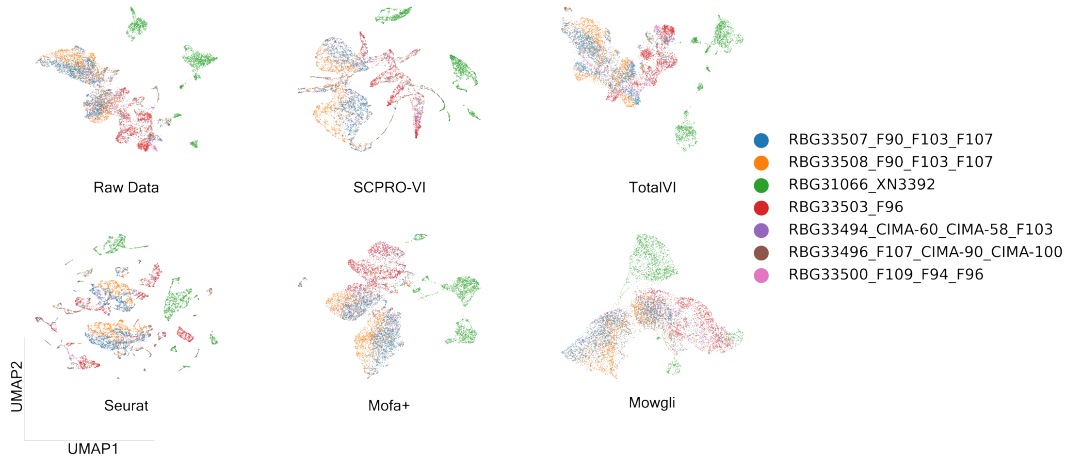
araçtır. Bu yöntem, derin öğrenme tabanlı modelimizi matris faktörizasyon tabanlı bir algoritmaya karşı test etmek için kullanılmıştır.

Karşılaştırmanın kapsamını genişletmek için iki algoritma daha kıyaslamaya dahil edildi. TotalVI [92], VAE modellerine dayanan bir üretken derin öğrenme modeli olması sebebiyle dahil edilmiştir. Algoritma, SCPRO-VI'deki özgün özelliklerin sağladığı potansiyel performans iyileştirmelerini değerlendirmemize olanak tanımaktadır. Son olarak, Çoklu-Omik Çizge Öğrenme (Mowgli) [90], Negatif Olmayan Matris Faktörizasyonunu Optimal Taşıma ile birleştiren güncel bir algoritma olarak testlere dahil edilmiştir. Mowgliyi geliştiren araştırmacılar algoritmalarının rakiplerine karşı verimliliğini test eden kapsamlı deneyler gerçekleştirmiştir. Mowgli'nin dahil edilmesiyle yüksek performanslı diğer algoritmalara karşı da öngörü oluşturulmuştur.

Bütünleştirme sonuçları, UMAP yöntemi kullanılarak görselleştirilmiş (Şekil 3.8A) ve 1-NN hücre eşleşmesi, ASW, NMI, ARI ve çizge bağlantısı metrikleri nümerik karşılaştırma için hesaplanmıştır (Şekil 3.8B). UMAP gösterimleri algoritmaların çözüm yaklaşımlarına dair iç görüler sağlayarak sayısal karşılaştırma sonuçlarının açıklanmasını kolaylaştırmaktadır. SCPRO-VI ve TotalVI, küme içi hücre uzaklıklarını en aza indirerek farklı türdeki hücre kümeleri arasındaki kesişimleri azaltmayı başarmıştır. Seurat algoritması, ortak lenfoid öncüller ve lenfoid soy kısıtlı öncüller gibi bazı hücre tipleri için heterojenliği artırmaktadır. Öte yandan algoritma, aynı tür hücreleri birbirinden uzak konumlarda yer alan alt kümelere ayırmaktadır. Bu küme içi parçalanma, farklı hücre türleri arasında kapsamlı bir ayırım gerektiren ileri analizlerde zayıf veya hatalı sonuçlara yol açabilmektedir.

Matris faktörizasyon tabanlı yöntemler olan Mofa+ ve Mowgli, farklı türdeki hücre kümeleri arasında zayıf bir ayırım ile benzer sonuçlar elde etmiştir. Ancak, Mofa+, aynı tür hücreleri bir araya getirme konusunda Mowgli'den daha iyi performans gösterirken, Mowgli hücre kümelerini büyük ölçüde karıştırmıştır. Öte yandan, algoritmaların hiçbiri farklı örneklerden gelen aynı tür hücre kümelerini bir araya getirememiştir (Şekil 3.9). Ancak, SCPRO-VI ve Mowgli, ham veride zayıf bağlantılara sahip olan farklı veri setlerindeki öncül B hücre kümelerini bir arada tutmayı başarmıştır.

SCPRO-VI, tüm metriklerde rakiplerine karşı en yüksek performans sonuçlarını elde etmiştir (Şekil 3.8B). Algoritmamız, ham verinin 1-NN hücre eşleşme skorunu %53



Şekil 3.9: Algoritmaların bütünleşik ölçümlerinin örnek etiketleri ile renklendirilmiş UMAP gösterimleri.

artırmıştır. En yakın rakipler TotalVI, Seurat ve Mofa+, yaklaşık %7 daha düşük 1-NN hücre eşleşme skorlarına sahiptir ve geliştirilen yöntemin %78 eşleşme skoruna ulaşamamıştır. Mowgli, UMAP’lerde görüldüğü üzere farklı tip hücre kümeleri arasındaki kesişmeler nedeniyle ham veri üzerinde kısıtlı iyileştirme göstermiştir.

Artan komşuluk boyutlarında hücre eşleşme kalitesi de test edilmiştir (Şekil 3.8C). Sonuçlar, komşu sayısı arttıkça tüm algoritmaların benzer performans düşüşleri yaşadığını göstermiştir. Ancak, 20 en yakın komşu dikkate alındığında, SCPRO-VI diğer algoritmaların 1-NN eşleşme için elde ettiği en iyi performanslara benzer bir eşleşme performansına sahip olmuştur. Saf yerel komşuluk, diğer metriklerde daha yüksek performans sonuçlarına yol açmaktadır çünkü benzer hücre türlerini gruplamak, dayanıklı bir bütünleştirme için esastır.

Siluet skoru, hücreler arası heterojenlik kalitesini göstermektedir ve kıyaslama veri setinde farklı hücre kümeleri arasındaki yüksek kesişme nedeniyle genellikle düşüktür. SCPRO-VI, %56 siluet skoru ile ikinci en iyi algoritma olan Mofa+’dan %4 daha iyi sonuçlarla en iyi performansı göstermiştir. TotalVI, %52 siluet skoru ile ortalama bir iyileşme göstermiştir. Mowgli ve Seurat ise sırasıyla %50 ve %49 ile benzer düşük sonuçlar elde etmiştir. Daha önce belirtildiği gibi, Seurat, aynı hücreleri birden fazla alt kümeye bölmüş ve bu alt kümeler birbirinden uzak konumlanmıştır. Bu uzak hücre kümeleri, algoritmaların siluet skorunu düşürmektedir çünkü ortalama siluet skoru

metriği, her hücre türü için bir küme merkezi dikkate alınarak hesaplanır. Son olarak, Mowgli, karışık hücre kümeleri nedeniyle en düşük performansı göstermiştir.

Algoritmamız, hücre kümelenmesinin SCPRO-VI gömmelerindeki gerçek hücre kümelenmesiyle uyumunu gösteren ARI ve NMI metriklerinde en iyi sonuçları üretmiştir. SCPRO-VI, ham veri performansını %2 artırmayı başaran tek algoritmadır, diğer tüm algoritmalar ise ham veriden daha kötü bir uyum göstermiştir. Mowgli hariç rakipler benzer performanslar sergilemiştir. Mowgli, hücre tiplerini ayırt edemediği için en düşük ARI skoruna sahip olmuştur. Ayrıca, Seurat hücre tipi içi parçalanma nedeniyle en düşük NMI skorunu elde etmiştir.

Çizge bağlantısı skorları, düşük NMI skorları ile aynı nedenden ötürü Seurat hariç tüm kıyaslama algoritmaları için çok yakındır. SCPRO-VI ve TotalVI, %90 çizge bağlantısı skoru ile ham verilere göre %1 daha iyi gömmeler üretmiş ve aynı tür hücrelerin ham ölçümlere göre daha bağlantılı kümeler oluşturacak şekilde bir araya toplandığını göstermiştir. Mofa+ ve Mowgli, aynı tür hücreleri bir araya getirmeyi başarsalar da %87 ile biraz daha düşük bağlantı skorları elde etmiştir.

Kıyaslama sonuçları, SCPRO-VI algoritmasının çoklu-omik veri bütünleştirme konusundaki etkinliğini göstermektedir. Algoritmamız, aynı tür hücrelerin küme içi uzaklıklarını azaltırken, farklı tür hücre kümeleri arasındaki heterojenliği artırmayı ölçen beş performans metriğinde yüksek performanslı kıyaslama algoritmalarını geride bırakmıştır. Ayrıca, SCPRO-VI'nin TotalVI'ye karşı belirgin üstünlüğü, yöntemin performansının sadece bir derin öğrenme yöntemi olmanın ötesine geçtiğini göstermektedir. Her bir modalitenin bireysel olarak ele alınması ve önceden hesaplanmış benzerlik çizgilerinin bütünleştirmeye dahil edilmesi algoritmanın performansını artırmaktadır. Yapılan deneyler ile SCPRO-VI algoritmasının çoklu-omik veri bütünleştirmenin iki temel amacını karşılama açısından kıyaslama algoritmalarına kıyasla en iyi performansa sahip olduğu gösterilmiştir.

SCPRO-VI bütünleştirmesi, hücre eşleşmelerine olan katkıyı hesaplamak için tekil modalitelerle de karşılaştırılmıştır (Şekil 3.8D). 1-NN komşuları, her modalitenin ölçümleri ile hücreler arasındaki kosinüs uzaklıkları kullanılarak hesaplanmıştır. Transkriptomik ile uzaklıklar hesaplanırken ham ölçümler yerine PCA gömmeleri kullanılmıştır. PCA gömmeleri, eşleşme performansını %54'ten %76'ya dramatik

bir şekilde artırmıştır.

Sonuçlar, önerilen yöntemin her iki modaliteden de yararlanarak bireysel modalite performanslarından daha yüksek performans elde edebildiğini göstermiştir. Ayrıca, transkriptomik verinin her hücre tipi için proteomikten daha fazla hücreler arası heterojenliği yansıttığı gözlemlenmiştir. Bu durum 176 protein ölçümüne karşılık 11361 gen ölçümü bulunması ile ilişkilendirilmektedir. Sonuçlar ayrıca, hücrelerin %5'inin (475 hücre) proteomik ile doğru eşleştirildiğini ancak hem transkriptomik hem de bütünleştirilmiş uzaklıklar tarafından yanlış eşleştirildiğini göstermiştir. Örneğin, 143 EPC hücresi ve 135 hematopoietik multipotent öncül hücresi, proteomik ölçümler ile doğru bir şekilde eşleştirilmiştir; ancak transkriptomik ile yalnızca sırasıyla 24 ve 29 hücre doğru eşleşmiştir (Çizelge 3.7).

SCPRO-VI bütünleştirmelerinde transkriptomik ilişkilerin proteomik ilişkilere daha baskın olduğu görülmektedir. Ancak, SCPRO-VI, bu anlaşmazlıkları etkili bir şekilde

Çizelge 3.7: Modalitelerin ve bütünleşik ölçümlerin hücreleri doğru eşleştirme sayıları.

Hücre Tipi	IPR	PR	IP	IR	I	P	R
Lenfoid soy kısıtlı progenitör hücre	270	8	24	382	44	23	26
Yaygın lenfoid progenitör	185	3	11	290	27	24	14
Proeritroblast	168	6	21	186	19	45	5
Megakaryosit-eritroid progenitör	185	3	14	181	21	30	10
Hematopoietik multipotent progenitör	1518	40	53	351	37	133	11
Eritroid progenitör	1040	48	41	396	48	120	12
Eritroblast	106		1	83	8	11	4
Prekürsör B hücresi	328	4	6	72	8	21	6
Miyeloid soy kısıtlı progenitör	232	4	9	226	20	26	10
Pro-B hücresi	550	7	18	214	14	28	8
Yaygın dendritik progenitör	147		2	60	6	8	3
Granülosit	61	1	1	121	7	6	5

I: Bütünleşik veri, P: Proteomik ölçümler, R: Transkriptomik ölçümler

yöneterek modalitelerin uzlaşısını %49'dan (4790 + 124 hücre) %75.5'e (4790 + 201 + 2562 hücre) çıkarmayı başarmıştır. Ayrıca, SCPRO-VI, her iki modalitenin de yanlış eşleştirdiği hücrelerin % 15'ini doğru eşleştirmiştir.

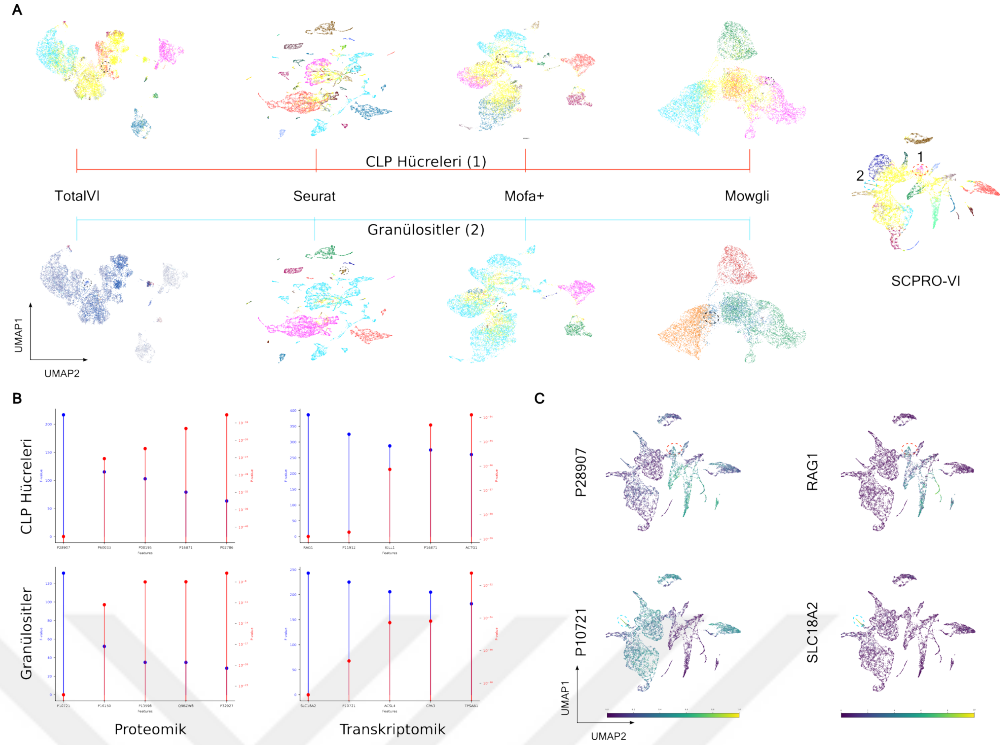
3.3.6. Özgün Hücre Alt Kümelerinin Analizi

Önerilen algoritmanın verimliliğini ve etkinliğini değerlendirmek için kıyaslama veri seti [99] üzerinde ileri analizler gerçekleştirilmiştir. İlk olarak, SCPRO-VI'nin bütünleştirilmiş gömmeleri üzerinde, belirgin hücre alt kümelerini ortaya çıkarmak için DBSCAN [73] algoritmasıyla yoğunluk tabanlı bir kümeleme gerçekleştirilmiştir. 50'den fazla hücreye sahip 18 hücre kümesi arasından, bir granülosit hücre alt kümesi ve bir ortak lenfoid öncül (CLP) hücre alt kümeleri seçilmiştir. Her iki alt küme de hücre tiplerinin ana kümelerinden ayrılmış tek tip hücre içeren alt kümelerdir.

Bu iki hücre alt kümesi, diğer algoritmaların bütünleştirilmiş gömmeleri ile elde edilen alt kümelerde de incelenmiştir. Bu analiz ile SCPRO-VI'nin ham verilerde tespit edilemeyen alt kümeleri ortaya çıkarma kabiliyetinin diğer bütünleştirme yöntemleriyle karşılaştırılması amaçlanmıştır. Granülosit hücre kümesini ham veride bile neredeyse ayrılabilir olduğu için CLP hücrelerinden daha kolay ayırt edilebilmektedir. Öte yandan, CLP alt kümesindeki hücreler, tüm CLP hücrelerinin merkezinde herhangi bir ayrışma olmaksızın yer almaktadır. Bu alt küme, granülosit hücre alt kümesine göre daha zor ayrıştırılabildiği için seçilmiştir.

Her iki hücre alt kümesi için algoritmaların optimal kümelemeleri ayrı ayrı hesaplanmıştır (Şekil 3.10A). Sistemik olarak her algoritma için 20 farklı kümeleme hesaplanmıştır ve bunlar bir F1 eşleşme skoruna göre sıralanmıştır. Bir kümeleme içindeki her küme, CLP hücre alt kümesiyle eşleştirilmiş ve ortak bulunan hücrelere göre F1 skoru hesaplanmıştır. En yüksek F1 skoruna sahip hücre kümesini içeren kümeleme, optimal olarak belirlenmiştir. Aynı işlem granülosit alt kümesi için de uygulanmış ve böylece her algoritma için iki optimal kümeleme elde edilmiştir (Şekil 3.10A). Hedef alt kümedeki hücreler daire içine alınmıştır.

Her iki granülosit ve CLP hücre kümeleri için önerilen algoritmanın F1 skorları 1'dir çünkü bu hücreler, başka hiçbir hücre içermeyen bağımsız hücre kümelerinde kümelenmiştir. Granülosit hücrelerine ait sonuçlar, Seurat algoritmasının bu hücreleri



Şekil 3.10: Bütünleştirilmiş MNC verilerinde ortaya çıkan hücre alt kümelerine ait analiz sonuçları. **A** Algoritmaların kendi optimal kümeleriyle renklendirilmiş bütünleştirilmiş gömmelerinin UMAP gösterimleri. **B** CLP (üst) ve granülosit (alt) hücre alt kümeleri için önemli proteinlerin (*P*) ve genlerin (*F*) değerlerinin Stem diyagramı. **C** SCPRO-VI'nin bütünleştirilmiş gömmelerinin CLP (üst) ve granülosit (alt) hücre alt kümeleri için en ayırt edici proteinin (sol) ve genin (sağ) ifade seviyelerine göre renklendirilmiş UMAP gösterimleri.

neredeyse saf bir hücre kümesinde %0.82 F1 skoru ile kümelemeyi başardığını göstermektedir. Seurat'ın beklenen başarısı algoritmanın hücreleri saf küçük kümelere ayırma eğiliminin bir sonucudur. TotalVI de %76 F1 skoru ile iyi bir performans göstermiştir. Bu VAE tabanlı yöntem, hücre kümelerindeki küçük varyasyonları yakalamada derin öğrenmenin gücünü ortaya koyarken, hücreler arası heterojenliği korumaktadır. Mowgli ve Mofa+, hücre kümelerini karıştırdıkları için granülosit alt kümesini ayırtıramamıştır (her iki algoritma için F1 skoru %0.02).

Granülosit hücre alt kümesinin biyolojik önemini incelemek için ayırt edici gen ve protein analizleri gerçekleştirilmiştir (Şekil 3.10B ve 3.10C). F istatistiği, öznel olarak açıklanan sınıflar arasındaki varyans derecesini nicel olarak ifade eder ve p-değeri, bu varyansın istatistiksel olarak anlamlı olup olmadığını değerlendirmektedir. F istatistiği büyük olduğunda, *p* değeri genellikle küçük olur ve bu da özelliğin iki

sınıfı ayırt etmede son derece önemli olduğunu gösterir. Örneğin, sonuçlar, P10721 proteini ve SLC18A2 geninin diğer özelliklere kıyasla önemli ölçüde daha düşük p değerlere sahip olduğunu göstermektedir, bu da bu özelliklerin granülosit alt kümesini ayırt etmede istatistiksel olarak en önemli öznitelikler olduğunu göstermektedir.

Granülosit hücrelerinde yüksek ifade edilen P10721 gen ve proteini, hücre tipi içi ayırt ediciler olarak ortaya çıkmış ve bu hücre alt kümesinin bazofil, doğal öldürücü (NK) veya eozinofil hücrelere farklılaştığını düşündürmektedir [114]. Ancak, SLC18A2 geninin diğer granülosit hücrelerindeki yukarı yönlü değişimi (Şekil 3.10C), bu hücrelerin bazofil olma olasılığını artırmaktadır [115]. Ek olarak, ayırt edici proteinler olan P16150 (CD43), P13598 (CD102) ve Q9BZW8 (CD244), bu hücrelerin bazofil hücre olma olasılığını desteklemektedir [114, 116]. Sonuç olarak, tespit edilen alt kümenin belirli bir görev için özelleştiği ve algoritmanın bu tür alt kümeleri ortaya çıkarma kabiliyetine sahip olduğu sonucuna varılmıştır.

Aynı işlemler daha zor tespit edilen bir alt küme üzerinde karşılaştırma algoritmalarının başarımını değerlendirmek için CLP hücre alt kümesi için tekrar edilmiştir. Granülosit alt kümesini tespit edemeyen Mowgli ve Mofa+ algoritmaları beklendiği gibi başarısız olmuştur. TotalVI, CLP alt kümesini ortaya çıkarmada %0.16 F1 skoru elde ederek başarısını sürdürememiştir. Seurat algoritması, %70 başarı ile ikinci en iyi algoritma olmuştur. Ancak, tüm CLP hücrelerinin tek bir büyük kümede kümelendiği gözlemlenmiştir; bu da Seurat'ın CLP hücrelerini ayırt edebildiği, ancak hedeflenen alt kümeyi geri kalanından ayıramadığı anlamına gelmektedir. Bu sonuçların tümü, SCPRO-VI'nin, bütünleştirilmiş verilerde ortaya çıkan ve diğer çoklu-omik veri bütünleştirme algoritmaları tarafından ortaya çıkarılamayan nadir hücre alt kümelerini bulma yeteneğine sahip olduğunu göstermektedir.

CLP alt kümesindeki ayırt edici genler ve proteinler de incelenmiştir (Şekil 3.10B). Belirlenen ayırt edici genler RAG1, P11912, P16871, ACTG1 ve IGLL1, ortak lenfoid öncül hücrelerde farklılaşmaya ve alt kümelerin oluşumuna neden olmaktadır [117–119]. Benzer şekilde, belirlenen ayırt edici proteinler P28907, P60033, P16871, P08195 ve P26010, CLP'lerde hücre farklılaşmasına katkı sağladığı literatürde rapor edilmiştir [120, 121]. En ayırt edici gen RAG1 ve en ayırt edici protein P12907, diğer CLP hücrelerinin aksine bu hücrelerde yukarı yönlü değişim göstermiştir (Şekil 3.10C). Ancak, hem RAG1 hem de P12907'un yukarı yönlü değişimi, tespit edilen alt kümedeki

hücreler için özgün olmadığı gözlemlenmiştir.

CLP alt kümesini daha iyi tanımlamak için Seurat v3 [35] algoritması kullanılarak yüksek değişken gen/protein (HVG/HVP) analizi gerçekleştirilmiştir. HVP analizinde en değişken beş gen ZNF300, ITGB7, LINC01954, DIABLO, EHD3 ve beş protein P49961, IgG2b, ICOSLG, SELL, Q9BQ51 belirlenmiştir. IgG2b'nin, CLP hücrelerinin farklılaşması sırasında önemli ifade seviyesi değişiklikleri gösterdiği rapor edilmiştir [122]. Ayrıca, tüm IgG2b, ICOSLG, PDCD1LG2 ve SELL proteinlerinin birlikte yüksek değişkenlik göstermeleri CLP hücrelerinin B hücrelerine farklılaştığını düşündürmektedir [122, 123]. Bu proteinler arasında yalnızca IgG2b, tüm CLP hücrelerinin en yüksek değişken 10 proteininden biri olarak listelenmiştir. Bu durum, bu özel hücre kümesinin kendi HVG'lerine sahip olduğunu göstermektedir. Tüm bu sonuçlar, nadir hücre popülasyonlarının ortaya çıkarabilen için güçlü çoklu-omik gömmeler üreterek önerilen algoritmanın çoklu-omik veri bütünleştirme konusundaki etkinliğini ortaya koymaktadır.

3.4. Sonuç

Tek hücrelerdeki omik katmanların bütünleşik analizi, hücresel çözünürlüğü artırarak, özelleşmiş hücre alt gruplarının tanımlanmasını mümkün kılmaktadır. Ancak, iki farklı veri uzayındaki ilişkilerin ortak bir uzayda gömülmesi, modalitelere özgü ilişkilerdeki dengesizliklerden kaynaklanan çatışmalar nedeniyle zordur. Dahası, etkili bir bütünleştirme, omik katmanlar arasındaki düzenleyici mekanizmalara dair daha derin iç görüler sağlayacak şekilde açıklanabilir ve yorumlanabilir olmalıdır.

Bu bölümde, eşleştirilmiş tek-hücre çoklu-omik verilerinin bütünleştirilmesi için özgün bir araç olan tek-hücre çoklu-omik dikey bütünleştirme (SCPRO-VI) yöntemi sunulmuştur. SCPRO-VI, yapılan çalışmada proteomik ve transkriptomik bütünleştirme üzerine yoğunlaşsa da, nümerik herhangi bir eşleştirilmiş modaliteyi bütünleştirme yeteneğine sahiptir. Algoritma, biyolojik bilgi içeren yeni bir uzaklık metriği kullanılarak hesaplanan omiklere özgü dengelenmiş uzaklıkları çarparak hücreler arasındaki bütünleştirilmiş uzaklıkları hesaplamaktadır. Daha sonra, SCPRO-VI, bu modaliteleri çoklu-görünüm varyasyonel çizge oto-kodlayıcı kullanarak bütünleştirilmiş bir veri uzayına gömmektedir.

SCPRO-VI, çoklu örneklerden elde edilmiş bir CITE-seq veri seti [72] kullanarak dayanıklılık, verimlilik ve etkinlik açısından titizlikle test edilmiştir. Sonuçlar, farklı tür hücre heterojenliğini hesaplamalı performans metriklerine dayanarak tüm örnek veri setlerinde rakip yöntemlere kıyasla üstün performansla arttırabildiğini göstermiştir. Ek olarak, analizler SCPRO-VI'nin küme içi heterojenliği artırarak literatürle uyumlu biyolojik olarak anlamlı hücre alt kümeleri ortaya çıkardığını ortaya koymuştur.

Daha düşük hücre türleri arası heterojenliğe sahip zorlu bir CITE-seq veri seti [99] kullanılarak SCPRO-VI mevcut yüksek performanslı algoritmalarla da karşılaştırılmıştır. Sonuçlar, SCPRO-VI'nin hem hesaplamalı metriklerde hem de özelleşmiş hücre alt kümelerini ortaya çıkarma yeteneğinde mevcut yöntemleri açıkça geride bıraktığını göstermiştir. Ek olarak, yöntem hem modaliteye özgü ilişkileri kullanarak hem de genişleyen komşuluklarda hücre eşleşmesinin saflığını rakiplerinden daha etkili bir şekilde koruyarak üstün bir performans sergilemiştir.

SCPRO-VI algoritması çoklu-omik veri bütünleştirmenin ötesinde, herhangi bir çoklu-modalite veri setinin bütünleştirilmesi için genel bir çözüm sunmaktadır. Algoritmanın temel prensibi, farklı modaliteler arasında ortak bir temsil öğrenerek, bilgi kaybını minimize ederken farklı veri kaynaklarının uyumlu hale getirilmesini sağlamaktır. Örneğin, tıbbi görüntüleme alanında, MRI ve CT taramalarının birlikte değerlendirilmesi veya finans sektöründe, hisse senedi fiyatları ile makroekonomik göstergelerin bütünleştirilerek tahmin modellerinin geliştirilmesi gibi uygulamalar SCPRO-VI'nin potansiyel kullanım alanlarına örnek olarak gösterilebilir. Ayrıca, doğal dil işleme ve veri madenciliği gibi alanlarda metin, ses ve görüntü verilerinin birlikte işlenmesi için SCPRO-VI, modaliteler arası bilgi transferini optimize etmek adına önemli avantajlar sunmaktadır.

Deneylerde gözlemlendiği üzere, yüksek başarımlı algoritmaların performansı farklı veri setleri arasında önemli ölçüde değişiklik gösterebilmektedir. Yapılan karşılaştırmalar, ilgili algoritmaların yazarları tarafından önerilen varsayılan parametreler kullanılarak gerçekleştirilmiştir. Bu durum algoritmaların genellenebilirliğini sınırlayabilmektedir. Ayrıca, veri seti oluşturucuları tarafından genellikle hesaplamalı olarak türetilen hücre tipi bilgisinin metrikler tarafından kullanılması, yanlış etiketlemeden kaynaklanan potansiyel doğruluk sorunlarını da beraberinde getirebilmektedir. Tüm algoritmalar aynı bilgiler çerçevesinde değerlendirilmiş olsa da, yanlış etiketlenmiş hücreler performans

metriklerini olumsuz etkileyebildiğini belirtmek önemlidir.

SCPRO-VI'nin dikkate değer bir kısıtı, özgün uzaklık metriği hesaplamasının getirdiği hesaplama yüküdür. Bu işlem, karesel karmaşıklığa sahiptir. Bu durum, milyonlarca hücre veya modalite başına binlerce özellik içeren veri setleri için zorluklar yaratacaktır. Ancak, bu sınırlama, GPU'larda verimli vektörel işlemler için hücrelerin ve özneliklerin partilere ayrılması gibi paralelleştirme yöntemleriyle hafifletilebilmektedir. Bir diğer zorluk ise modalite öznelikleri arasındaki ilişkileri için kapsamlı ilişki bilgisinin eksikliğidir. Örneğin, protein-protein etkileşim verileri proteomik için yaygın olarak mevcutken, gen-gen etkileşim veritabanları, bireysel çalışmalardaki gelişmelere rağmen [124] henüz yeterince gelişmemiştir.

SCPRO-VI'yi daha da geliştirmek için çeşitli iyileştirmeler yapılabilir. Mevcut durumda, bütünleştirilmiş uzaklıklar basitlik adına dengelenmiş uzaklıkların çarpılmasıyla hesaplanmaktadır. Omiklere özgü uzaklıkları güvenilirliklerine göre dinamik olarak ağırlıklandıran gelişmiş bir algoritmanın geliştirilmesi performansı artırabilir. Ayrıca, bir ön işleme adımıyla belirlenen uyarlanabilir komşuluk boyutunun eklenmesi, uzaklık hesaplamalarını ve benzerlik çizgesi oluşturmayı optimize ederek algoritmanın performansı artırabilir.

Bu genişletmelerin ötesinde, en heyecan verici gelecek çalışmalardan biri eksik modalitelerle başa çıkmak için yapılacak çalışmalardır. scRNA-seq veri setlerine kıyasla, çoklu-omik çalışmaların sayısı ve kapsamı hala sınırlıdır. Çoklu-omik veriler üzerinde eğitilen bir üretici modül ile gömme modelinin geliştirilmesi, proteomik veriler mevcut olmadığında bile bütünleştirmeyi mümkün kılabilir. Bu tür bir genişletme, SPCRO-VI'nin çoklu-omik heterojenliği kapsamlı tek-hücre transkriptomik veri setlerine yansıtmasına olanak tanıyarak uygulanabilirliğini ve etkisini artıracaktır.

3.5. Veri ve Kod Erişilebilirliği

Bu çalışmada analiz edilen veriler, Human Cell Atlas (HCA) veritabanından [107] elde edilmiştir. HIV veri setine <https://explore.data.humancellatlas.org/projects/3ce9ae94-c469-419a-9637-5d138a4e642f> adresinden erişilebilir ve ilgili meta-veriler, Gene Expression Omnibus (GEO) veritabanında GSE164378 erişim numarası altında bulunmaktadır [125]. Mononükleer

hücre verilerine <https://explore.data.humancellatlas.org/projects/04ad400c-58cb-40a5-bc2b-2279e13a910b> adresinden erişilebilir ve meta-veriler GEO'da GSE166895 erişim numarası altında sunulmaktadır. Protein-protein etkileşim skorları, STRING veritabanından <https://stringdb-downloads.org/download/protein.links.v12.0/9606.protein.links.v12.0.txt.gz> adresinden indirilmiştir [100]. Yolak verileri ise Reactome veritabanından https://reactome.org/download/current/UniProt2Reactome_All_Levels.txt adresinden alınmıştır [101].

Bu makaledeki sonuçları yeniden üretmek için gerekli Python kodu <https://github.com/single-cell-proteomic/SCPRO-VI> adresinde mevcuttur.



4. VİRAL PATOGENEZ ANALİZİ

4.1. Giriş

Virüslerin konak organizma üzerindeki hastalık yapıcı etkilerini, bu etkilerin moleküler, hücresel ve organizma düzeyindeki değişimlerini ve bu değişimleri önleyici veya tedavi edici mekanizmalarına dair tüm çalışmalar viral patogenezi oluşturmaktadır. Virüs kaynaklı hastalıkların yol açtığı ciddi sağlık sorunları ve pandemik potansiyelleri, bu alandaki çalışmaların önemini daha da artırmaktadır. Virüslerin konak hücrelerle etkileşimi ve konak organizmada yayılım mekanizmalarının anlaşılması, yeni tanı ve tedavi yaklaşımlarının geliştirilmesine olanak sağlamaktadır. Ancak, virüslerin karmaşık yaşam döngüsü sebebiyle farklı hücre tipleri ile heterojen ilişkiler kurması, bu süreçlerin anlaşılmasını zorlaştırmaktadır.

Tek-hücre omik veri kümelerinin bütünleştirilmesi, viral patogenezi çalışmalarında enfeksiyon süreçlerini hücresel düzeyde daha ayrıntılı inceleme fırsatı sunmaktadır. Geleneksel yöntemler, büyük ölçüde toplu dizileme verilerine dayandığından, bireysel hücrelerin enfeksiyona verdiği tepkilerin heterojenliğini yakalamakta yetersiz kalmaktadır. Tek-hücre seviyesindeki dizileme teknolojilerinin gelişmesi, bu eksikliği gidermekte ve virüs-hücre etkileşimlerinin detaylı analizine olanak tanımaktadır. Enfeksiyon süreçlerinin moleküler düzeyde anlaşılması, farklı hücre tiplerinin enfeksiyona verdiği yanıtları ortaya çıkarmakta ve enfeksiyon dinamiklerinin çözümlenmesine katkı sağlamaktadır.

Bu çalışmada, viral patogenezi anlamak amacıyla hasta hücrelerinin bağışıklık aktiflik seviyeleri belirlenerek farklılaşan hücre kümelerindeki ayırt edici biyobelirteçler incelenmiştir. Farklılaşan altkümeleri keşfetmek amacıyla çoklu-omik veriler geliştirilen varyasyonel çizge oto-kodlayıcı tabanlı SCPRO-VI yöntemi ile bütünleştirilmiştir. Hücre enfeksiyon seviyelerinin tespitinde, PHISTO veri tabanındaki [134] etkileşimlerden elde edilen virüs hedef proteinlerinin yukarı/aşağı yönlü regülasyonları, protein ölçümler üzerinde z-skor hesabı ile analiz edilmiştir. Bu skorlar, literatürdeki enfeksiyon sonrası beklenen değişimlerle karşılaştırılmış, her bir hücrenin eşleşen değişim yönü sayısı enfeksiyon seviyesi olarak belirlenmiştir.

Yüksek enfeksiyon seviyesine sahip altkümelerde ayrıştırıcı gen / protein analiziyle biyobelirteçler tespit edilmiştir.

Geliştirilen yöntem HIV virüsü ile enfekte olan hastalardan elde edilen proteomik ve transkriptomik ölçümler barındıran tek-hücre çoklu-omik CITE-seq veri kümesi üzerinde test edilmiştir. Gerçekleştirilen testlerde diğer CD4 T hücrelerinden belirgin şekilde daha aktif olmasıyla ayrılan sitotoksik CD4 T hücre altkümesi tespit edilmiştir. HIV enfeksiyonunun kontrol altında tutulmasına yardımcı olduğu bildirilen bu hücre alt kümesi [135] üzerinde önemli ayırıcı biyobelirteç analizi yürütülmüştür. Tespit edilen genlerin HIV enfeksiyonuna karşı bağışıklıkta aktif rol alan hücrelerde yoğun ifade edildiği gözlemlenmiştir. Dahası tespit edilen bazı genlerin sadece bu hücre altkümesinde yüksek ifade seviyesine sahip olduğu gözlemlenmiştir. Geliştirilen yöntem ve elde edilen sonuçlara ait detaylar sonraki bölümlerde sunulmaktadır.

4.1.1. Benzer Çalışmalar

Viral patogenezi analizinde yaygın olarak viral genom analizleri, mikroskopik görüntülemeler ve yeni nesil dizileme ile yüksek çıktılı omik verilerin ölçüm ve analiz teknikleri [126] virüslerin konak hücre üzerindeki etkilerini anlamak için sıklıkla kullanılmaktadır. Bununla birlikte, bu yöntemler çoğunlukla toplu dizileme ile elde edilen verilerle sınırlıdır ve bireysel hücrelerin enfeksiyona verdiği yanıtların heterojenliğini anlamakta yetersiz kalmaktadır. Tek-hücre seviyeli dizileme teknolojilerinin gelişmesi, viral patogenezi analizinde yeni ufuklar açmış ve tek-hücre düzeyinde detaylı analizlerin gerçekleştirilmesine olanak tanımıştır [127].

Enfeksiyonun farklı evrelerinde konak canlıının farklı bağışıklık cevapları verdiği ve virüsün sebep olduğu moleküler değişimlerin hücreden hücreye değişiklik gösterdiği bilinmektedir [128]. Tek-hücre dizileme teknolojileri, hücreler arası heterojenliği ortaya çıkararak virüs-hücre etkileşimlerinin moleküler detaylarla incelenmesine olanak sağlamaktadır. Hücreler arasındaki bu heterojen yapının oluşmasına sebep olan biyomolekül farklılıkları virüs kaynaklı gen ifade seviyesi değişim modellerini hücre seviyesinde tespit etmeyi sağlamaktadır.

Son yıllarda yapılan çalışmalarda tek-hücre transkriptom dizileme (scRNA-seq) teknolojisi enfekte ve enfekte olmayan hücre popülasyonları arasındaki transkripsiyonel

farklılıkları incelemek için yaygın olarak kullanılmaktadır [129]. Virüslerin konak hücre metabolizması üzerindeki düzenleyici etkilerinin analizi [130], hücresel bağışıklık yanıtlarının hücre tipine bağlı manipülasyonlarının incelenmesi [131] ve enfeksiyonun farklı aşamalarında aktifleşen genetik yolların belirlenmesi için tek-hücre seviyeli transkriptom ölçümleri kullanılmaktadır [132].

Tek-hücre teknolojilerinin viral patogenezin açıklanmasına sağladığı bir diğer katkı farklı tip hücrelerdeki bağışıklık yanıtlarının analiz edilebilmesidir. Virüslerin neden olduğu enfeksiyonlar sırasında, konak organizmanın bağışıklık sistemi, doğuştan gelen ve kazanılmış yanıtlar yoluyla enfeksiyonla mücadele eder. Bu süreçte, doğal öldürücü (NK) hücreler, makrofajlar, dendritik hücreler ve T hücreleri gibi farklı bağışıklık hücrelerinin virüse verdiği yanıtlar kritik bir rol oynamaktadır. Tek-hücre seviyeli ölçümler sayesinde farklı tipteki hücrelerin enfeksiyonu kontrol altına almak için faydalandığı biyomolekülleri tespit etmek mümkün olmaktadır [133].

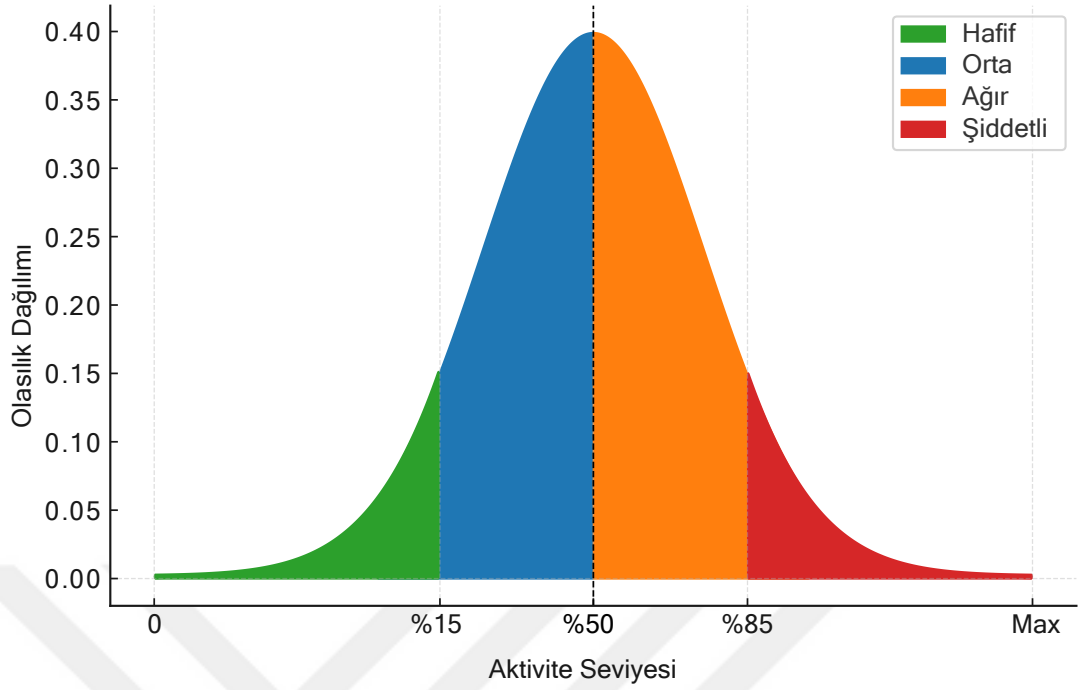
Geliştirilen tek-hücre çoklu-omik dizileme yöntemleri, transkriptomik veriye ek olarak proteomik, epigenomik, uzamsal vb. ölçümleri de aynı hücrelerden eş zamanlı olarak elde etmeyi mümkün hale getirmiştir [82]. Çoklu-omik veriler bütünleştirilerek kapsamlı analizler gerçekleştirilmektedir. Bütünleşik ölçümler viral enfeksiyonun hücresel ve moleküler etkilerini çok yönlü bir bakış açısıyla incelemeyi mümkün hale getirmektedir.

4.2. Yöntem

4.2.1. Hücre Viral Aktivite Seviyelerinin Belirlenmesi

Hücrelerin viral aktivite seviyeleri belirlenirken virüs proteinlerinin hedef aldığı konak proteinleri üzerindeki değişimler incelenmiştir. Hedef proteinlerin enfeksiyon sonrasında yukarı ya da aşağı yönlü regülasyon göstermesi beklenmektedir. Beklenen değişimleri gösteren hücreler enfeksiyon ya da bağışıklık yoluyla viral patogeneze bakımından aktif kabul edilmiştir.

Hedef proteinleri tespit etmek için patojen ile konak arasındaki protein – protein etkileşimleri kullanılmıştır. Analiz edilen tek-hücre veri kümesinde bulunan proteinler arasından veri setinin elde edildiği hastaları enfekte eden virüsün konak canlıda etkileşim kurduğu proteinler hedef protein olarak belirlenmiştir. Her bir hedef proteinin



Şekil 4.1: Viral aktivite sınıf dağılım diyagramı.

enfeksiyon sonrası regülasyon yönü çeşitli veri tabanlarından [107, 136, 137] ve literatür çalışmalarından derlenerek tespit edilmiştir.

Her bir tek-hücrenin viral aktivite seviyesi o hücrede beklenen regülasyon değişimini gösteren hedef proteinlerin sayısı olarak hesaplanmıştır (Şekil 1.3C). Hedef proteinlerin ilgili hücredeki değişim yönü z-skör yöntemi kullanılarak hesaplanmaktadır:

$$Z_i = \frac{X_i^{\text{norm}} - \mu_{\text{protein}}}{\sigma_{\text{protein}}} \quad (4.1)$$

Formül 4.1’de, Z_i i hücresinin her bir proteini için z-skör değerlerini tutan vektörü göstermektedir. Hücreler arasındaki ölçüm derinliği farkını ortadan kaldırmak amacıyla öncelikle her bir hücre kendi içerisinde normalize edilmiştir. X_i^{norm} , i hücreindeki ham protein ölçümlerinin o hücredeki toplam protein miktarına bölünmesi ile elde edilen normalize edilmiş değer vektörünü göstermektedir. μ tüm hücrelerdeki normalize edilmiş ölçümler kullanılarak her bir protein için hesaplanmış ortalama ölçümler vektörüdür. Benzer şekilde normalize edilmiş ölçümler ile hesaplanan proteinlere ait standart sapma değer vektörü σ ile gösterilmektedir.

Z-skör değerleri alttan ve üstten sınırlandırılarak sınır dışında kalan ölçümler yukarı ya

da aşağı olarak etiketlenmiştir. Sınır değerleri belirlenirken $\mu + -2\sigma$ (%95) yaklaşımı kullanılmıştır. Her bir hücredeki yukarı aşağı yönlü değişim gösteren proteinler literatürde beklenen değişim yönü ile karşılaştırılarak, beklenen yönlü değişim gösteren proteinlerin toplam sayısı o hücrenin viral aktivite seviyesi olarak belirlenmektedir.

Benzer aktiflik seviyesine sahip hücrelerin birlikte değerlendirilebilmesi ve etiketlerin sadeleştirilmesi amacıyla aktiflik seviyeleri 4 sınıfa ayrılmıştır: Hafif, orta, ağır ve şiddetli. Aktiflik seviyeleri sınıflara ayrılırken Şekil 4.1'dekine bezer bir normal dağılıma yakınsanmaktadır. Diyagramda max en yüksek aktivite seviyesinin değerini belirtmektedir ve % ile gösterilen değerler max değeri ile çarpıldığında elde edilen değerleri göstermektedir.

Sınıf dağılımları belirlenirken hafif ve şiddetli sınıfları hücrelerin en fazla %15'ini içerecek şekilde aktivite seviyeleri sınırlandırılmıştır. Hafif ve şiddetli sınıfları dışında kalan aktivite seviyeleri eşit iki parçaya bölünerek orta ve ağır sınıflarına atanmıştır. Tek sayıda aktivite seviyesi kalması durumunda orta sınıfında daha fazla aktivite seviyesi kalacak biçimde seviyeler iki sınıfa bölünmüştür.

4.2.2. Viral Patogenezde Aktif Hücre Altkümelerinin Tespiti

Viral patogenezde aktif rol oynayan hücre altkümelerinin tespiti için bütünleştirilmiş çoklu-omik tek-hücre kümeleri kullanılmaktadır. Analiz edilen çoklu-omik tek-hücre veri kümesi geliştirilen tek-hücre dikey bütünleştirme yöntemi (SCPRO-VI) kullanılarak bütünleştirilmektedir (Şekil 1.3B).

Bütünleşik veri üzerinde DBSCAN [73] kümeleme algoritması çalıştırılarak altkümeler elde edilmektedir (Şekil 1.3D). İki hücrenin aynı kümede olabilmesi için izin verilen maksimum uzaklığı temsil eden epsilon parametresi için optimal değer her bir hücrenin en yakın 5. komşusuna olan uzaklıkların %95'inden yüksek olan ilk uzaklık değeri olarak belirlenmektedir. Viral etkileşimler yönünden aktif hücre altkümelerinin tespiti için hesaplanan viral aktivite seviyelerinden faydalanılarak her bir hücre altkümesi için aşağıdaki formül ile bir aktiflik değeri hesaplanmaktadır:

$$Altküme\ skoru(i) = \frac{\sum (4 * \text{şiddetli} + 3 * \text{ağır} + 2 * \text{orta} + \text{hafif})}{\text{Toplam Hücre Sayısı (i)}} \quad (4.2)$$

Formül 4.2’de, i alt kümesinin aktivite seviyesi, altkümenin her bir ağırlık seviyesindeki hücre sayılarının ağırlıklı ortalaması alınarak hesaplanmaktadır. Hücre altkümeleri elde edilen aktiflik skoruna göre sıralanarak en aktif hücre altkümeleri belirlenmektedir.

Viral patogeneizde aktif altkümeler ileri analizlere tabi tutulmadan önce hücre tipi bakımından da değerlendirilmiştir. Çok fazla tipte hücre barındıran altkümeler enfeksiyonda veya bağışıklıkta tüm hücre tiplerinde benzer değişimler gösteren üst seviye biyobelirteçlerin tespiti hedefiyle incelenmektedir. Öte yandan tek ya da az sayıda tip hücre barındıran altkümeler enfeksiyon veya bağışıklık için özelleşmiş hücreler olarak değerlendirilmektedir. Bu altkümelerde yapılan analizler hastalık alt tiplerini belirlemeyi ya da enfeksiyonla mücadelede etkin biyobelirteçleri tespit etmeyi hedeflemektedir.

4.2.3. Hücre Altkümelerinde Biyobelirteç Analizi

Viral patogeneizde aktif hücre altkümelerinde bulunan biyobelirteçler, ilgili altkümedeki hücreler ile kalan diğer tüm hücreler arasında ayırt edici öznelik analizi ile tespit edilmektedir. Ayırt edici öznelik analizi hücrelerdeki ham tekil omik ölçümler kullanılarak gerçekleştirilmektedir. Analiz edilen omik veri tipinde ilgili altkümeyi diğer tüm hücreleri kapsayan kümeden ayırmada etkin olan önemli biyolojik öznelikler (genler, proteinler vb.) viral patogeneizde etkin biyobelirteçler olarak tanımlanmaktadır.

Omik veri içerisindeki önemli özneliklerin tespiti için her bir öznelik ilgili altküme ile diğer tüm hücrelerin altkümesi arasında F testi ve p değeri hesabına tabi tutulmaktadır. F testi ilgili özneliğin kümeler arası varyansını küme içi varyansına oranlayarak bir önem değeri belirlemektedir. Yüksek F skoruna sahip öznelikler daha yüksek kümeler arası farklılık ve daha düşük küme içi değişiklik gösterdiği için altkümeyi diğer tüm hücrelerden ayırmada daha önemli kabul edilmektedir.

Öznelikler F skorlarına göre büyükten küçüğe sıralandıktan sonra p değerleri ile filtrelenmektedir. Özneliklerin p değerleri her iki küme arasında ilgili özneliğin değer dağılımı arasında ne kadar farklılık olduğunu göstermektedir. Hedeflenen biyobelirteçlerin viral aktivitesi yüksek olan hücre kümelerinde farklı karakteristik göstermesi beklendiğinden p değeri düşük öznelikler daha önemli kabul edilmektedir. F skoru kümeler arası varyasyon farkını, p değeri ise bu varyasyon farkının ne kadar önemli olduğunu ölçümlendiği için doğaları gereği F skoru yüksek özneliklerde p

değerinin düşük olması beklenir. Bu sebeple öznitelikler öncelikle F skorlarına göre sıralanmakta ve ardından belirli bir p değerinin altında kalanlar viral patogeneze önemli biyobelirteçler olarak belirlenmektedir.

4.3. Sonuçlar

4.3.1. Deneylerde Kullanılan Veri Kümesi

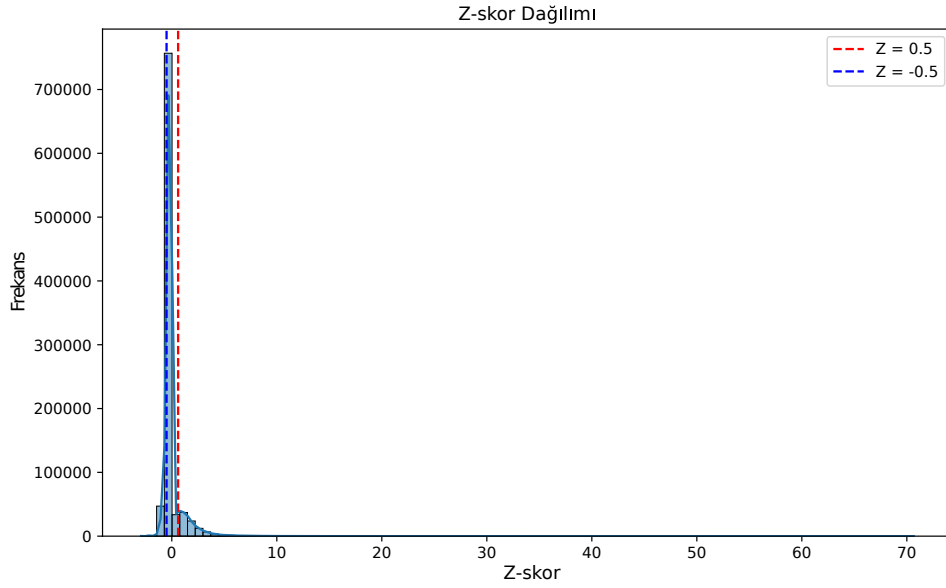
Viral patogeneze analizi için geliştirilen dikey bütünleştirme algoritmasının testlerinde kullanılan HIV alt veri kümesi kullanılmıştır [72]. Bu veri kümesi HIV virüs enfeksiyonuna sahip 8 hastadan elde edilen hücrelerden dizilenen 188 protein ve 20380 gene sahip ölçümler içermektedir. Hesaplama zorlukları sebebiyle veri kümesindeki 167000 hücre arasından 6000 hücre, hücre tiplerine göre hücre sayısı oranları ana kümeyle benzer olacak şekilde örneklenmiştir.

4.3.2. Hücrelerin Viral Aktivite Seviyelerinin Belirlenmesi

Veri kümesindeki proteinler öncelikle PHISTO [134] veri tabanından elde edilen HIV-1 virüsüne ait protein – protein etkileşimlerinde hedef olarak bildirilen proteinler kullanılarak filtrelenmiştir. 188 protein arasından 33 adet protein virüs hedefi olarak tespit edilmiştir. Çizelge A.3’de bu proteinler, beklenen regülasyon yönleri ve referans alınan literatür çalışması sunulmaktadır. Literatürden elde edilen bilgiler doğrultusunda 15 proteinde yukarı, 18 proteinde aşağı yönlü değişim beklenmektedir. 33 protein üzerinde z-skor analizi yapılarak her bir hücrede bulunan proteinlerin yukarı/aşağı değişim yönleri tespit edilmiştir (Şekil 4.2).

Z-skor analizinde eşik değeri olarak 0.5 değeri kullanılmıştır. İşlem sonucunda toplamda proteinlerin 121 bin ölçümde yukarı yönlü değişim, 112 bin ölçümde aşağı yönlü değişim gösterdiği tespit edilmiştir. Hücre bazlı değişimler sayılarak hücre viral aktivite seviyeleri hesaplanmıştır. Hücrelerin aktivite seviyeleri 1 ile 16 arasında değişmektedir. Bu değerler normal dağılıma uygun şekilde dağıtılarak 1 – 4 hafif, 5 – 7 orta, 8 – 12 ağır ve 13 – 16 şiddetli olacak şekilde aktivite sınıfları belirlenmiştir.

Veri kümesi UMAP yöntemi ile görselleştirilerek hücreler elde edilen aktivite sınıfı etiketleri ile renklendirilmiştir (Şekil 4.3). UMAP projeksiyonundan da anlaşıldığı üzere



Şekil 4.2: Proteinlere ait z-skor dağılım grafiği.

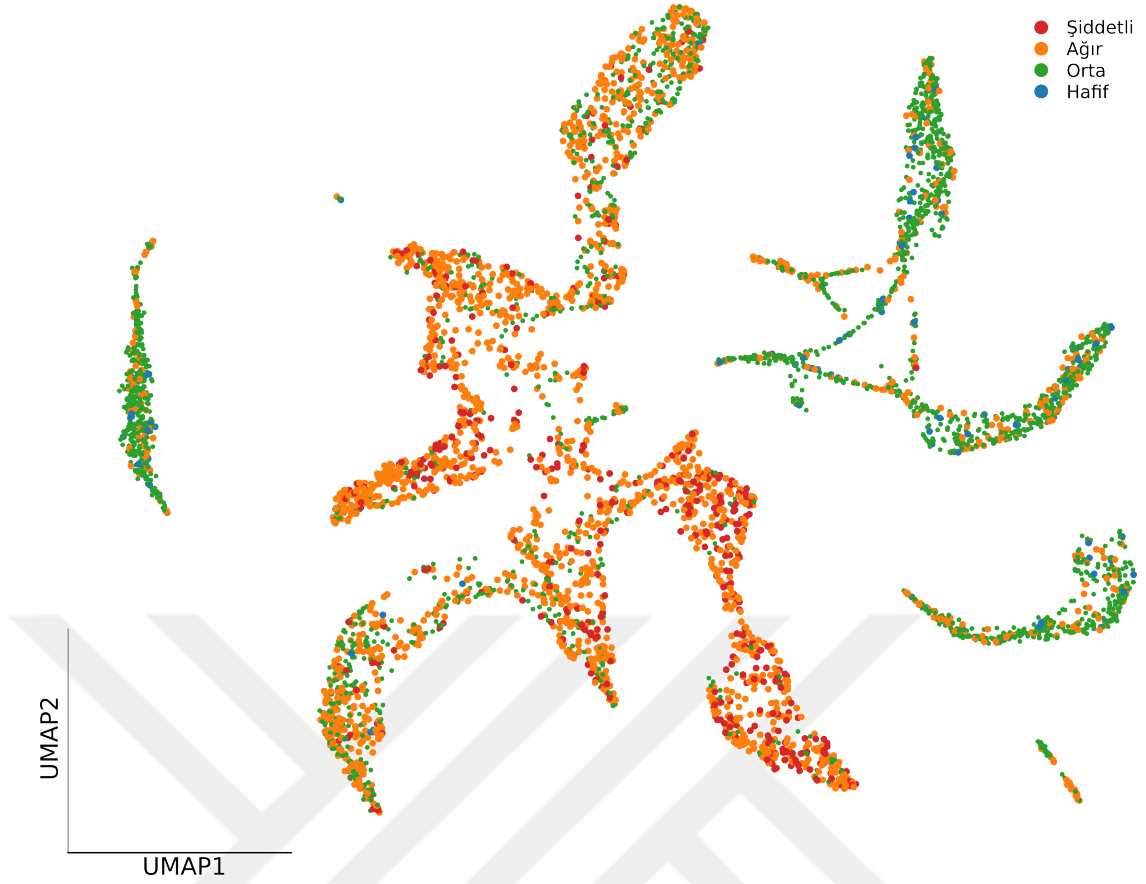
ağır ve şiddetli hücreler birbirleri ile hafif ve orta seviye aktif hücreler de kendi aralarında kümelenmektedir. Bu kümelenmeleri daha iyi yorumlayabilmek adına aktivite seviyeleri hücre tipleri bakımından analiz edilmiştir.

4.3.3. Viral Aktivite Seviyelerinin Analizi

Aktivite seviyeleri hücrelerin hücre tiplerine göre incelenerek hangi hücre tiplerinin HIV enfeksiyonu patogenizinde daha aktif olduğu tespit edilmiştir. Her bir tipteki hücreler dört farklı aktivite seviyesinin her biri için sayılarak hücre tipi aktivite seviyelerinin pasta grafikleri oluşturulmuştur (Şekil 4.4).

Elde edilen sonuçlar HIV patogeneğinde beklenen protein değişimleri açısından doğal öldürücü (NK) tipteki hücrelerin en aktif hücre tipi olduğunu göstermiştir. Elde edilen sonuçlara göre bu hücrelerin %87,4'ü ağır ve şiddetli aktiflik seviyesine sahip olduğunu göstermiştir. Bu durum doğal öldürücü hücrelerin HIV virüsü ile enfekte olmuş hücreleri tanıma ve yok etmede esas rolü üstlenmelerine dayandırılmıştır [138].

CD4 T hücreleri en aktif ikinci hücre tipi olarak gözlemlenmiştir. Bu hücrelerin %81,6'sı ağır ve şiddetli aktiviteye sahiptir. CD4 T hücreleri HIV virüsünün bulaşışında ve virüsün vücutta yayılmasında temel hedef hücre tipi olarak bilinmektedir [139]. Bu durum CD4



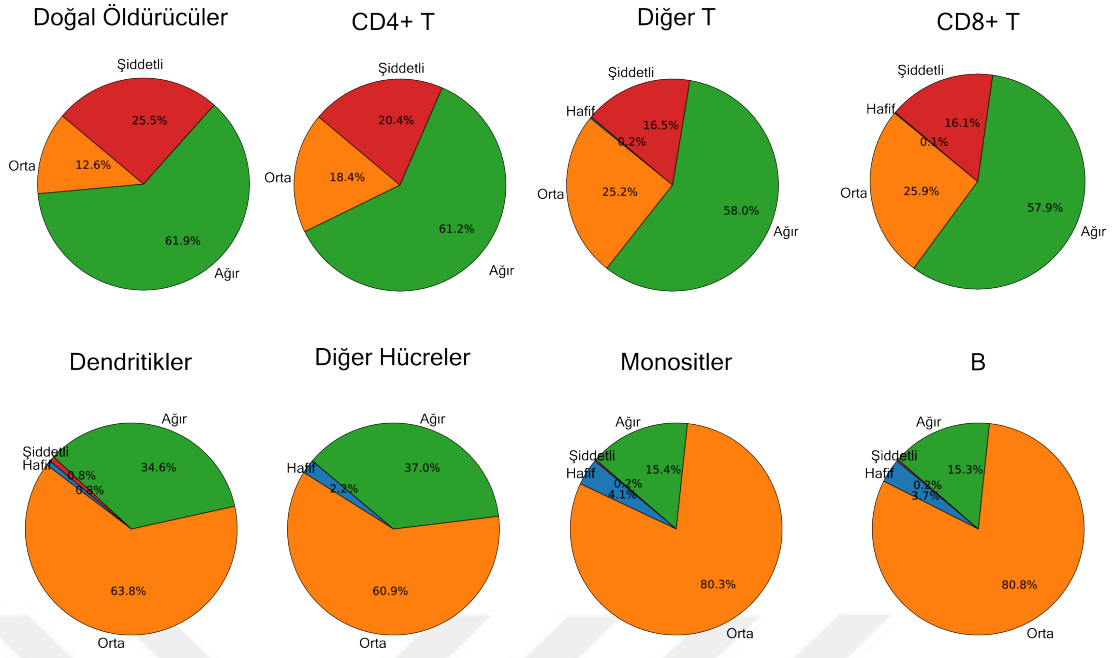
Şekil 4.3: Bütünleştirilmiş verinin aktivite sınıfı ile renklendirilmiş UMAP gösterimi.

T hücrelerinin en aktif ikinci sınıf olmasının temel sebebi olarak kabul edilmiştir.

CD8 T ve diğer T hücreleri benzer karakteristik göstererek %74 ağır / şiddetli hücreye sahip oldukları belirlenmiştir. Bu sitotoksik hücreler doğal öldürücü hücreler ile beraber enfeksiyonun kontrol altında tutulmasında önemli rol oynamaktadır [135]. Doğal öldürücülerle benzer olarak enfeksiyonun kontrol altında tutulmasında görev aldıkları için viral patogeneizde aktif oldukları düşünülmektedir.

Dendritik, monosit, B ve diğer tipteki hücreler HIV enfeksiyonunda beklenen değişimleri karşılama bakımından oldukça pasif karakteristik gösterdikleri gözlemlenmiştir. Monosit hücreleri CD4 T hücreleri gibi HIV enfeksiyonun erken dönem hedefleri olarak bildirilmiştir [140]. Öte yandan hücre altkümeleri deneyleri sırasında gözlemlediğimiz üzere CD4 T hücrelerinin erken dönem (naif) örneklerinde de aktiflik seviyesi düşük olarak tespit edilmiştir.

B hücrelerinin dolaylı yoldan HIV enfeksiyonunda görev aldığı literatür çalışmalarında bildirilmiştir [141]. Yeni ortaya atılan laboratuvar ortamında gözlemlenmiş zayıf



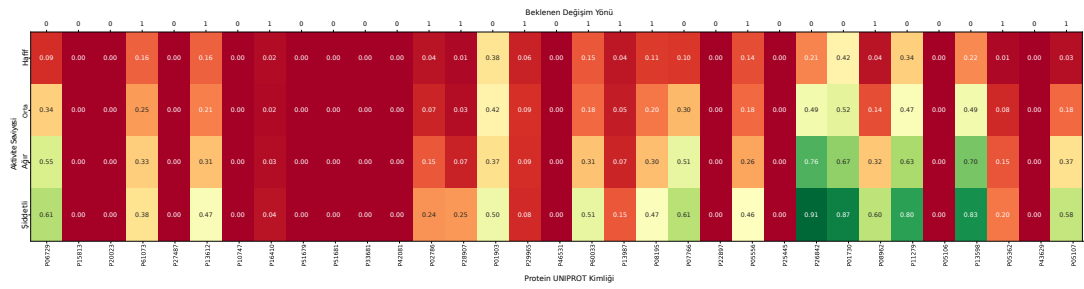
Şekil 4.4: Hücre tiplerine göre aktivite sınıfı dağılımlarının pasta grafik olarak gösterimi.

etkileşimlere rağmen B hücrelerinin doğrudan HIV virüsü ile etkileşimi olmayışı bu hücrelerin düşük aktivite seviyesine sahip olmasını açıklamaktadır. Dendritik hücrelerinin bağışıklık sisteminin HIV enfeksiyonu cevabında aktif rol oynadığı bildirilmiştir [142]. Öte yandan bu hücrelerin neden şüpheye yer bırakmayacak şekilde düşük aktiflik seviyesi gösterdiğine dair bir kanıt bu çalışma kapsamında tespit edilememiştir. Hedef proteinlerin bu hücrelerdeki görevleri ve beklenen değişim yönleri gelecek çalışmaları kapsamında detaylı olarak incelenecektir.

4.3.4. Hedef Proteinler Üzerinde Viral Patogenez Analizi

Hedef proteinlerin hücre tipi, örneklenme zamanı ve viral aktivite yönünden ayrıran hücrelerde gösterdiği farklılıklar analiz edilmiştir. Elde edilen sonuçlar ile virüsün etkileşim kurduğu bu proteinlerin viral patogenezin hangi mekanizmasında görev aldığına dair daha ileri bilgiler elde edilebilmektedir.

Şekil 4.5'te verilen ısı haritası incelendiğinde farklı viral aktivite seviyesine sahip hücre gruplarında hedef proteinlerin beklenen yukarı/aşağı yönlü değişimleri karşılama oranlarının farklılık gösterdiği görülmektedir. Her bir proteinin beklenen değişimi bir A aktivite sınıfında karşılama oranı hesaplanırken; A aktivite sınıfına sahip hücrelerde



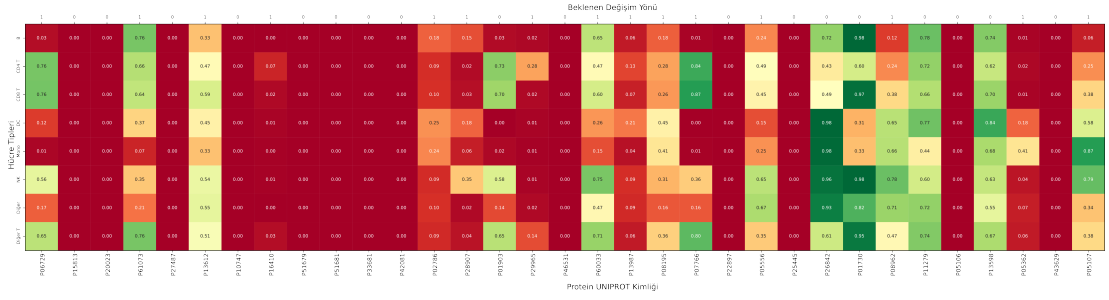
Şekil 4.5: Hedef proteinlerin aktivite sınıflarına göre beklenen ifade seviyesi değişimini karşılama oranlarının ısı haritası.

proteinin beklenen değişimi gösterdiği hücre sayısı / ilgili aktivite sınıfındaki toplam hücre sayısı formülü ile hesaplanmaktadır. Hedef proteinler üzerinde yapılan bu analiz ile sadece ağır veya şiddetli aktiviteye sahip hücrelerde beklenen değişimleri gösteren proteinleri tespit etmek mümkün olmaktadır.

Analiz sonuçları incelendiğinde genel olarak birçok hedef proteinin hafiften şiddetliye doğru beklenen değişimi sağlama oranı kademeli olarak artmaktadır. Bunların yanında P01903 ve P29965 proteinleri tüm aktivite seviyelerinde benzer oranda hücrelerde beklenen değişimi göstermektedir. P01903 proteininin değişim oranları hücre tiplerine göre incelendiğinde de viral aktivitesi yüksek olan tüm hücre tiplerinde benzer karakteristik göstermektedir. Bu durum P01903 proteininin HIV-1 enfeksiyonunun kontrolünü sağlayan hücrelerde ortak olarak bulunması ve potansiyel aşı hedefi olması ile açıklanabilmektedir [143].

P29965 proteini hücre tiplerine göre incelendiğinde sadece CD4 T ve diğer T hücrelerinde beklenen değişimi gösterdiği saptanmıştır. Bu proteinin sitotoksik T hücrelerinin üretimini desteklediği bildirilmiştir [144]. Sitotoksik CD4 T hücrelerinin enfeksiyona karşı aktif rol oynadığı literatürde bildirilmektedir [135]. Tez kapsamında gerçekleştirilen ve ilerleyen bölümlerde detaylı anlatılan altküme analizlerinde de sitotoksik CD4 T hücre altkümesinin viral aktivite seviyesi en yüksek altküme olarak ayrıştığı gözlemlenmiştir.

P28907 ve P08962 proteinleri de diğer proteinlerden farklı karakteristik göstererek şiddetli aktiviteye sahip hücrelerde diğer hücrelerden çok daha yüksek oranda beklenen yukarı yönlü değişimi göstermektedirler. P08962 proteini hücre tipleri bakımından incelendiğinde T ve B hücrelerinde beklenen değişimi düşük oranda gösterirken doğal



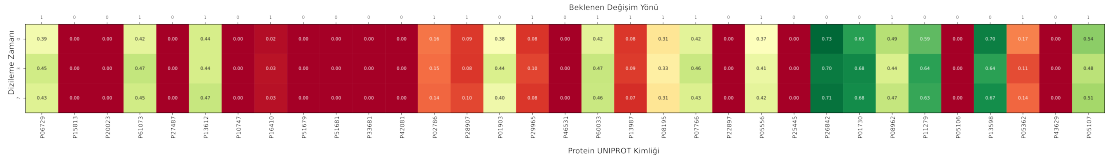
Şekil 4.6: Hedef proteinlerin hücre tiplerine göre beklenen ifade seviyesi değişimini karşılama oranlarının ısı haritası.

öldürücü, monosit, dendritik ve diğer tip hücrelerde yüksek oranda karşılamaktadır. Hücre tiplerinin aktivite seviyeleri göz önüne alındığında şiddetli aktif hücrelerdeki bu artışın doğal öldürücü hücrelerden kaynaklandığını söylemek mümkündür.

P28907 proteini de P08962 proteini ile benzer şekilde şiddetli aktif iki hücre grubundan biri olan T hücrelerinde düşük diğer şiddetli aktif doğal öldürücü hücrelerde ise yüksek oranda beklenen değişimi göstermiştir. Bununla beraber bu protein sadece doğal öldürücü hücrelerde yüksek oranda yukarı yönlü değişim göstermektedir. Yapılan çalışmalar P28907 proteininin doğal öldürücü hücrelerde yukarı yönlü değişime sahip olmasının CD4 T ve CD8 T hücrelerinin profilasyonunu baskıladığını ortaya koymaktadır [145]. Bu durum proteini potansiyel bir tedavi hedefi yapmaktadır.

İleri analizler sonuçlarında ayrıca P15813, P20023, P10747 gibi 13 proteinin tamamında sıfır ölçüm gözlemlenmiştir. Bu proteinlerin tamamının aşağı yönlü değişim göstermesi beklenmektedir. Bu gözlem öncelikle ilgili proteinler için literatürden derlenen değişim yönü beklentisinin yanlış olabileceğini düşündürmektedir. İlgili proteinlere ait ölçümler incelendiğinde toplam hücre sayısının %10'dan daha az hücrede ölçümlenebildikleri gözlemlenmiştir. Bu durum ilgili proteinlerin aşağı yönlü değişim gösterdiğini sağlıklı bir şekilde tespit etmeyi imkânsız hale getirmektedir.

Hedef proteinlerin beklenen değişimleri gösterme oranları hücre tiplerine göre de analiz edilmiştir (Şekil 4.6). Birçok protein değişken oranlarla tüm hücre tiplerinde beklenen değişimleri göstermekle beraber bazı proteinler sadece viral patogeneze aktif hücre tiplerinde beklenen değişimleri göstermektedir. Daha önceden bahsedilen P01903 proteini sadece aktif hücre tiplerinde aşağı yönlü değişim göstermekle beraber, bu hücrelerdeki tüm aktivite seviyeleri için benzer oranlar sergilemektedir.



Şekil 4.7: Hedef proteinlerin örnek dizileme zamanına göre beklenen ifade seviyesi değişimini karşılama oranlarının ısı haritası.

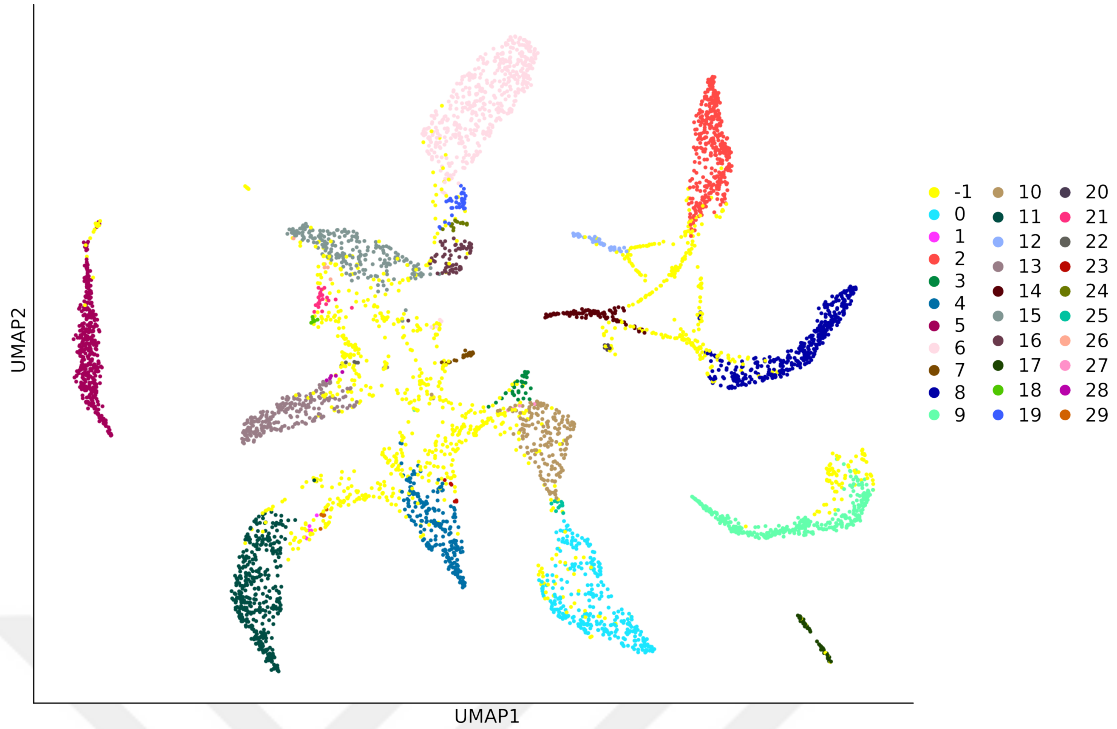
P07766 proteini tüm T hücre tiplerinde yüksek oranda beklenen yukarı yönlü değişimi göstermektedir. Tüm T hücrelerinde sentezlenmesinin yanı sıra bu proteini sentezleyen T hücrelerinin azalması enfeksiyonun ilerlediğinin ve muhtemel ölümle sonuçlanabileceğinin önemli bir göstergesi olabileceği bildirilmiştir [146].

P16410 proteini sadece çok az miktardaki CD4 T hücrelerinde beklenen yukarı yönlü değişimi göstermiştir. Bu protein yüksek virüs yüküne sahip CD4 T hücrelerinde sentezlenerek bu hücrelerin aktivitelerini baskılamaktadır [147]. Bu sebeple P16410 proteini sentezleyen hücrelerin ileri analizi ile HIV-1 patogenezinin açıklanmasına potansiyel katkı sağlanabileceği düşünülmektedir ve gelecek çalışmalar arasındadır.

Hedef proteinler üzerinde aşı sonrası değişimleri gözlemlemek adına hücrelerin örneklenme zamanlarına göre de beklenen değişimleri gösterme oranları incelenmiştir (Şekil 4.7). 12 Protein aşılama sonrası üçüncü günde beklenen değişimleri daha yüksek oranda göstermiştir. Bu proteinlerin 5'i yedinci günde aşılama öncesi seviyeleri geri dönerken 7'si mevcut artışı korumuştur.

7 proteinin beklenen değişimleri gösterme oranı üçüncü günde azalmıştır. Bu proteinlerden P02786 hariç tamamı yedinci günde tekrar artış eğilimi göstermiştir. Proteinin aşılama sonrası neden daha az yukarı yönlü değişim gösterdiğine dair bir delil bulunamamıştır ve gelecek çalışmalar arasında yer almaktadır.

P13612 proteini üçüncü günde değişim gösterme oranında bir değişiklik olmamasına rağmen yedince gün örneklerinde daha fazla sayıda hücrede yukarı yönlü değişim gösterdiği tespit edilmiştir. Bu proteinin HIV-1 enfeksiyonundan korunma ve tedavide önemli olduğu literatür çalışmalarında bildirilmektedir [148]. Protein üzerindeki bu değişimin aşıya olumlu cevap veren hastalardan kaynaklanması muhtemeldir. Bu hücrelerin incelenmesi ile HIV-1 patogenezinin açıklanmasına potansiyel katkı sağlanabileceği düşünülmektedir ve gelecek çalışmalar arasında yer almaktadır.



Şekil 4.8: Bütünleşik verinin gözetimsiz kümeleme etiketleriyle renklendirilmiş UMAP gösterimi.

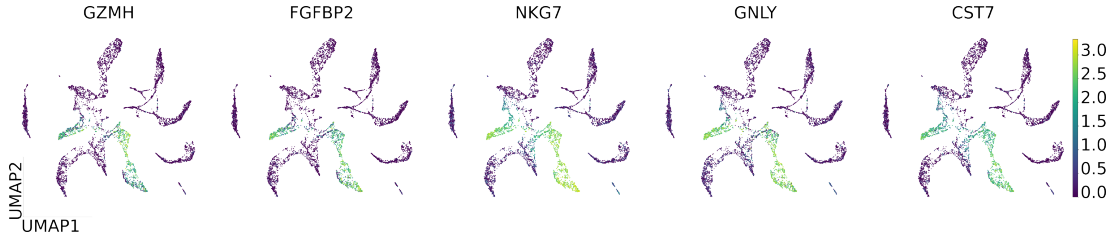
4.3.5. Viral Patogenezde Aktif Hücre Altkümelerinin İncelenmesi

Test veri kümesi SCPRO-VI algoritması kullanılarak bütünleştirilmiştir. Bütünleşik veri DBSCAN kümeleme yöntemi kullanılarak kümelendirilmiştir. Epsilon değeri 0.003 olarak belirlenmiştir. Kümeleme sonucunda 29 hücre kümesi elde edilmiştir (Şekil 4.8).

Elde edilen hücre alt kümelerin aktivite seviyeleri hesaplandığında sırasıyla 10, 3, 0 ve 13 kümelerinin 3 ve üzeri aktivite seviyesine sahip olduğu tespit edilmiştir. 0 altkümesinin tamamı doğal öldürücü hücrelerden oluşmaktadır. Bu hücrelerin yüksek aktivite gösterdiği önceki analizler ile sabittir.

3 kümesi çoğunluğu CD8 T hücrelerinden oluşan ve toplamda 36 hücre içeren küçük bir hücre kümesidir. Bu hücrelerin 20 tanesi efektör hafıza CD8 T hücrelerinden oluşmaktadır. Bu CD8 T hücre alt kümesinin viral rezervuarları temizleyebildiği ve virüsle aktif savaştığı bildirilmiştir [149, 150].

13 kümesi %17'si şiddetli aktif olmak üzere 235 hücre barındırmaktadır. Bu hücre kümesinin %32'si CD8 T hücrelerinden oluşurken geri kalanı doğal öldürücü hücrelerdir. CD8 T hücreleri 3 kümesinde olduğu gibi efektör hafıza hücreleridir. Her iki hücre



Şekil 4.9: Bütünleşik verinin sitotoksik CD4 T hücre altkümesini diğer CD4 T hücrelerinden ayırmada en etkili 5 genin ifade seviyeleri ile renklendirilmiş UMAP gösterimleri.

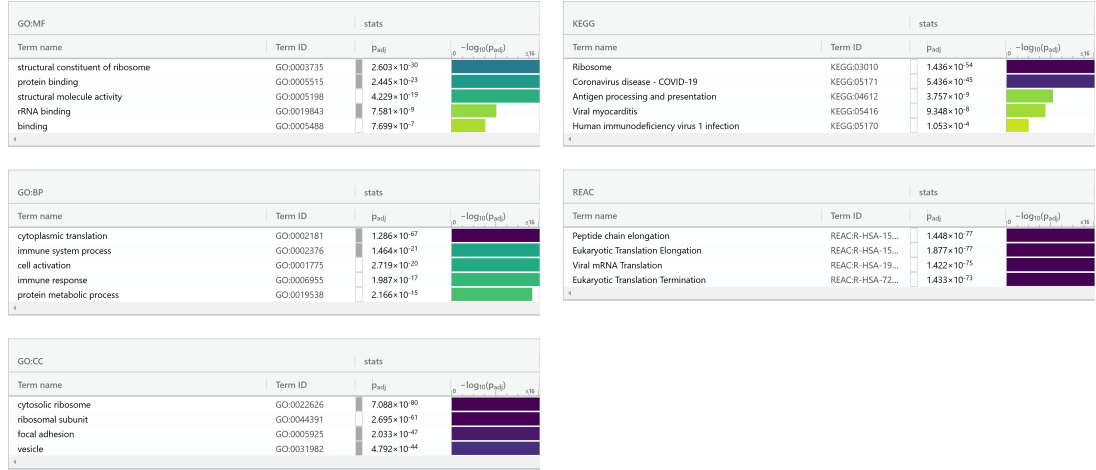
tipinin de HIV-1 enfeksiyonu ile doğrudan savaştığı önceki analizlerde bildirilmiştir.

10 hücre kümesinin %42'si şiddetli aktif hücrelerden, kalan %58'i ise ağır aktif hücrelerden oluşmaktadır. İçerdiği 175 hücrenin 32'si sitotoksik CD4 T hücre tipindedir ve kalan hücreler efektör hafıza CD8 T hücreleridir. Değerildiği üzere, sitotoksik CD4 T hücrelerinin HIV-1 virüs hedefi CD4 T hücrelerinden farklı olarak enfeksiyona karşı diğer bağışıklık hücreleri ile işbirliği kurarak savaştığı bildirilmiştir [135].

4.3.6. Viral Patogeneizde Aktif Altkümelerde Biyobelirteç Analizi

Altküme analizinde tespit edilen sitotoksik CD4 T altkümesinde (alküme 10) biyobelirteç analizi gerçekleştirilmiştir. Bu altkümenin analiz için seçilmesindeki amaç HIV-1 virüsünün enfeksiyonda temel olarak hedef aldığı CD4 T hücrelerinin hangi genler ile ayrışarak enfeksiyonla savaşan hücrelere dönüştüğünün tespit edilmesidir.

Analiz öncelikle sadece CD4 T hücreleri arasında gerçekleştirilerek hedef hücre altkümesini diğer CD4 T hücrelerinden ayırtıran genler tespit edilmiştir (Şekil 4.9). Filtrelemede kullanılan p değeri için varsayılan değer olarak 5×10^{-5} kabul edilmiştir. Bu değer düşürülerek daha az sayıda ve yüksek ayırt ediciliğe sahip biyobelirteçler elde edilebilmektedir. Benzer şekilde p eşik değeri arttırıldığında daha fazla sayıda ayırt edici Biyobelirteç tespit etmek mümkündür. Analiz sonucunda 470 gen eşik p değerinin üzerinde kalarak ayırtıcı olarak tespit edilmiştir. Bunların arasından en ayırtıcı 20 gen Çizelge A.4'de verilmiştir. Şekil 4.9'da sunulan en ayırtıcı 5 gene ait ifade seviyeleri ile renklendirilen bütünleşik veriye ait UMAP gösterimleri, bu genlerin enfeksiyonla doğrudan savaşan hücrelerin tamamında aktif olduğunu göstermektedir. Öte yandan sadece diğer CD4 T hücreleri değil, kalan tüm hücrelerde bu genlerin

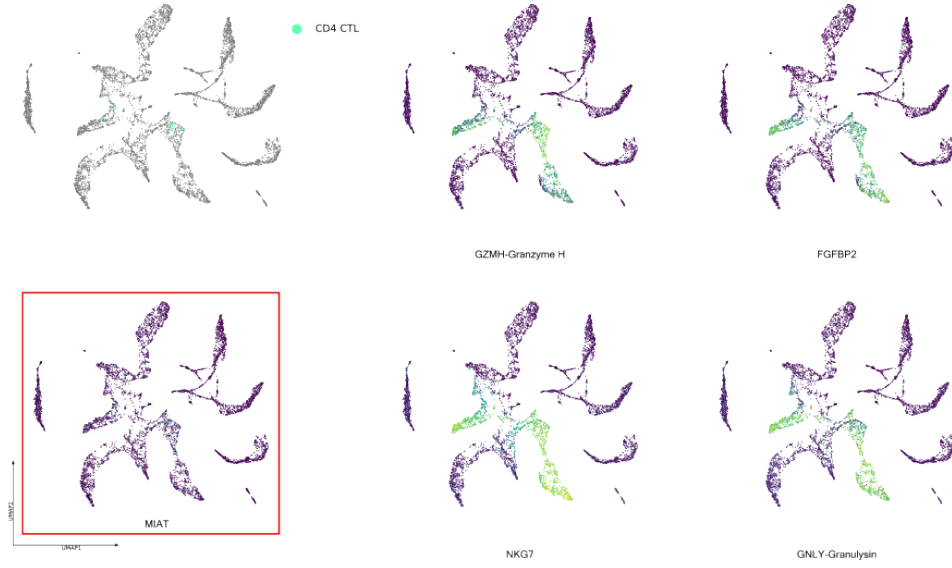


Şekil 4.10: CD4 T hücre altkümesini diğer CD4 T hücrelerinden ayırmada en etkili genler ile gen ontoloji zenginleştirme analiz sonuçları.

düşük ifade seviyelerine sahip olduğu gözlemlenmiştir. Bu durum ilgili genlerin HIV-1 enfeksiyonu ile savaşmada aktif rol oynadığını göstermektedir [151].

Tespit edilen 100 en ayırıştırıcı gen üzerinde gen ontoloji zenginleştirme analizi gerçekleştirilmiştir (Şekil 4.10). Hücresel bileşen analizi genlerin yoğun olarak hücre içi sınıfta –sitoplazmada- bulunduğu ve görev aldığını göstermiştir. Genlerin biyolojik fonksiyon zenginleştirmeleri incelendiğinde bağışıklık sistemi ve hücre aktivasyonunda sitoplazmada bulunan yollarda görev alarak sitoplazmada gerçekleşen RNA translasyonlarında görev almaktadırlar. Bu genlerin moleküler fonksiyon zenginleştirmeleri de destekleyici biçimde protein ve rRNA bağlanmalarında aktif rol oynayarak protein sentezine katkı sağladıklarını göstermektedir. REAC ve KEGG veri tabanlarından elde edilen yolak bilgileri de yine bu genlerin viral patogeneze etkili proteinlerin translasyonlarında görev aldıklarını göstermektedir.

Ayırıştırıcı gen analizi ilgili hücre altkümesinde bulunan hücreleri kalan tüm hücrelerden ayıran biyobelirteçleri tespit etmek üzere tekrarlanmıştır. Analiz sonucunda 227 ayırıştırıcı gen tespit edilmiştir ve bunların ilk 20'si Çizelge A.4'de verilmiştir. En ayırıştırıcı 5 gen incelendiğinde (Şekil 4.11) CD4 T hücreleri arasında yapılan analizde de ayırıştırıcı olan GZMH ve NKG7 genleri bu hücreleri kalan tüm hücrelerden ayırmada da önemli olarak tespit edilmiştir. Dahası önceki analizde dokuzuncu en ayırıştırıcı gen olan CCL5 geni de tüm sitotoksik hücrelerde yüksek ifade seviyesine sahip olmasının yanında bu alt küme için dördüncü en ayırıştırıcı gen olarak belirlenmiştir.



Şekil 4.11: Bütünleşik verinin sitotoksik CD4 T hücre altkumesini diğer tüm hücrelerden ayırmada en etkili 5 genin ifade seviyeleri ile renklendirilmiş UMAP gösterimleri.

Ayrıştırıcı MIAT ve TRGC2 genleri diğer ayrıştırıcı üç genden farklı olarak sadece sitotoksik CD4 T hücrelerinde ifade edilmektedir. Literatürde MIAT geni ile HIV-1 enfeksiyonu arasında ilişki kuran bir çalışmaya rastlanmamıştır. Bununla beraber bu genin sitotoksik T hücrelerinde yüksek ifade seviyesine sahip olduğunu tespit eden güncel bir çalışma bulunmaktadır [152]. Benzer şekilde TRGC2 geni de sitotoksik T hücrelerinde yüksek ifade seviyeli olarak bildirilmiştir [153].

4.4. Sonuç

Bu bölümde yapılan çalışmalar, viral patogeneze dinamiklerini anlamada tek-hücre seviyeli çoklu-omik veri analizinin potansiyelini ortaya koymaktadır. Geliştirilen viral aktivite seviyelerini belirleme yöntemi, enfeksiyon sonrası hücrelerdeki proteinlerde beklenen yukarı/aşağı yönlü değişimleri temel alarak, enfeksiyona karşı yanıt veren hücre popülasyonlarının doğru bir şekilde sınıflandırılmasını sağlamaktadır. Elde edilen hücre aktivite seviyeleri, viral patogenezin anlaşılması ve enfeksiyon sebebiyle özelleşen hücre altkümelerinin tespiti açısından önemli bir yenilik sunmaktadır.

SCPRO-VI tek-hücre çoklu-omik veri bütünleştirme algoritması, transkriptomik ve proteomik verilerin birleştirilmesini sağlayarak viral patogeneze özgü hücresel

altkümeleri ortaya çıkarmada etkin bir yaklaşım sunmaktadır. Örneğin, sitotoksik CD4 T hücreleri gibi enfeksiyonla doğrudan ilişkili altkümelerin belirlenmesi, enfeksiyon sürecinde hücreler arasındaki artan heterojen yapının daha iyi anlaşılmasını sağlamaktadır. Bu altkümelerdeki genler ve proteinler üzerinde gerçekleştirilen biyobelirteç analizleri, enfeksiyonun gelişimi ve bağışıklık tepkisinde rol oynayan ayırt edici özniteliklerin tespit edilmesini mümkün kılmıştır.

Elde edilen bulgular, viral patogeneze ilişkili genlerin ayırt edici özelliklerini ortaya koymuştur. Örneğin, GZMH, NKG7 ve CCL5 gibi genlerin sitotoksik hücrelerde yüksek ifade seviyelerine sahip olması, bu genlerin enfeksiyonla savaşta oynadığı önemli rolü desteklemektedir. Ayrıca, sitotoksik CD4 T hücrelerinde öne çıkan genlerin enfeksiyon sürecindeki bağışıklık işlevlerini aydınlatmada kritik bilgiler sunmaktadır.

Çalışma sonucunda elde edilen bulgular, gelecekteki araştırmalar için heyecan verici yeni yollar sunmaktadır. Öncelikle, hedef proteinlerin viral aktivite özellikleri daha ayrıntılı bir şekilde incelenecek ve patogeneze rollerinin daha iyi anlaşılması sağlanacaktır. Ayrıca, viral aktivite düzeylerini belirleme algoritmasının hücre tipine özgü düzenleme değişimlerini destekleyecek şekilde geliştirilmesi öncelikli hedeflerimizdendir. Böylece enfeksiyon dinamiklerinin daha ayrıntılı bir şekilde anlaşılması mümkün olacaktır.

Tespit edilen hücre altkümeleri ve biyobelirteçler ileri analizlere tabi tutularak ve viral patogeneze özgü katkıları bağışıklık tepkilerindeki rolleri bakımından değerlendirilecektir. Son olarak, uçtan uca biyobelirteç tespit yöntemi, farklı veri setlerine uygulanarak yöntemin dayanıklılığı ve yeniden uyarlanabilirliği değerlendirilecektir. Bu hedefler, metodolojilerimizi geliştirmeyi ve viral patogeneze alanında ilerlemelere katkıda bulunmayı amaçlamaktadır.

5. SONUÇ

Yatay ve dikey veri bütünleştirme süreçlerindeki temel zorlukları ele alan SCPRO-HI ve SCPRO-VI algoritmaları, yüksek boyutlu tek-hücre verilerinin anlamlı biyolojik varyasyonları koruyarak bütünleştirilmesini sağlayan sağlam bir çerçeve sunmaktadır. Bu tezde sunulan sonuçlar, önerilen tek-hücre çoklu-omik veri bütünleştirme yöntemlerinin veri kalitesini artırmada, biyolojik anlamı korumada ve ileri analiz kapasitesini geliştirmede etkili olduğunu göstererek bu savı desteklemektedir.

Yatay bütünleştirme yöntemi SCPRO-HI, tek-hücre proteomik verilerindeki toplu etki gürültüsünü etkili bir şekilde ortadan kaldırarak farklı deneylerden elde edilen veri kümelerinin biyolojik çeşitliliğini kaybetmeden uyumlu hale getirilmesini sağlamaktadır. Sonuçlar, SCPRO-HI'nin siluet skorlarını önemli ölçüde iyileştirdiğini ve böylece daha saf ve ayırık hücre kümeleri elde edildiğini göstermektedir. Ayrıca, alternatif yöntemlere kıyasla yüksek değişken proteinleri daha iyi koruyarak biyolojik anlamı daha güvenilir bir şekilde yansıtmaktadır. Bu iyileştirmeler, farklı örnekler arasında daha güvenilir karşılaştırmalı analizlerin gerçekleştirilmesine olanak tanıyarak teknik etkilerin gölgeleyebileceği biyolojik örüntülerin ortaya çıkarılmasını sağlamaktadır.

Dikey bütünleştirme yöntemi SCPRO-VI, tek-hücre seviyesinde transkriptomik ve proteomik ölçümleri bütünleştirerek çoklu-omik verileri etkin bir şekilde kullanmayı sağlamaktadır. Gerçekleştirilen deneylerde bütünleşik verinin ham veriye göre daha yüksek ARI skoru ve en yakın komşu eşleştirme performansına sahip olması, yöntemin hücresel heterojenliği belirginleştirdiğini göstermektedir. SCPRO-VI, çoklu-omik özelliklerin biyolojik açıdan anlamlı bir şekilde bütünleştirilmesini güvence altına alarak hücre yaşam döngülerinin ve etkileşimlerinin daha doğru bir şekilde karakterize edilmesini sağlamaktadır. Çoklu-omik gömmelerin analizi, birden fazla moleküler katmanın eşzamanlı olarak değerlendirilmesinin avantajlarını vurgulayarak hücresel kimliğin ve işlevlerin daha kapsamlı bir şekilde anlaşılmasını mümkün kılmaktadır.

Bütünleştirilmiş veri kümelerinin biyolojik geçerliliğini göstermek amacıyla, işlenmiş tek-hücre verisi kullanılarak viral patogenezi analizi gerçekleştirilmiştir. Çalışma, yüksek viral aktivite seviyelerine sahip farklı hücre alt popülasyonlarını belirleyerek viral enfeksiyon dinamikleri ile ilişkili önemli biyobelirteçleri ortaya çıkarmıştır. Sonuçlar,

proteomik ve transkriptomik verilerin birlikte değerlendirilmesinin enfeksiyonla ilişkili hücresel durumları tespit etme kapasitesini artırdığını göstermekte ve viral patogeneзде rol oynayan hem bilinen hem de aday biyobelirteçlerin keşfini sağlamaktadır. Yüksek aktivite gösteren hücre alt kümelerinin tanımlanması, konak-patojen etkileşimlerine ve potansiyel terapötik hedeflerin belirlenmesine dair değerli iç görüler sunmaktadır.

Bu tez, tek-hücre veri bütünleştirme için yeni hesaplamalı yaklaşımlar sunarak biyoinformatik alanına katkıda bulunmaktadır. Mevcut yöntemlerin sınırlamalarını ele almakta ve önerilen yöntemlerin gerçek biyolojik problemlere uygulanabilirliğini ortaya koymaktadır. Önerilen metodolojiler hem hesaplamalı olarak verimli hem de biyolojik açıdan anlamlıdır, bu da onları tek-hücre omik verileri ile çalışan araştırmacılar için değerli araçlar haline getirmektedir. Heterojen veri kümelerinin daha iyi bütünleştirilmesiyle geliştirilen algoritmalar, daha doğru hastalık modelleme, biyobelirteç keşfi ve kişiselleştirilmiş tıp uygulamalarına zemin hazırlamaktadır. Bulgular ayrıca çoklu-omik bütünleştirmenin, tekil modalite analizleriyle tam olarak açıklanamayan karmaşık biyolojik süreçlerin aydınlatılmasındaki önemini vurgulamaktadır. Gelecekteki çalışmalar, bu yöntemleri daha da geliştirerek bütünleştirme stratejilerini rafine edebilir, ek veri modalitelerini araştırabilir ve farklı hastalık modelleri ve deneysel koşullar üzerinden bütünleştirilmiş veri kümelerinin biyolojik etkilerini daha geniş kapsamlı olarak doğrulayabilir.

Geliştirilen bütünleştirme algoritmalarının kullanımı biyolojik veri ile sınırlı değildir. SCPRO-HI temel olarak yatay bütünleştirme prensiplerine dayanmaktadır ve hava durumu veya sensör ağları gibi aynı türdeki farklı kaynaklardan elde edilen verilerin gürültüden arındırılması için de kullanılabilir. Öte yandan SCPRO-VI yöntemi temelde farklı modaliteler arasında ortak bir temsil öğrenerek, bilgi kaybını minimize ederken farklı veri kaynaklarının uzlaştırılmasını (ing. data reconciliation) sağlamayı amaçlamaktadır. Bu yöntem çoklu-omik veri bütünleştirmenin ötesinde herhangi bir çoklu-modalite veri setinin bütünleştirilmesi için genel bir çözüm sunmaktadır.

SCPRO-VI çoklu veri ile desteklenen sistemlerde modaliteler arası bilgi transferini optimize etmek adına avantaj sağlama potansiyeline sahiptir. Ekonomik ve sosyal veriler arasındaki ilişkilerin analiz edilmesi bu kapsamda değerlendirilebilir. Örneğin, ekonomik göstergeler (enflasyon, faiz oranları, döviz kurları) ile sosyal faktörler (eğitim düzeyi, yaş grupları, politik eğilimler, savaşlar) arasındaki ilişkilerin belirlenmesi

iin dikey btnleřtirme gereklidir. Benzer řekilde, lkelerin ekonomik durumlarını anlamak iin yalnızca finansal gstergelerin yanında siyasal ve sosyal deėiřkenlerin de dikkate alınması nemlidir. SCPRO-VI algoritması, farklı modalitelerden gelen bilgiler btnleřtirilerek lkelerin ekonomik durumunu kapsamlı biimde analiz etme imkanı saėlamaktadır. SCPRO-HI ve SCPRO-VI yntemlerinin genelleřebilir yapısı sayesinde hem akademik arařtırmalar hem de gerek dnya problemlerinde ok ynl analizler ile karar verme srelerine katkı saėlama potansiyeli sunulmaktadır.



KAYNAKLAR

- [1] Sanger F., Nicklen S., Coulson A. R., (1977), "DNA sequencing with chain-terminating inhibitors", *Proceedings of the National Academy of Sciences of the United States of America*, 74(12), 5463–5467.
- [2] Bentley D. R., Balasubramanian S., Swerdlow H. P., Smith G. P., Milton J., Brown C. G., Hall K. P., Evers D. J., Barnes C. L., Bignell H. R., Boutell J. M., Bryant J., Carter R. J., Keira Cheetham R., Cox A. J., Ellis D. J., Flatbush M. R., Gormley N. A., Humphray S. J., Irving L. J., Karbelashvili M. S., Kirk S. M., Li H., Liu X., Maisinger K. S., Murray L. J., Obradovic B., Ost T., Parkinson M. L., Pratt M. R., Rasolonjatovo I. M. J., Reed M. T., Rigatti R., Rodighiero C., Ross M. T., Sabot A., Sankar S. V., Scally A., Schroth G. P., Smith M. E., Smith V. P., Spiridou A., Torrance P. E., Tzonev S. S., Vermaas E. H., Walter K., Wu X., Zhang L., Alam M. D., Anastasi C., Aniebo I. C., Bailey D. M. D., Bancarz I. R., Banerjee S., Barbour S. G., Baybayan P. A., Benoit V. A., Benson K. F., Bevis C., Black P. J., Boodhun A., Brennan J. S., Bridgham J. A., Brown R. C., Brown A. A., Buermann D. H., Bundu A. A., Burrows J. C., Carter N. P., Castillo N., Chiara E. Catenazzi M., Chang S., Neil Cooley R., Crake N. R., Dada O. O., Diakoumakos K. D., Dominguez-Fernandez B., Earnshaw D. J., Egbujor U. C., Elmore D. W., Etchin S. S., Ewan M. R., Fedurco M., Fraser L. J., Fuentes Fajardo K. V., Scott Furey W., George D., Gietzen K. J., Goddard C. P., Golda G. S., Granieri P. A., Green D. E., Gustafson D. L., Hansen N. F., Harnish K., Haudenschield C. D., Heyer N. I., Hims M. M., Ho J. T., Horgan A. M., Hoschler K., Hurwitz S., Ivanov D. V., Johnson M. Q., James T., Huw Jones T. A., Kang G.-D., Kerelska T. H., Kersey A. D., Khrebtukova I., Kindwall A. P., Kingsbury Z., Kokko-Gonzales P. I., Kumar A., Laurent M. A., Lawley C. T., Lee S. E., Lee X., Liao A. K., Loch J. A., Lok M., Luo S., Mammen R. M., Martin J. W., McCauley P. G., McNitt P., Mehta P., Moon K. W., Mullens J. W., Newington T., Ning Z., Ling Ng B., Novo S. M., O'Neill M. J., Osborne M. A., Osnowski A., Ostadan O., Paraschos L. L., Pickering L., Pike A. C., Pike A. C., Chris Pinkard D., Pliskin D. P., Podhasky J., Quijano V. J., Racz C., Rae V. H., Rawlings S. R., Chiva Rodriguez A., Roe P. M., Rogers J., Rogert Bacigalupo M. C., Romanov N., Romieu A., Roth R. K., Rourke N. J., Ruediger S. T., Rusman E., Sanches-Kuiper R. M., Schenker M. R., Seoane J. M., Shaw R. J., Shiver M. K., Short S. W., Sizto N. L., Sluis J. P., Smith M. A., Ernest Sohna Sohna J., Spence E. J., Stevens K., Sutton N., Szajkowski L., Tregidgo C. L., Turcatti G., vandeVondele S., Verhovsky Y., Virk S. M., Wakelin S., Walcott G. C., Wang J., Worsley G. J., Yan J., Yau L., Zuerlein M., Rogers J., Mullikin J. C., Hurles M. E., McCooke N. J., West J. S., Oaks F. L., Lundberg P. L., Klenerman D., Durbin R., Smith A. J., (2008), "Accurate whole human genome sequencing using reversible terminator chemistry", *Nature*, 456(7218), 53–59. Publisher: Nature Publishing Group.
- [3] Nuwaysir E. F., Bittner M., Trent J., Barrett J. C., Afshari C. A., (1999), "Microarrays and toxicology: the advent of toxicogenomics", *Molecular Carcinogenesis*, 24(3), 153–159.
- [4] Bechtel W., (2005), "Discovering Cell Mechanisms: The Creation of Modern Cell

Biology”, Cambridge Studies in Philosophy and Biology, Cambridge University Press, Cambridge.

- [5] Baxevas A. D., Bader G. D., Wishart D. S., (2020), “Bioinformatics”, John Wiley & Sons. Google-Books-ID: PBxZxAEACAAJ.
- [6] Picelli S., Björklund A. K., Faridani O. R., Sagasser S., Winberg G., Sandberg R., (2013), “Smart-seq2 for sensitive full-length transcriptome profiling in single cells”, *Nature Methods*, 10(11), 1096–1098. Publisher: Nature Publishing Group.
- [7] Chen G., Ning B., Shi T., (2019), “Single-Cell RNA-Seq Technologies and Related Computational Data Analysis”, *Frontiers in Genetics*, 10, 317.
- [8] Jackson C. A., Vogel C., (2022), “New horizons in the stormy sea of multimodal single-cell data integration”, *Molecular Cell*, 82(2), 248–259.
- [9] Cochain C., Vafadarnejad E., Arampatzi P., Pelisek J., Winkels H., Ley K., Wolf D., Saliba A. E., Zernecke A., (2018), “Single-cell RNA-seq reveals the transcriptional landscape and heterogeneity of aortic macrophages in murine atherosclerosis”, *Circulation research*, 122(12), 1661–1674.
- [10] Raser J. M., O’Shea E. K., (2005), “Noise in gene expression: origins, consequences, and control”, *Science (New York, N.Y.)*, 309(5743), 2010–2013.
- [11] Adil A., Kumar V., Jan A. T., Asger M., (2021), “Single-Cell Transcriptomics: Current Methods and Challenges in Data Acquisition and Analysis”, *Frontiers in Neuroscience*, 15, 591122.
- [12] Cusanovich D. A., Daza R., Adey A., Pliner H. A., Christiansen L., Gunderson K. L., Steemers F. J., Trapnell C., Shendure J., (2015), “Multiplex single cell profiling of chromatin accessibility by combinatorial cellular indexing”, *Science (New York, N.Y.)*, 348(6237), 910–914.
- [13] Karemaker I. D., Vermeulen M., (2018), “Single-Cell DNA Methylation Profiling: Technologies and Biological Applications”, *Trends in Biotechnology*, 36(9), 952–965.
- [14] Kaya-Okur H. S., Wu S. J., Codomo C. A., Pledger E. S., Bryson T. D., Henikoff J. G., Ahmad K., Henikoff S., (2019), “CUT&Tag for efficient epigenomic profiling of small samples and single cells”, *Nature Communications*, 10(1), 1930. Publisher: Nature Publishing Group.
- [15] Ramani V., Deng X., Qiu R., Gunderson K. L., Steemers F. J., Distech C. M., Noble W. S., Duan Z., Shendure J., (2017), “Massively multiplex single-cell Hi-C”, *Nature Methods*, 14(3), 263–266. Publisher: Nature Publishing Group.
- [16] Xie H., Ding X., (2022), “The Intriguing Landscape of Single-Cell Protein Analysis”, *Advanced Science (Weinheim, Baden-Wurttemberg, Germany)*, 9(12), e2105932.
- [17] Stoeckius M., Hafemeister C., Stephenson W., Houck-Loomis B., Chattopadhyay P. K., Swerdlow H., Satija R., Smibert P., (2017), “Simultaneous epitope and transcriptome measurement in single cells”, *Nature Methods*, 14(9), 865–868. Publisher: Nature Publishing Group.

- [18] Chung H., Parkhurst C. N., Magee E. M., Phillips D., Habibi E., Chen F., Yeung B. Z., Waldman J., Artis D., Regev A., (2021), “Joint single-cell measurements of nuclear proteins and RNA in vivo”, *Nature Methods*, 18(10), 1204–1212. Publisher: Nature Publishing Group.
- [19] Slavov N., (2021), “Single-cell protein analysis by mass spectrometry”, *Current Opinion in Chemical Biology*, 60, 1–9.
- [20] Petelski A. A., Emmott E., Leduc A., Huffman R. G., Specht H., Perlman D. H., Slavov N., (2021), “Multiplexed single-cell proteomics using SCoPE2”, *Nature Protocols*, 16(12), 5398–5425. Publisher: Nature Publishing Group.
- [21] Cheung T. K., Lee C.-Y., Bayer F. P., McCoy A., Kuster B., Rose C. M., (2021), “Defining the carrier proteome limit for single-cell proteomics”, *Nature Methods*, 18(1), 76–83. Publisher: Nature Publishing Group.
- [22] Bock C., Farlik M., Sheffield N. C., (2016), “Multi-Omics of Single Cells: Strategies and Applications”, *Trends in Biotechnology*, 34(8), 605–608.
- [23] Ma A., McDermaid A., Xu J., Chang Y., Ma Q., (2020), “Integrative Methods and Practical Challenges for Single-Cell Multi-omics”, *Trends in Biotechnology*, 38(9), 1007–1022.
- [24] Hu Y., An Q., Guo Y., Zhong J., Fan S., Rao P., Liu X., Liu Y., Fan G., (2019), “Simultaneous Profiling of mRNA Transcriptome and DNA Methylome from a Single Cell”, *Methods in Molecular Biology* (Clifton, N.J.), 1979, 363–377.
- [25] Chen S., Lake B. B., Zhang K., (2019), “High-throughput sequencing of the transcriptome and chromatin accessibility in the same cell”, *Nature Biotechnology*, 37(12), 1452–1457. Publisher: Nature Publishing Group.
- [26] Rodriguez-Meira A., O’Sullivan J., Rahman H., Mead A. J., (2020), “TARGET-Seq: A Protocol for High-Sensitivity Single-Cell Mutational Analysis and Parallel RNA Sequencing”, *STAR protocols*, 1(3), 100125.
- [27] Mimitou E. P., Cheng A., Montalbano A., Hao S., Stoeckius M., Legut M., Roush T., Herrera A., Papalexi E., Ouyang Z., Satija R., Sanjana N. E., Koralov S. B., Smibert P., (2019), “Multiplexed detection of proteins, transcriptomes, clonotypes and CRISPR perturbations in single cells”, *Nature Methods*, 16(5), 409–412. Publisher: Nature Publishing Group.
- [28] Peterson V. M., Zhang K. X., Kumar N., Wong J., Li L., Wilson D. C., Moore R., McClanahan T. K., Sadekova S., Klappenbach J. A., (2017), “Multiplexed quantification of proteins and transcripts in single cells”, *Nature Biotechnology*, 35(10), 936–939. Publisher: Nature Publishing Group.
- [29] Xia C., Fan J., Emanuel G., Hao J., Zhuang X., (2019), “Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression”, *Proceedings of the National Academy of Sciences of the United States of America*, 116(39), 19490–19499.
- [30] Lee J., Hyeon D. Y., Hwang D., (2020), “Single-cell multiomics: technologies and data analysis methods”, *Experimental & Molecular Medicine*, 52(9), 1428–1442. Publisher: Nature Publishing Group.

- [31] Dixit A., Parnas O., Li B., Chen J., Fulco C. P., Jerby-Arnon L., Marjanovic N. D., Dionne D., Burks T., Raychowdhury R., Adamson B., Norman T. M., Lander E. S., Weissman J. S., Friedman N., Regev A., (2016), “Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens”, *Cell*, 167(7), 1853–1866.e17.
- [32] Zheng G. X. Y., Terry J. M., Belgrader P., Ryvkin P., Bent Z. W., Wilson R., Ziraldo S. B., Wheeler T. D., McDermott G. P., Zhu J., Gregory M. T., Shuga J., Montesclaros L., Underwood J. G., Masquelier D. A., Nishimura S. Y., Schnall-Levin M., Wyatt P. W., Hindson C. M., Bharadwaj R., Wong A., Ness K. D., Beppu L. W., Deeg H. J., McFarland C., Loeb K. R., Valente W. J., Ericson N. G., Stevens E. A., Radich J. P., Mikkelsen T. S., Hindson B. J., Bielas J. H., (2017), “Massively parallel digital transcriptional profiling of single cells”, *Nature Communications*, 8(1), 14049. Publisher: Nature Publishing Group.
- [33] Saunders A., Macosko E. Z., Wysoker A., Goldman M., Krienen F. M., de Rivera H., Bien E., Baum M., Bortolin L., Wang S., Goeva A., Nemesh J., Kamitaki N., Brumbaugh S., Kulp D., McCarroll S. A., (2018), “Molecular Diversity and Specializations among the Cells of the Adult Mouse Brain”, *Cell*, 174(4), 1015–1030.e16.
- [34] Li Y., Li X., Chen H., Sun K., Li H., Zhou Y., Wang J., Bai F., Yang F., (2022), “Single-cell RNA sequencing reveals the multi-cellular ecosystem in different radiological components of pulmonary part-solid nodules”, *Clinical and Translational Medicine*, 12(2), e723.
- [35] Stuart T., Butler A., Hoffman P., Hafemeister C., Papalexi E., Mauck W. M., Hao Y., Stoeckius M., Smibert P., Satija R., (2019), “Comprehensive Integration of Single-Cell Data”, *Cell*, 177(7), 1888–1902.e21.
- [36] Stuart T., Satija R., (2019), “Integrative single-cell analysis”, *Nature Reviews Genetics*, 20(5), 257–272.
- [37] Korsunsky I., Millard N., Fan J., Slowikowski K., Zhang F., Wei K., Baglaenko Y., Brenner M., Loh P.-r., Raychaudhuri S., (2019), “Fast, sensitive and accurate integration of single-cell data with Harmony”, *Nature Methods*, 16(12), 1289–1296.
- [38] Saelens W., Cannoodt R., Todorov H., Saeys Y., (2019), “A comparison of single-cell trajectory inference methods”, *Nature Biotechnology*, 37(5), 547–554.
- [39] Angelidis I., Simon L. M., Fernandez I. E., Strunz M., Mayr C. H., Greiffo F. R., Tsitsiridis G., Ansari M., Graf E., Strom T.-M., Nagendran M., Desai T., Eickelberg O., Mann M., Theis F. J., Schiller H. B., (2019), “An atlas of the aging lung mapped by single cell transcriptomics and deep tissue proteomics”, *Nature Communications*, 10(1), 963.
- [40] Longo S. K., Guo M. G., Ji A. L., Khavari P. A., (2021), “Integrating single-cell and spatial transcriptomics to elucidate intercellular tissue dynamics”, *Nature Reviews Genetics*, 22(10), 627–644.
- [41] Argelaguet R., Cuomo A. S. E., Stegle O., Marioni J. C., (2021), “Computational

- principles and challenges in single-cell data integration”, *Nature Biotechnology*, 39(10), 1202–1215.
- [42] Leek J. T., (2014), “svaseq: removing batch effects and other unwanted noise from sequencing data”, *Nucleic Acids Research*, 42(21), e161.
 - [43] Hicks S. C., Townes F. W., Teng M., Irizarry R. A., (2018), “Missing data and technical variability in single-cell RNA-sequencing experiments”, *Biostatistics*, 19(4), 562–578.
 - [44] Luecken M. D., Büttner M., Chaichoompu K., Danese A., Interlandi M., Mueller M. F., Strobl D. C., Zappia L., Dugas M., Colomé-Tatché M., Theis F. J., (2022), “Benchmarking atlas-level data integration in single-cell genomics”, *Nature Methods*, 19(1), 41–50.
 - [45] Lähnemann D., Köster J., Szczurek E., McCarthy D. J., Hicks S. C., Robinson M. D., Vallejos C. A., Campbell K. R., Beerenwinkel N., Mahfouz A., Pinello L., Skums P., Stamatakis A., Attolini C. S.-O., Aparicio S., Baaijens J., Balvert M., Barbanson B. d., Cappuccio A., Corleone G., Dutilh B. E., Florescu M., Gurjev V., Holmer R., Jahn K., Lobo T. J., Keizer E. M., Khatri I., Kielbasa S. M., Korbel J. O., Kozlov A. M., Kuo T.-H., Lelieveldt B. P. F., Mandoiu I. I., Marioni J. C., Marschall T., Mölder F., Niknejad A., Rączkowska A., Reinders M., Ridder J. d., Saliba A.-E., Somarakis A., Stegle O., Theis F. J., Yang H., Zelikovsky A., McHardy A. C., Raphael B. J., Shah S. P., Schönhuth A., (2020), “Eleven grand challenges in single-cell data science”, *Genome Biology*, 21(1), 31.
 - [46] Sharma P. V., Thaïss C. A., (2020), “Host-Microbiome Interactions in the Era of Single-Cell Biology”, *Frontiers in Cellular and Infection Microbiology*, 10, 569070.
 - [47] Penaranda C., Hung D. T., (2019), “Single-Cell RNA Sequencing to Understand Host-Pathogen Interactions”, *ACS infectious diseases*, 5(3), 336–344.
 - [48] Mahalingam R., Dharmalingam P., Santhanam A., Kotla S., Davuluri G., Karmouty-Quintana H., Ashrith G., Thandavarayan R. A., (2021), “Single-cell RNA sequencing analysis of SARS-CoV-2 entry receptors in human organoids”, *Journal of Cellular Physiology*, 236(4), 2950–2958.
 - [49] Westermann A. J., Vogel J., (2021), “Cross-species RNA-seq for deciphering host–microbe interactions”, *Nature Reviews Genetics*, 22(6), 361–378. Publisher: Nature Publishing Group.
 - [50] Blattman S. B., Jiang W., Oikonomou P., Tavazoie S., (2020), “Prokaryotic single-cell RNA sequencing by in situ combinatorial indexing”, *Nature Microbiology*, 5(10), 1192–1201. Publisher: Nature Publishing Group.
 - [51] Imdahl F., Vafadarnejad E., Homberger C., Saliba A.-E., Vogel J., (2020), “Single-cell RNA-sequencing reports growth-condition-specific global transcriptomes of individual bacteria”, *Nature Microbiology*, 5(10), 1202–1206. Publisher: Nature Publishing Group.
 - [52] Wimmers F., Donato M., Kuo A., Ashuach T., Gupta S., Li C., Dvorak M., Foecke M. H., Chang S. E., Hagan T., De Jong S. E., Maecker H. T., van der Most R., Cheung P., Cortese M., Bosinger S. E., Davis M., Roupheal N.,

- Subramaniam S., Yosef N., Utz P. J., Khatri P., Pulendran B., (2021), “The single-cell epigenomic and transcriptional landscape of immunity to influenza vaccination”, *Cell*, 184(15), 3915–3935.e21.
- [53] Collora J. A., Liu R., Pinto-Santini D., Ravindra N., Ganoza C., Lama J. R., Alfaro R., Chiarella J., Spudich S., Mounzer K., Tebas P., Montaner L. J., van Dijk D., Duerr A., Ho Y.-C., (2022), “Single-cell multiomics reveals persistence of HIV-1 in expanded cytotoxic T cell clones”, *Immunity*, 55(6), 1013–1031.e7.
- [54] Meckiff B. J., Ramírez-Suástegui C., Fajardo V., Chee S. J., Kusnadi A., Simon H., Eschweiler S., Grifoni A., Pelosi E., Weiskopf D., Sette A., Ay F., Seumois G., Ottensmeier C. H., Vijayanand P., (2020), “Imbalance of Regulatory and Cytotoxic SARS-CoV-2-Reactive CD4⁺ T Cells in COVID-19”, *Cell*, 183(5), 1340–1353.e16.
- [55] Forcato M., Romano O., Biciato S., (2021), “Computational methods for the integrative analysis of single-cell data”, *Briefings in Bioinformatics*, 22(3), bbaa042.
- [56] Haghverdi L., Lun A. T. L., Morgan M. D., Marioni J. C., (2018), “Batch effects in single-cell RNA-sequencing data are corrected by matching mutual nearest neighbors”, *Nature Biotechnology*, 36(5), 421–427.
- [57] Yang Y., Li G., Qian H., Wilhelmsen K. C., Shen Y., Li Y., (2021), “SMNN: batch effect correction for single-cell RNA-seq data via supervised mutual nearest neighbor detection”, *Briefings in Bioinformatics*, 22(3), bbaa097.
- [58] Zhang F., Wu Y., Tian W., (2019), “A novel approach to remove the batch effect of single-cell data”, *Cell Discovery*, 5(1), 1–4.
- [59] Lin Y., Ghazanfar S., Wang K. Y. X., Gagnon-Bartsch J. A., Lo K. K., Su X., Han Z.-G., Ormerod J. T., Speed T. P., Yang P., Yang J. Y. H., (2019), “scMerge leverages factor analysis, stable expression, and pseudoreplication to merge multiple single-cell RNA-seq datasets”, *Proceedings of the National Academy of Sciences*, 116(20), 9775–9784.
- [60] Yan X., Zheng R., Wu F., Li M., (2023), “CLAIRE: contrastive learning-based batch correction framework for better balance between batch mixing and preservation of cellular heterogeneity”, *Bioinformatics*, 39(3), btad099.
- [61] Polański K., Young M. D., Miao Z., Meyer K. B., Teichmann S. A., Park J.-E., (2020), “BBKNN: fast batch alignment of single cell transcriptomes”, *Bioinformatics*, 36(3), 964–965.
- [62] Lopez R., Regier J., Cole M. B., Jordan M. I., Yosef N., (2018), “Deep generative modeling for single-cell transcriptomics”, *Nature Methods*, 15(12), 1053–1058.
- [63] Johansen N., Quon G., (2019), “scAlign: a tool for alignment, integration, and rare cell identification from scRNA-seq data”, *Genome Biology*, 20(1), 166.
- [64] Lotfollahi M., Naghipourfar M., Theis F. J., Wolf F. A., (2020), “Conditional out-of-distribution generation for unpaired data using transfer VAE”, *Bioinformatics*, 36(Supplement_2), i610–i617.

- [65] Xu C., Lopez R., Mehlman E., Regier J., Jordan M. I., Yosef N., (2021), “Probabilistic harmonization and annotation of single-cell transcriptomics data with deep generative models”, *Molecular Systems Biology*, 17(1), e9620.
- [66] Bennett H. M., Stephenson W., Rose C. M., Darmanis S., (2023), “Single-cell proteomics enabled by next-generation sequencing or mass spectrometry”, *Nature Methods*, 20(3), 363–374.
- [67] Zhu B., Chen S., Bai Y., Chen H., Liao G., Mukherjee N., Vazquez G., McIlwain D. R., Tzankov A., Lee I. T., Matter M. S., Goltsev Y., Ma Z., Nolan G. P., Jiang S., (2023), “Robust single-cell matching and multimodal analysis using shared and distinct features”, *Nature Methods*, 1–12.
- [68] Kingma D. P., Welling M., (2022), “Auto-Encoding Variational Bayes”.
- [69] Hie B., Bryson B., Berger B., (2019), “Efficient integration of heterogeneous single-cell transcriptomes using Scanorama”, *Nature Biotechnology*, 37(6), 685–691.
- [70] Grønbech C. H., Vording M. F., Timshel P. N., Sønderby C. K., Pers T. H., Winther O., (2020), “scVAE: variational auto-encoders for single-cell gene expression data”, *Bioinformatics*, 36(16), 4415–4422.
- [71] McInnes L., Healy J., Melville J., (2020), “UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction”, arXiv:1802.03426 [cs, stat].
- [72] Hao Y., Hao S., Andersen-Nissen E., Mauck W. M., Zheng S., Butler A., Lee M. J., Wilk A. J., Darby C., Zager M., Hoffman P., Stoeckius M., Papalexi E., Mimitou E. P., Jain J., Srivastava A., Stuart T., Fleming L. M., Yeung B., Rogers A. J., McElrath J. M., Blish C. A., Gottardo R., Smibert P., Satija R., (2021), “Integrated analysis of multimodal single-cell data”, *Cell*, 184(13), 3573–3587.e29.
- [73] Ester M., Kriegel H.-P., Sander J., Xu X., (1996), A density-based algorithm for discovering clusters in large spatial databases with noise, in “Proceedings of the Second International Conference on Knowledge Discovery and Data Mining”, KDD’96, AAAI Press, Portland, Oregon, 226–231.
- [74] Rahmah N., Sitanggang I. S., (2016), “Determination of Optimal Epsilon (Eps) Value on DBSCAN Algorithm to Clustering Data on Peatland Hotspots in Sumatra”, *IOP Conference Series: Earth and Environmental Science*, 31(1), 012012.
- [75] Kullback S., Leibler R. A., (1951), “On Information and Sufficiency”, *The Annals of Mathematical Statistics*, 22(1), 79–86.
- [76] Pedregosa F., Varoquaux G., Gramfort A., Michel V., Thirion B., Grisel O., Blondel M., Prettenhofer P., Weiss R., Dubourg V., Vanderplas J., Passos A., Cournapeau D., Brucher M., Perrot M., Duchesnay E., (2011), “Scikit-learn: Machine Learning in Python”, *J. Mach. Learn. Res.*, 12(null), 2825–2830.
- [77] Rahil Z., Leylek R., Schürch C. M., Chen H., Bjornson-Hooper Z., Christensen S. R., Gherardini P. F., Bhate S. S., Spitzer M. H., Fragiadakis G. K., Mukherjee N., Kim N., Jiang S., Yo J., Gaudilliere B., Affrime M., Bock B., Hensley S. E., Idoyaga J., Aghaeepour N., Kim K., Nolan G. P., McIlwain D. R., (2020), “Landscape of

- coordinated immune responses to H1N1 challenge in humans”, *The Journal of Clinical Investigation*, 130(11), 5800–5816.
- [78] Bjornson-Hooper Z. B., Fragiadakis G. K., Spitzer M. H., Chen H., Madhiredy D., Hu K., Lundsten K., McIlwain D. R., Nolan G. P., (2022), “A Comprehensive Atlas of Immunological Differences Between Humans, Mice, and Non-Human Primates”, *Frontiers in Immunology*, 13, 867015.
 - [79] Lotfollahi M., Naghipourfar M., Luecken M. D., Khajavi M., Büttner M., Wagenstetter M., Avsec Z., Gayoso A., Yosef N., Interlandi M., Rybakov S., Misharin A. V., Theis F. J., (2022), “Mapping single-cell data to reference atlases by transfer learning”, *Nature Biotechnology*, 40(1), 121–130.
 - [80] Leduc A., Huffman R. G., Cantlon J., Khan S., Slavov N., (2022), “Exploring functional protein covariation across single cells using nPOP”, *Genome Biology*, 23(1), 1–31.
 - [81] Cao Z.-J., Gao G., (2022), “Multi-omics single-cell data integration and regulatory inference with graph-linked embedding”, *Nature Biotechnology*, 40(10), 1458–1466. Publisher: Nature Publishing Group.
 - [82] Baysoy A., Bai Z., Satija R., Fan R., (2023), “The technological landscape and applications of single-cell multi-omics”, *Nature Reviews Molecular Cell Biology*, 24(10), 695–713. Number: 10 Publisher: Nature Publishing Group.
 - [83] Rautenstrauch P., Vlot A. H. C., Saran S., Ohler U., (2022), “Intricacies of single-cell multi-omics data integration”, *Trends in Genetics*, 38(2), 128–139. Publisher: Elsevier.
 - [84] Amodio M., Krishnaswamy S., (2018), MAGAN: Aligning biological manifolds, in J. Dy A. Krause, eds, “Proceedings of the 35th International Conference on Machine Learning”, Vol. 80 of *Proceedings of Machine Learning Research*, PMLR, 215–223.
 - [85] Wang X., Wu X., Hong N., Jin W., (2024), “Progress in single-cell multimodal sequencing and multi-omics data integration”, *Biophysical Reviews*, 16(1), 13–28.
 - [86] Wang X., Sun Z., Zhang Y., Xu Z., Xin H., Huang H., Duerr R. H., Chen K., Ding Y., Chen W., (2020), “BREM-SC: a bayesian random effects mixture model for joint clustering single cell multi-omics data”, *Nucleic Acids Research*, 48(11), 5814–5824.
 - [87] Singh R., Hie B. L., Narayan A., Berger B., (2021), “Schema: metric learning enables interpretable synthesis of heterogeneous single-cell modalities”, *Genome Biology*, 22(1), 131.
 - [88] Argelaguet R., Arnol D., Bredikhin D., Deloro Y., Velten B., Marioni J. C., Stegle O., (2020), “MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data”, *Genome Biology*, 21(1), 111.
 - [89] Liu J., Gao C., Sodico J., Kozareva V., Macosko E. Z., Welch J. D., (2020), “Jointly defining cell types from multiple single-cell datasets using LIGER”, *Nature Protocols*, 15(11), 3632–3662. Publisher: Nature Publishing Group.

- [90] Huizing G.-J., Deutschmann I. M., Peyré G., Cantini L., (2023), “Paired single-cell multi-omics data integration with Mowgli”, *Nature Communications*, 14(1), 7711. Publisher: Nature Publishing Group.
- [91] Dou J., Liang S., Mohanty V., Miao Q., Huang Y., Liang Q., Cheng X., Kim S., Choi J., Li Y., Li L., Daher M., Basar R., Rezvani K., Chen R., Chen K., (2022), “Bi-order multimodal integration of single-cell data”, *Genome Biology*, 23(1), 112.
- [92] Gayoso A., Steier Z., Lopez R., Regier J., Nazor K. L., Streets A., Yosef N., (2021), “Joint probabilistic modeling of single-cell multi-omic data with totalVI”, *Nature Methods*, 18(3), 272–282. Publisher: Nature Publishing Group.
- [93] Zuo C., Chen L., (2021), “Deep-joint-learning analysis model of single cell transcriptome and open chromatin accessibility data”, *Briefings in Bioinformatics*, 22(4), bbaa287.
- [94] Qattous H., Azzeh M., Ibrahim R., Abed Al-Ghafer I., Al Sorkhy M., Alkhateeb A., (2024), “PaCMAP-embedded convolutional neural network for multi-omics data integration”, *Heliyon*, 10(1), e23195.
- [95] Wen H., Ding J., Jin W., Wang Y., Xie Y., Tang J., (2022), Graph Neural Networks for Multimodal Single-Cell Data Integration, in “Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining”, KDD '22, Association for Computing Machinery, New York, NY, USA, 4153–4163.
- [96] Athaya T., Ripan R. C., Li X., Hu H., (2023), “Multimodal deep learning approaches for single-cell multi-omics data integration”, *Briefings in Bioinformatics*, 24(5), bbad313.
- [97] John C. R., Watson D., Barnes M. R., Pitzalis C., Lewis M. J., (2020), “Spectrum: fast density-aware spectral clustering for single and multi-omic data”, *Bioinformatics*, 36(4), 1159–1166.
- [98] Chen H., Ryu J., Vinyard M. E., Lerer A., Pinello L., (2024), “SIMBA: single-cell embedding along with features”, *Nature Methods*, 21(6), 1003–1013. Publisher: Nature Publishing Group.
- [99] Jardine L., Webb S., Goh I., Quiroga Londoño M., Reynolds G., Mather M., Olabi B., Stephenson E., Botting R. A., Horsfall D., Engelbert J., Maunder D., Mende N., Murnane C., Dann E., McGrath J., King H., Kucinski I., Queen R., Carey C. D., Shrubsole C., Poyner E., Acres M., Jones C., Ness T., Coulthard R., Elliott N., O’Byrne S., Haltalli M. L. R., Lawrence J. E., Lisgo S., Balogh P., Meyer K. B., Prigmore E., Ambridge K., Jain M. S., Efremova M., Pickard K., Creasey T., Bacardit J., Henderson D., Coxhead J., Filby A., Hussain R., Dixon D., McDonald D., Popescu D.-M., Kowalczyk M. S., Li B., Ashenberg O., Tabaka M., Dionne D., Tickle T. L., Slyper M., Rozenblatt-Rosen O., Regev A., Behjati S., Laurenti E., Wilson N. K., Roy A., Göttgens B., Roberts I., Teichmann S. A., Haniffa M., (2021), “Blood and immune development in human fetal bone marrow and Down syndrome”, *Nature*, 598(7880), 327–331.
- [100] Szklarczyk D., Morris J. H., Cook H., Kuhn M., Wyder S., Simonovic M., Santos A., Doncheva N. T., Roth A., Bork P., Jensen L. J., von Mering C., (2017), “The STRING database in 2017: quality-controlled protein–protein

- association networks, made broadly accessible”, *Nucleic Acids Research*, 45(D1), D362–D368.
- [101] Milacic M., Beavers D., Conley P., Gong C., Gillespie M., Griss J., Haw R., Jassal B., Matthews L., May B., Petryszak R., Ragueneau E., Rothfels K., Sevilla C., Shamovsky V., Stephan R., Tiwari K., Varusai T., Weiser J., Wright A., Wu G., Stein L., Hermjakob H., D’Eustachio P., (2023), “The Reactome Pathway Knowledgebase 2024”, *Nucleic Acids Research*, 52(D1), D672.
 - [102] Wolf F. A., Angerer P., Theis F. J., (2018), “SCANPY: large-scale single-cell gene expression data analysis”, *Genome Biology*, 19(1), 15.
 - [103] Hamilton W. L., Ying R., Leskovec J., (2017), Inductive representation learning on large graphs, in “Proceedings of the 31st International Conference on Neural Information Processing Systems”, NIPS’17, Curran Associates Inc., Red Hook, NY, USA, 1025–1035.
 - [104] Xiao C., Chen Y., Meng Q., Wei L., Zhang X., (2024), “Benchmarking multi-omics integration algorithms across single-cell RNA and ATAC data”, *Briefings in Bioinformatics*, 25(2), bbae095.
 - [105] Traag V., Waltman L., Eck N. J. v., (2019), “From Louvain to Leiden: guaranteeing well-connected communities”, *arXiv:1810.08473*.
 - [106] Kolberg L., Raudvere U., Kuzmin I., Adler P., Vilo J., Peterson H., (2023), “g:Profiler-interoperable web service for functional enrichment analysis and gene identifier mapping (2023 update)”, *Nucleic Acids Research*, 51(W1), W207–W212.
 - [107] Regev A., Teichmann S. A., Lander E. S., Amit I., Benoist C., Birney E., Bodenmiller B., Campbell P., Carninci P., Clatworthy M., Clevers H., Deplancke B., Dunham I., Eberwine J., Eils R., Enard W., Farmer A., Fugger L., Göttgens B., Hacohen N., Haniffa M., Hemberg M., Kim S., Klenerman P., Kriegstein A., Lein E., Linnarsson S., Lundberg E., Lundeberg J., Majumder P., Marioni J. C., Merad M., Mhlanga M., Nawijn M., Netea M., Nolan G., Pe’er D., Phillipakis A., Ponting C. P., Quake S., Reik W., Rozenblatt-Rosen O., Sanes J., Satija R., Schumacher T. N., Shalek A., Shapiro E., Sharma P., Shin J. W., Stegle O., Stratton M., Stubbington M. J. T., Theis F. J., Uhlen M., van Oudenaarden A., Wagner A., Watt F., Weissman J., Wold B., Xavier R., Yosef N., Human Cell Atlas Meeting Participants (2017), “The Human Cell Atlas”, *eLife*, 6, e27041. Publisher: eLife Sciences Publications, Ltd.
 - [108] Bredikhin D., Kats I., Stegle O., (2022), “MUON: multimodal omics analysis framework”, *Genome Biology*, 23(1), 42.
 - [109] Cui A., Li B., Wallace M. S., Gonye A. L. K., Oetheimer C., Patel H., Tonnerre P., Holmes J. A., Lieb D., Yao B. S., Ma A., Roberts K., Damasio M., Chen J. H., Piou D., Carlton-Smith C., Brown J., Mylvaganam R., Hon Fung J. M., Sade-Feldman M., Aneja J., Gustafson J., Epstein E. T., Salloum S., Brisac C., Thabet A., Kim A. Y., Lauer G. M., Hacohen N., Chung R. T., Alatrakchi N., (2024), “Single-cell atlas of the liver myeloid compartment before and after cure of chronic viral hepatitis”, *Journal of Hepatology*, 80(2), 251–267.
 - [110] Wang S., Zhang Q., Hui H., Agrawal K., Karris M. A. Y., Rana T. M., (2020), “An

- atlas of immune cell exhaustion in HIV-infected individuals revealed by single-cell transcriptomics”, *Emerging Microbes & Infections*, 9(1), 2333–2347.
- [111] Phaahla N. G., Lassaunière R., Da Costa Dias B., Waja Z., Martinson N. A., Tiemessen C. T., (2019), “Chronic HIV-1 Infection Alters the Cellular Distribution of FcγRIIIa and the Functional Consequence of the FcγRIIIa-F158V Variant”, *Frontiers in Immunology*, 10, 735.
- [112] Poonia B., Kijak G. H., Pauza C. D., (2010), “High Affinity Allele for the Gene of FCGR3A Is Risk Factor for HIV Infection and Progression”, *PLOS ONE*, 5(12), e15562. Publisher: Public Library of Science.
- [113] Chen Y., Li X., Liu S., Ao W., Lin J., Li Z., Wu S., Ye H., Han X., Li D., (2023), “An atlas of immune cell transcriptomes in human immunodeficiency virus-infected immunological non-responders identified marker genes that control viral replication”, *Chinese Medical Journal*, 136(22), 2694–2705.
- [114] Uhlén M., Karlsson M. J., Hober A., Svensson A.-S., Scheffel J., Kotol D., Zhong W., Tebani A., Strandberg L., Edfors F., Sjöstedt E., Mulder J., Mardinoglu A., Berling A., Ekblad S., Dannemeyer M., Kanje S., Rockberg J., Lundqvist M., Malm M., Volk A.-L., Nilsson P., Månberg A., Dodig-Crnkovic T., Pin E., Zwahlen M., Oksvold P., von Feilitzen K., Häussler R. S., Hong M.-G., Lindskog C., Ponten F., Katona B., Vu J., Lindström E., Nielsen J., Robinson J., Ayoglu B., Mahdessian D., Sullivan D., Thul P., Danielsson F., Stadler C., Lundberg E., Bergström G., Gummesson A., Voldborg B. G., Tegel H., Hober S., Forsström B., Schwenk J. M., Fagerberg L., Sivertsson A., (2019), “The human secretome”, *Science Signaling*, 12(609), eaaz0274.
- [115] Anlauf M., Schäfer M. K.-H., Schwark T., von Wurmb-Schwark N., Brand V., Sipos B., Horny H.-P., Parwaresch R., Hartschuh W., Eiden L. E., Klöppel G., Weihe E., (2006), “Vesicular monoamine transporter 2 (VMAT2) expression in hematopoietic cells and in patients with systemic mastocytosis”, *The Journal of Histochemistry and Cytochemistry: Official Journal of the Histochemistry Society*, 54(2), 201–213.
- [116] Arock M., Schneider E., Boissan M., Tricottet V., Dy M., (2002), “Differentiation of human basophils: an overview of recent advances and pending questions”, *Journal of Leukocyte Biology*, 71(4), 557–564.
- [117] Welner R. S., Esplin B. L., Garrett K. P., Pelayo R., Luche H., Fehling H. J., Kincade P. W., (2009), “Asynchronous RAG-1 expression during B lymphopoiesis”, *Journal of immunology (Baltimore, Md. : 1950)*, 183(12), 7768.
- [118] Luger D., Yang Y.-a., Raviv A., Weinberg D., Banerjee S., Lee M.-J., Trepel J., Yang L., Wakefield L. M., (2013), “Expression of the B-Cell Receptor Component CD79a on Immature Myeloid Cells Contributes to Their Tumor Promoting Effects”, *PLOS ONE*, 8(10), e76115. Publisher: Public Library of Science.
- [119] Kaiser F. M. P., Janowska I., Menafrà R., de Gier M., Korzhenevich J., Pico-Knijnenburg I., Khatrı I., Schulz A., Kuijpers T. W., Lankester A. C., Konstantinidis L., Erlacher M., Kloet S., van Schouwenburg P. A., Rizzi M., van der Burg M., (2023), “IL-7 receptor signaling drives human B-cell progenitor differentiation and expansion”, *Blood*, 142(13), 1113–1130.

- [120] Hoebeke I., De Smedt M., Stolz F., Pike-Overzet K., Staal F. J. T., Plum J., Leclercq G., (2007), “T-, B- and NK-lymphoid, but not myeloid cells arise from human CD34(+)CD38(-)CD7(+) common lymphoid progenitors expressing lymphoid-specific genes”, *Leukemia*, 21(2), 311–319.
- [121] Liang K. L., Laurenti E., Taghon T., (2023), “Circulating IRF8-expressing CD123+CD127+ lymphoid progenitors: key players in human hematopoiesis”, *Trends in Immunology*, 44(9), 678–692.
- [122] Guo H., Barberi T., Suresh R., Friedman A. D., (2018), “Progression from the Common Lymphoid Progenitor to B/Myeloid PreproB and ProB Precursors during B Lymphopoiesis Requires C/EBP α ”, *Journal of Immunology (Baltimore, Md.: 1950)*, 201(6), 1692–1704.
- [123] Morgan D., Tergaonkar V., (2022), “Unraveling B cell trajectories at single cell resolution”, *Trends in Immunology*, 43(3), 210–229.
- [124] Cui T., El Mekkaoui K., Reinval J., Havulinna A. S., Marttinen P., Kaski S., (2022), “Gene–gene interaction detection with deep learning”, *Communications Biology*, 5(1), 1–12. Publisher: Nature Publishing Group.
- [125] Clough E., Barrett T., (2016), The Gene Expression Omnibus Database, in E. Mathé S. Davis, eds, “Statistical Genomics: Methods and Protocols”, *Methods in Molecular Biology*, Springer, New York, NY, 93–110.
- [126] Tang Y.-W., Hindiyeh M., Liu D., Sails A., Spearman P., Zhang J.-R., (2023), “Molecular Medical Microbiology”, Vol. 1, Academic Press. Google-Books-ID: 6EsiEAAAQBAJ.
- [127] Huang W., Wang D., Yao Y.-F., (2021), “Understanding the pathogenesis of infectious diseases by single-cell RNA sequencing”, *Microbial Cell*, 8(9), 208–222.
- [128] Wang C., Forst C. V., Chou T.-w., Geber A., Wang M., Hamou W., Smith M., Sebra R., Zhang B., Zhou B., Ghedin E., (2020), “Cell-to-Cell Variation in Defective Virus Expression and Effects on Host Responses during Influenza Virus Infection”, *mBio*, 11(1), e02880–19.
- [129] Cai Y., Dai Y., Wang Y., Yang Q., Guo J., Wei C., Chen W., Huang H., Zhu J., Zhang C., Zheng W., Wen Z., Liu H., Zhang M., Xing S., Jin Q., Feng C. G., Chen X., (2020), “Single-cell transcriptomics of blood reveals a natural killer cell subset depletion in tuberculosis”, *EBioMedicine*, 53, 102686.
- [130] Zhang J., Liu J., Yuan Y., Huang F., Ma R., Luo B., Xi Z., Pan T., Liu B., Zhang Y., Zhang X., Luo Y., Wang J., Zhao M., Lu G., Deng K., Zhang H., (2020), “Two waves of pro-inflammatory factors are released during the influenza A virus (IAV)-driven pulmonary immunopathogenesis”, *PLoS pathogens*, 16(2), e1008334.
- [131] Gierahn T. M., Wadsworth M. H., Hughes T. K., Bryson B. D., Butler A., Satija R., Fortune S., Love J. C., Shalek A. K., (2017), “Seq-Well: portable, low-cost RNA sequencing of single cells at high throughput”, *Nature Methods*, 14(4), 395–398.
- [132] Aulicino A., Rue-Albrecht K. C., Preciado-Llanes L., Napolitani G., Ashley N., Cribbs A., Koth J., Lagerholm B. C., Ambrose T., Gordon M. A., Sims D., Simmons

- A., (2018), “Invasive Salmonella exploits divergent immune evasion strategies in infected and bystander dendritic cell subsets”, *Nature Communications*, 9(1), 4883.
- [133] Steuerman Y., Cohen M., Peshes-Yaloz N., Valadarsky L., Cohn O., David E., Frishberg A., Mayo L., Bacharach E., Amit I., Gat-Viks I., (2018), “Dissection of Influenza Infection In Vivo by Single-Cell RNA Sequencing”, *Cell Systems*, 6(6), 679–691.e4.
- [134] Durmuş Tekir S., Çakır T., Ardiç E., Sayılırbaş A. S., Konuk G., Konuk M., Sarıyer H., Uğurlu A., Karadeniz I., Özgür A., Sevilgen F. E., Ülgen K. O., (2013), “PHISTO: pathogen–host interaction search tool”, *Bioinformatics*, 29(10), 1357–1358.
- [135] Sanchez-Martinez A., Perdomo-Celis F., Acevedo-Saenz L., Rugeles M. T., Velilla P. A., (2019), “Cytotoxic CD4+ T-cells during HIV infection: Targets or weapons?”, *Journal of Clinical Virology: The Official Publication of the Pan American Society for Clinical Virology*, 119, 17–23.
- [136] Sayers E. W., Bolton E. E., Brister J. R., Canese K., Chan J., Comeau D., Connor R., Funk K., Kelly C., Kim S., Madej T., Marchler-Bauer A., Lanczycki C., Lathrop S., Lu Z., Thibaud-Nissen F., Murphy T., Phan L., Skripchenko Y., Tse T., Wang J., Williams R., Trawick B., Pruitt K., Sherry S., (2021), “Database resources of the National Center for Biotechnology Information”, *Nucleic Acids Research*, 50(D1), D20–D26.
- [137] The UniProt Consortium (2025), “UniProt: the Universal Protein Knowledgebase in 2025”, *Nucleic Acids Research*, 53(D1), D609–D617.
- [138] Flórez-Álvarez L., Hernandez J. C., Zapata W., (2018), “NK Cells in HIV-1 Infection: From Basic Science to Vaccine Strategies”, *Frontiers in Immunology*. Publisher: Frontiers.
- [139] Doitsh G., Greene W. C., (2016), “Dissecting How CD4 T Cells Are Lost During HIV Infection”, *Cell Host & Microbe*, 19(3), 280–291. Publisher: Elsevier.
- [140] Kumar A., Abbas W., Herbein G., (2014), “HIV-1 latency in monocytes/macrophages”, *Viruses*, 6(4), 1837–1860.
- [141] Hu Z., Luo Z., Wan Z., Wu H., Li W., Zhang T., Jiang W., (2015), “HIV-associated memory B cell perturbations”, *Vaccine*, 33(22), 2524–2529.
- [142] Nasi A., Amu S., Göthlin M., Jansson M., Nagy N., Chiodi F., Réthi B., (2017), “Dendritic Cell Response to HIV-1 Is Controlled by Differentiation Programs in the Cells and Strain-Specific Properties of the Virus”, *Frontiers in Immunology*, 8, 244.
- [143] Nyanhete T. E., Frisbee A. L., Bradley T., Faison W. J., Robins E., Payne T., Freel S. A., Sawant S., Weinhold K. J., Wiehe K., Haynes B. F., Ferrari G., Li Q.-J., Moody M. A., Tomaras G. D., (2019), “HLA class II-Restricted CD8+ T cells in HIV-1 Virus Controllers”, *Scientific Reports*, 9(1), 10165. Publisher: Nature Publishing Group.
- [144] Kornbluth R. S., (2000), “The emerging role of CD40 ligand in HIV infection”, *Journal of Leukocyte Biology*, 68(3), 373–382.

- [145] Qian S., Xiong C., Wang M., Zhang Z., Fu Y., Hu Q., Ding H., Han X., Shang H., Jiang Y., (2022), “CD38+CD39+ NK cells associate with HIV disease progression and negatively regulate T cell proliferation”, *Frontiers in Immunology*. Publisher: Frontiers.
- [146] Gebremicael G., Gebreegziabxier A., Kassa D., (2022), “Low transcriptomic of PTPRCv1 and CD3E is an independent predictor of mortality in HIV and tuberculosis co-infected patient”, *Scientific Reports*, 12(1), 10133.
- [147] Kaufmann D. E., Walker B. D., (2009), “PD-1 and CTLA-4 inhibitory cosignaling pathways in HIV infection and the potential for therapeutic intervention”, *Journal of Immunology* (Baltimore, Md.: 1950), 182(10), 5891–5897.
- [148] Arthos J., Cicala C., Nawaz F., Byraredy S. N., Villinger F., Santangelo P. J., Ansari A. A., Fauci A. S., (2018), “The Role of Integrin $\alpha 4\beta 7$ in HIV Pathogenesis and Treatment”, *Current HIV/AIDS reports*, 15(2), 127–135.
- [149] Gulzar N., Copeland K. F. T., (2004), “CD8+ T-cells: function and response to HIV infection”, *Current HIV research*, 2(1), 23–37.
- [150] Collins D. R., Gaiha G. D., Walker B. D., (2020), “CD8+ T cells in HIV control, cure and prevention”, *Nature Reviews Immunology*, 20(8), 471–482. Publisher: Nature Publishing Group.
- [151] Wang X.-M., Zhang J.-Y., Xing X., Huang H.-H., Xia P., Dai X.-P., Hu W., Zhang C., Song J.-W., Fan X., Wu F.-Y., Liu F.-H., Ke Y., Zhao Y., Jiang T.-J., Wang L.-F., Jiao Y.-M., Xu R.-N., Jin L., Shi M., Bai F., Wang F.-S., (2022), “Global transcriptomic characterization of T cells in individuals with chronic HIV-1 infection”, *Cell Discovery*, 8(1), 1–17. Publisher: Nature Publishing Group.
- [152] Elmasri R. A., Rashwan A. A., Gaber S. H., Rostom M. M., Karousi P., Yasser M. B., Kontos C. K., Youness R. A., (2024), “Puzzling out the role of MIAT LncRNA in hepatocellular carcinoma”, *Non-coding RNA research*, 9(2), 547–559.
- [153] Pizzolato G., Kaminski H., Tosolini M., Franchini D.-M., Pont F., Martins F., Valle C., Labourdette D., Cadot S., Quillet-Mary A., Poupot M., Laurent C., Ysebaert L., Meraviglia S., Dieli F., Merville P., Milpied P., Déchanet-Merville J., Fournié J.-J., (2019), “Single-cell RNA sequencing unveils the shared and the distinct cytotoxic hallmarks of human TCRV $\delta 1$ and TCRV $\delta 2$ $\gamma\delta$ T lymphocytes”, *Proceedings of the National Academy of Sciences of the United States of America*, 116(24), 11906–11915.
- [154] Lopalco L., (2010), “CCR5: From Natural Resistance to a New Anti-HIV Strategy”, *Viruses*, 2(2), 574–600.
- [155] Pawlak E. N., Dirk B. S., Jacob R. A., Johnson A. L., Dikeakos J. D., (2018), “The HIV-1 accessory proteins Nef and Vpu downregulate total and cell surface CD28 in CD4+ T cells”, *Retrovirology*, 15(1), 6.
- [156] Matheson N. J., Sumner J., Wals K., Rapiteanu R., Weekes M. P., Vigan R., Weinelt J., Schindler M., Antrobus R., Costa A. S. H., Frezza C., Clish C. B., Neil S. J. D., Lehner P. J., (2015), “Cell Surface Proteomic Map of HIV Infection Reveals Antagonism of Amino Acid Metabolism by Vpu and Nef”, *Cell Host & Microbe*, 18(4), 409–423.

- [157] Girard A., Vergnon-Miszczycha D., Depincé-Berger A.-E., Roblin X., Lutch F., Lambert C., Rochereau N., Bourlet T., Genin C., Paul S., (2016), “Brief Report: A High Rate of $\beta 7^+$ Gut-Homing Lymphocytes in HIV-Infected Immunological Nonresponders is Associated With Poor CD4 T-Cell Recovery During Suppressive HAART”, *Journal of Acquired Immune Deficiency Syndromes* (1999), 72(3), 259–265.
- [158] Kassa D., Ran L., Jager W. d., van den Broek T., Jacobi R., Mekonen M., Messele T., Haks M. C., Ottenhoff T. H. M., van Baarle D., (2016), “Discriminative expression of whole blood genes in HIV patients with latent and active TB in Ethiopia”, *Tuberculosis*, 100, 25–31.
- [159] Pinti M., Pedrazzi J., Benatti F., Sorrentino V., Nuzzo C., Cavazzuti V., Biswas P., Petrusca D. N., Mussini C., De Rienzo B., Cossarizza A., (1999), “Differential down-regulation of CD95 or CD95L in chronically HIV-infected cells of monocytic or lymphocytic origin: cellular studies and molecular analysis by quantitative competitive RT-PCR”, *FEBS letters*, 458(2), 209–214.
- [160] Liechti T., Kadelka C., Braun D. L., Kuster H., Böni J., Robbiani M., Günthard H. F., Trkola A., (2019), “Widespread B cell perturbations in HIV-1 infection afflict naive and marginal zone B cells”, *The Journal of Experimental Medicine*, 216(9), 2071–2090.
- [161] Fu E., Pan L., Xie Y., Mu D., Liu W., Jin F., Bai X., (2015), “Tetraspanin CD63 is a regulator of HIV-1 replication”, *International Journal of Clinical and Experimental Pathology*, 8(2), 1184–1198.
- [162] Negash M., Chanyalew M., Girma T., Alemu F., Alcantara D., Towler B., Davey G., Boyton R. J., Altmann D. M., Howe R., Newport M. J., (2024), “Evidence for immune activation in pathogenesis of the HLA class II associated disease, podocniosis”, *Nature Communications*, 15(1), 2020.
- [163] Oravec T., Roderiquez G., Koffi J., Wang J., Ditto M., Bou-Habib D. C., Lusso P., Norcross M. A., (1995), “CD26 expression correlates with entry, replication and cytopathicity of monocytopathic HIV–1 strains in a T–cell line”, *Nature Medicine*, 1(9), 919–926. Publisher: Nature Publishing Group.
- [164] Avalos B., Kulbe J. R., Ford M. K., Laird A. E., Walter K., Mante M., Florio J. B., Boustani A., Chaillon A., Schlachetzki J. C. M., Sundermann E. E., Volsky D. J., Rissman R. A., Ellis R. J., Letendre S. L., Iudicello J., Fields J. A., (2024), “Cannabis Use and Cannabidiol Modulate HIV-Induced Alterations in TREM2 Expression: Implications for Age-Related Neuropathogenesis”, *Viruses*, 16(10), 1509.
- [165] Haller C., Müller B., Fritz J. V., Lamas-Murua M., Stolp B., Pujol F. M., Keppler O. T., Fackler O. T., (2014), “HIV-1 Nef and Vpu are functionally redundant broad-spectrum modulators of cell surface receptors, including tetraspanins”, *Journal of Virology*, 88(24), 14241–14257.
- [166] Boulet S., Song R., Kamya P., Bruneau J., Shoukry N. H., Tsoukas C. M., Bernard N. F., (2010), “HIV protective KIR3DL1 and HLA-B genotypes influence NK cell function following stimulation with HLA-devoid cells”, *Journal of Immunology* (Baltimore, Md.: 1950), 184(4), 2057–2064.

- [167] Iglesias-Ussel M., Vandergeeten C., Marchionni L., Chomont N., Romerio F., (2013), “High Levels of CD2 Expression Identify HIV-1 Latently Infected Resting Memory CD4⁺ T Cells in Virally Suppressed Subjects”, *Journal of Virology*, 87(16), 9148–9158.
- [168] Bellini N., Ye C., Ajibola O., Murooka T. T., Lodge R., Cohen E. A., (2024), “Downregulation of miRNA-26a by HIV-1 Enhances CD59 Expression and Packaging, Impacting Virus Susceptibility to Antibody-Dependent Complement-Mediated Lysis”, *Viruses*, 16(7), 1076. Number: 7 Publisher: Multidisciplinary Digital Publishing Institute.
- [169] Lafrenie R. M., Wahl L. M., Epstein J. S., Hewlett I. K., Yamada K. M., Dhawan S., (1996), “HIV-1-Tat modulates the function of monocytes and alters their interactions with microvessel endothelial cells. A mechanism of HIV pathogenesis”, *Journal of Immunology* (Baltimore, Md.: 1950), 156(4), 1638–1645.
- [170] Fiorelli V., Barillari G., Toschi E., Sgadari C., Monini P., Stürzl M., Ensoli B., (1999), “IFN-gamma induces endothelial cells to proliferate and to invade the extracellular matrix in response to the HIV-1 Tat protein: implications for AIDS-Kaposi’s sarcoma pathogenesis”, *Journal of Immunology* (Baltimore, Md.: 1950), 162(2), 1165–1170.
- [171] Jelicic K., Cimbro R., Nawaz F., Huang D. W., Zheng X., Yang J., Lempicki R. A., Pascuccio M., Van Ryk D., Schwing C., Hiatt J., Okwara N., Wei D., Roby G., David A., Hwang I. Y., Kehrl J. H., Arthos J., Cicala C., Fauci A. S., (2013), “The HIV-1 envelope protein gp120 impairs B cell proliferation by inducing TGF- β 1 production and FcRL4 expression”, *Nature Immunology*, 14(12), 1256–1265.
- [172] Chen N., McCarthy C., Drakesmith H., Li D., Cerundolo V., McMichael A. J., Screaton G. R., Xu X.-N., (2006), “HIV-1 down-regulates the expression of CD1d via Nef”, *European Journal of Immunology*, 36(2), 278–286.
- [173] Lee J.-H., Schierer S., Blume K., Dindorf J., Wittki S., Xiang W., Ostalecki C., Koliha N., Wild S., Schuler G., Fackler O. T., Saksela K., Harrer T., Baur A. S., (2016), “HIV-Nef and ADAM17-Containing Plasma Extracellular Vesicles Induce and Correlate with Immune Pathogenesis in Chronic HIV Infection”, *EBioMedicine*, 6, 103–113.
- [174] Crawley A. M., Angel J. B., (2012), “The influence of HIV on CD127 expression and its potential implications for IL-7 therapy”, *Seminars in Immunology*, 24(3), 231–240.
- [175] CREERY D., WEISS W., LIM W. T., AZIZ Z., ANGEL J. B., KUMAR A., (2004), “Down-regulation of CXCR-4 and CCR-5 expression by interferon- γ is associated with inhibition of chemotaxis and human immunodeficiency virus (HIV) replication but not HIV entry into human monocytes”, *Clinical and Experimental Immunology*, 137(1), 156–165.
- [176] Lambelé M., Koppensteiner H., Symeonides M., Roy N. H., Chan J., Schindler M., Thali M., (2015), “Vpu Is the Main Determinant for Tetraspanin Downregulation in HIV-1-Infected Cells”, *Journal of Virology*, 89(6), 3247–3255.
- [177] KUMAR A., ANGEL B., AUCCOIN S., CREERY W., DAFTARIAN M.,

- CAMERON D., FILION L., DIAZ-MITOMA F., (1999), “Dysregulation of B7.2 (CD86) expression on monocytes of HIV-infected individuals is associated with altered production of IL-2”, *Clinical and Experimental Immunology*, 117(1), 84–91.
- [178] Zhang W., Zhou M., Chen C., Wu S., Wang L., Xia B., Liu J., Ma X., Pan T., Zhang H., Li L., Liu B., (n.d.), “Identification of CD98 as a Novel Biomarker for HIV-1 Permissiveness and Latent Infection”, *mBio*, 13(6), e02496–22.
- [179] Wolthers K. C., Otto S. A., Lens S. M., Kolbach D. N., van Lier R. A., Miedema F., Meyaard L., (1996), “Increased expression of CD80, CD86 and CD70 on T cells from HIV-infected individuals upon activation in vitro: regulation by CD4+ T cells”, *European Journal of Immunology*, 26(8), 1700–1706.
- [180] Leng Q., Bentwich Z., Magen E., Kalinkovich A., Borkow G., (2002), “CTLA-4 upregulation during HIV infection: association with anergy and possible target for therapeutic intervention”, *AIDS (London, England)*, 16(4), 519–529.
- [181] Neri F., Giolo G., Potestà M., Petrini S., Doria M., (2011), “CD4 downregulation by the human immunodeficiency virus type 1 Nef protein is dispensable for optimal output and functionality of viral particles in primary T cells”, *The Journal of General Virology*, 92(Pt 1), 141–150.
- [182] Sugden S. M., Pham T. N. Q., Cohen E. A., (2017), “HIV-1 Vpu Downmodulates ICAM-1 Expression, Resulting in Decreased Killing of Infected CD4+ T Cells by NK Cells”, *Journal of Virology*, 91(8), e02442–16.
- [183] Kutscher S., Allgayer S., Dembek C. J., Bogner J. R., Protzer U., Goebel F. D., Erfle V., Cosma A., (2010), “MVA-nef induces HIV-1-specific polyfunctional and proliferative T-cell responses revealed by the combination of short- and long-term immune assays”, *Gene Therapy*, 17(11), 1372–1383.
- [184] Sukegawa S., Miyagi E., Bouamr F., Farkašová H., Strebel K., (2018), “Mannose Receptor 1 Restricts HIV Particle Release from Infected Macrophages”, *Cell reports*, 22(3), 786–795.
- [185] Hayasaka H., Kobayashi D., Yoshimura H., Nakayama E. E., Shioda T., Miyasaka M., (2015), “The HIV-1 Gp120/CXCR4 axis promotes CCR7 ligand-dependent CD4 T cell migration: CCR7 homo- and CCR7/CXCR4 hetero-oligomer formation as a possible mechanism for up-regulation of functional CCR7”, *PloS One*, 10(2), e0117454.

ÖZGEÇMİŞ

Mehmet Burak KOCA ilk ve orta dereceli okulları Erzincan'da tamamladıktan sonra Yalova Üniversitesi Bilgisayar Mühendisliği Bölümünde lisans eğitimine başladı. 2015 Yılında Bilgisayar Mühendisliği Bölümünden lisans derecesini alarak Bilgisayar Mühendisi unvanını kazandı. 2016 ve 2017 yıllarında çeşitli yazılım firmalarında stajyer ve yazılımcı olarak çalıştı. 2018 yılı Şubat ayında Gebze Teknik Üniversitesi Fen Bilimleri Enstitüsü Bilgisayar Mühendisliği Anabilim Dalında başladığı yüksek lisans eğitimini Ocak 2020'de tamamladı ve aynı alanda doktora eğitimine başladı. Yüksek lisans ve doktora eğitimi boyunca çeşitli biyolojik problemlere çizge ve yapay zeka tabanlı çözüm yöntemleri geliştirdi. Doktora sırasında takım kaptanı olarak katıldığı 2023 TEKNOFEST Sağlıkta Yapay Zeka yarışmasında birincilik ve en iyi sunum ödülünü kazandı. Doktora tezinin bir kısmını oluşturan çalışmalar 1002-A programı kapsamında TÜBİTAK tarafından araştırma projesi olarak desteklendi. 2018 yılı Şubat ayından beri Gebze Teknik Üniversitesi Bilgisayar Mühendisliği Bölümünde araştırma görevlisi olarak çalışmaktadır.

TEZ ÇALIŞMASI KAPSAMINDA YAPILAN YAYINLAR

Koca, M. B., Sevilgen, F. E., (2024), "Integration of single-cell proteomic datasets through distinctive proteins in cell clusters", *Proteomics*, 24(7), 2300282.

Koca, M. B., Sevilgen, F. E., (2024), "Explainable Graph Learning for Multimodal Single-Cell Data Integration", *bioRxiv*, 2024.12.06.627151.

Koca, M.B., Sevilgen, F. E., (2024), "Comparative Analysis of Fusion Techniques for Integrating Single-cell Multiomics Datasets", in "32st Signal Processing and Communications Applications Conference (SIU)", IEEE, Tarsus, Türkiye.

Koca, M. B., Sevilgen, F. E., (2024), "Multi-omic Single-cell Data Analysis for Biomarker Identification in Viral Pathogenesis", in "17th The International Symposium on Health Informatics and Bioinformatics", İstanbul, Türkiye.

Koca, M.B., Sevilgen, F. E., (2023), "Alignment of Single-cell Datasets with the Hierarchical Cell Matching Algorithm", in "31st Signal Processing and Communications Applications Conference (SIU)", IEEE, İstanbul, Türkiye.

Koca, M. B., Sevilgen, F. E., (2023), "Horizontal Integration of Single-cell Proteomics Datasets", in "9. Turkish Medical World Congress PhD Consortium on Artificial Intelligent in Health", İstanbul, Türkiye.

EKLER

Ek-A Çizelgeler

Çizelge A.1: Çoklu-tür tek-hücre proteomik veri kümelerindeki protein listeleri.

Protein ID	İnsan (H1N1)	İnsan (IFN γ)	Rhesus	Cyno
CD45	Mevcut	Mevcut	Mevcut	Mevcut
STAT3	Mevcut	Mevcut	Mevcut	Mevcut
CD61	Mevcut	Mevcut	Mevcut	Mevcut
CD7	Mevcut	Mevcut	Mevcut	Mevcut
CD33	Mevcut	Mevcut	Mevcut	Mevcut
CD11c	Mevcut	Mevcut	Mevcut	Mevcut
MAPKAPK2	Mevcut	Mevcut	Mevcut	Mevcut
CREB	Mevcut	Mevcut	Mevcut	Mevcut
PLCg2	Mevcut	Mevcut	Mevcut	Mevcut
STAT1	Mevcut	Mevcut	Mevcut	Mevcut
CD123	Mevcut	Mevcut	Mevcut	Mevcut
STAT5	Mevcut	Mevcut	Mevcut	Mevcut
CD14	Mevcut	Mevcut	Mevcut	Mevcut
Erk1.2	Mevcut	Mevcut	Mevcut	Mevcut
CD11b	Mevcut	Mevcut	Mevcut	Mevcut
p38	Mevcut	Mevcut	Mevcut	Mevcut
CD8	Mevcut	Mevcut	Mevcut	Mevcut
CD4	Mevcut	Mevcut	Mevcut	Mevcut
CD3	Mevcut	Mevcut	Mevcut	Mevcut
CD66	Mevcut	Mevcut	Mevcut	Mevcut
CD16	Mevcut	Mevcut	Mevcut	Mevcut
Zap70.Syk	Mevcut	Mevcut	Mevcut	Mevcut
TBK1	Mevcut	Mevcut	Mevcut	Mevcut
CD1c	Mevcut	Mevcut	Mevcut	Mevcut
I κ Ba	Mevcut	Mevcut	Mevcut	Mevcut
BDCA3	Mevcut	Mevcut	Mevcut	Mevcut
STAT6	Mevcut	Mevcut	Mevcut	Mevcut
CD45RA	Mevcut	Mevcut	Mevcut	Mevcut
Fox.P3	Mevcut	Mevcut	Mevcut	Mevcut
CD161	Mevcut	Mevcut	Mevcut	Mevcut
Ki67	Mevcut	Mevcut	Mevcut	Mevcut
STAT4	Mevcut	Mevcut	Mevcut	Mevcut

Protein ID	İnsan (H1N1)	İnsan (IFN γ)	Rhesus	Cyno
CCR7	Mevcut	Mevcut	Mevcut	Mevcut
X4EBP1	Mevcut	Mevcut	Mevcut	Mevcut
CD19	Mevcut	Mevcut	Mevcut	Mevcut
IgM	Mevcut	Mevcut	Mevcut	Mevcut
CD56	Mevcut	Mevcut	Mevcut	Mevcut
HLA.DR	Mevcut	Mevcut	Mevcut	Mevcut
CD235a	Mevcut	Mevcut	Yok	Mevcut
CD27	Mevcut	Yok	Yok	Yok

Çizelge A.2: Farklı uzaklık metrikleri ile hücre eşleştirme başarımları.

Gün	Hasta	Kosinüs	Korelasyon	Minkowski	Öklid	Manhattan
0	1	0.7596	0.7785	0.7691	0.7691	0.4956
0	2	0.8188	0.7996	0.7390	0.7390	0.3777
0	3	0.8710	0.8574	0.8312	0.8312	0.7952
0	4	0.8302	0.8231	0.8038	0.8038	0.5033
0	5	0.8865	0.8735	0.8179	0.8179	0.7207
0	6	0.9179	0.9063	0.8418	0.8418	0.8680
0	7	0.8962	0.8878	0.8312	0.8312	0.7327
0	8	0.8543	0.8472	0.7871	0.7871	0.7013
3	1	0.8153	0.8051	0.7896	0.7896	0.6224
3	2	0.7937	0.7807	0.7268	0.7268	0.4142
3	3	0.8786	0.8750	0.8633	0.8633	0.8303
3	4	0.8093	0.8049	0.7869	0.7869	0.6358
3	5	0.8939	0.8878	0.8396	0.8396	0.6259
3	6	0.8742	0.8579	0.8523	0.8523	0.5193
3	7	0.8783	0.8701	0.8376	0.8376	0.6583
3	8	0.8156	0.8041	0.7528	0.7528	0.5277
7	1	0.8583	0.8513	0.8336	0.8336	0.7130
7	2	0.7961	0.7770	0.7071	0.7071	0.6113
7	3	0.8478	0.8397	0.8192	0.8192	0.6266
7	4	0.8282	0.8106	0.7853	0.7853	0.4932
7	5	0.8810	0.8637	0.8167	0.8167	0.4879
7	6	0.9383	0.9308	0.8875	0.8875	0.7331
7	7	0.8979	0.8917	0.8656	0.8656	0.7594
7	8	0.8052	0.7920	0.7545	0.7545	0.5019
Ortalama		0.8519	0.8423	0.8058	0.8058	0.6242

Çizelge A.3: Patojen Hedef proteinleri ve beklenen değişim yönleri.

Proteinler	Değişim Yönü	Referans
P51681	Aşağı	[154]
P10747	Aşağı	[155]
P26842	Aşağı	[156]
P13612	Yukarı	[157]
P07766	Yukarı	[158]
P25445	Aşağı	[159]
P20023	Aşağı	[160]
P08962	Yukarı	[161]
P01903	Aşağı	[162]
P27487	Aşağı	[163]
P05106	Aşağı	[164]
P02786	Aşağı	[165]
P43629	Aşağı	[166]
P06729	Yukarı	[167]
P13987	Yukarı	[168]
P05107	Yukarı	[169]
P05556	Yukarı	[170]
P28907	Yukarı	[171]
P15813	Aşağı	[172]
P05362	Yukarı	[173]
P11279	Aşağı	[174]
P51679	Aşağı	[175]
P60033	Yukarı	[176]
P42081	Aşağı	[177]
P08195	Yukarı	[178]
P33681	Yukarı	[179]
P46531	Aşağı	[156]
P16410	Yukarı	[180]
P01730	Aşağı	[181]
P13598	Aşağı	[182]
P29965	Yukarı	[183]
P22897	Aşağı	[184]
P61073	Yukarı	[185]

Çizelge A.4: Sitotoksik CD4 T altkütmesi için ayırtedici genlerin artan önem sıralı listeleri. En önemli 5 gen koyu olarak verilmiştir.

CD4 hücreleri arası	Tüm hücreler arası
ENSG00000150687	ENSG00000069667
ENSG00000166710	ENSG00000197540
ENSG00000227507	ENSG00000163421
ENSG00000163421	ENSG00000198668
ENSG00000115956	ENSG00000054654
ENSG00000171476	ENSG00000235162
ENSG00000227191	ENSG00000115523
ENSG00000142634	ENSG00000160654
ENSG00000205336	ENSG00000171476
ENSG00000145649	ENSG00000077984
ENSG00000116667	ENSG00000008517
ENSG00000271503	ENSG00000166710
ENSG00000169554	ENSG00000185697
ENSG00000180644	ENSG00000145649
ENSG00000100453	ENSG00000137441
ENSG00000077984	ENSG00000227191
ENSG00000115523	ENSG00000271503
ENSG00000105374	ENSG00000105374
ENSG00000137441	ENSG00000225783
ENSG00000100450	ENSG00000100450