**TC.**
**HACETTEPE ÜNİVERSİTESİ**
**SAĞLIK BİLİMLERİ ENSTİTÜSÜ**

# HETEROJEN BİYOMEDİKAL VERİNİN BİLGİ ÇİZGELERİ VE DERİN ÖĞRENME TABANLI ANALİZİ İLE PROTEİN FONKSİYONLARININ OTOMATİK TAHMİNİ

**Erva ULUSOY**

**Biyoinformatik Programı**
**YÜKSEK LİSANS TEZİ**

**ANKARA**
**2023**

**TC.**
**HACETTEPE ÜNİVERSİTESİ**
**SAĞLIK BİLİMLERİ ENSTİTÜSÜ**

# HETEROJEN BİYOMEDİKAL VERİNİN BİLGİ ÇİZGELERİ VE DERİN ÖĞRENME TABANLI ANALİZİ İLE PROTEİN FONKSİYONLARININ OTOMATİK TAHMİNİ

**Erva ULUSOY**

**Biyoinformatik Programı**
**YÜKSEK LİSANS TEZİ**

**TEZ DANIŞMANI**
**Doç. Dr. Tunca DOĞAN**

**ANKARA**
**2023**

**HACETTEPE ÜNİVERSİTESİ**
**SAĞLIK BİLİMLERİ ENSTİTÜSÜ**

# HETEROJEN BİYOMEDİKAL VERİNİN BİLGİ ÇİZGELERİ VE DERİN ÖĞRENME TABANLI ANALİZİ İLE PROTEİN FONKSİYONLARININ OTOMATİK TAHMİNİ

**Öğrenci: Erva Ulusoy**
**Danışman: Doç. Dr. Tunca Doğan**

Bu tez çalışması 08.06.2023 tarihinde jürimiz tarafından "Biyoinformatik Programı"nda yüksek lisans tezi olarak kabul edilmiştir.

| | |
|---|---|
| **Jüri Başkanı:** | *Dr. Öğr. Üyesi İdil Yet*<br>*(Hacettepe Üniversitesi)* |
| **Tez Danışmanı:** | *Doç. Dr. Tunca Doğan*<br>*(Hacettepe Üniversitesi)* |
| **Üye:** | *Dr.Öğr.Üyesi Aybar Can Acar*<br>*(Orta Doğu Teknik Üniversitesi)* |

Bu tez Hacettepe Üniversitesi Lisansüstü Eğitim-Öğretim ve Sınav Yönetmeliğinin ilgili maddeleri uyarınca yukarıdaki jüri tarafından uygun bulunmuştur.

*Prof. Dr. Müge YEMİŞÇİ ÖZKAN*
**Enstitü Müdürü**

## YAYIMLAMA VE FİKRİ MÜLKİYET HAKLARI BEYANI

Enstitü tarafından onaylanan lisansüstü tezimin/raporumun tamamını veya herhangi bir kısmını, basılı (kağıt) ve elektronik formatta arşivleme ve aşağıda verilen koşullarla kullanıma açma iznini Hacettepe Üniversitesine verdiğimi bildiririm. Bu izinle Üniversiteye verilen kullanım hakları dışındaki tüm fikri mülkiyet haklarım bende kalacak, tezimin tamamının ya da bir bölümünün gelecekteki çalışmalarda (makale, kitap, lisans ve patent vb.) kullanım hakları bana ait olacaktır.
Tezin kendi orijinal çalışmam olduğunu, başkalarının haklarını ihlal etmediğimi ve tezimin tek yetkili sahibi olduğumu beyan ve taahhüt ederim. Tezimde yer alan telif hakkı bulunan ve sahiplerinden yazılı izin alınarak kullanılması zorunlu metinlerin yazılı izin alınarak kullandığımı ve istenildiğinde suretlerini Üniversiteye teslim etmeyi taahhüt ederim.

Yükseköğretim Kurulu tarafından yayınlanan ***"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*** kapsamında tezim aşağıda belirtilen koşullar haricince YÖK Ulusal Tez Merkezi / H.Ü. Kütüphaneleri Açık Erişim Sisteminde erişime açılır.

- Enstitü / Fakülte yönetim kurulu kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren 2 yıl ertelenmiştir. (1)
- Enstitü / Fakülte yönetim kurulunun gerekçeli kararı ile tezimin erişime açılması mezuniyet tarihimden itibaren ... ay ertelenmiştir. (2)
- Tezimle ilgili gizlilik kararı verilmiştir. (3)

…… /……/……

Erva Ulusoy

---

[1] *"Lisansüstü Tezlerin Elektronik Ortamda Toplanması, Düzenlenmesi ve Erişime Açılmasına İlişkin Yönerge"*

*(1) Madde 6. 1. Lisansüstü tezle ilgili patent başvurusu yapılması veya patent alma sürecinin devam etmesi durumunda, tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulu** iki yıl süre ile tezin erişime açılmasının ertelenmesine karar verebilir.*

*(2) Madde 6. 2. Yeni teknik, materyal ve metotların kullanıldığı, henüz makaleye dönüşmemiş veya patent gibi yöntemlerle korunmamış ve internetten paylaşılması durumunda 3. şahıslara veya kurumlara haksız kazanç imkanı oluşturabilecek bilgi ve bulguları içeren tezler hakkında tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulunun** gerekçeli kararı ile altı ayı aşmamak üzere tezin erişime açılması engellenebilir.*

*(3) Madde 7. 1. Ulusal çıkarları veya güvenliği ilgilendiren, emniyet, istihbarat, savunma ve güvenlik, sağlık vb. konulara ilişkin lisansüstü tezlerle ilgili gizlilik kararı, **tezin yapıldığı kurum** tarafından verilir *. Kurum ve kuruluşlarla yapılan işbirliği protokolü çerçevesinde hazırlanan lisansüstü tezlere ilişkin gizlilik kararı ise, **ilgili kurum ve kuruluşun önerisi** ile **enstitü** veya **fakültenin** uygun görüşü üzerine **üniversite yönetim kurulu** tarafından verilir. Gizlilik kararı verilen tezler Yükseköğretim Kuruluna bildirilir.*
*Madde 7.2. Gizlilik kararı verilen tezler gizlilik süresince enstitü veya fakülte tarafından gizlilik kuralları çerçevesinde muhafaza edilir, gizlilik kararının kaldırılması halinde Tez Otomasyon Sistemine yüklenir.*

***** Tez **danışmanının** önerisi ve **enstitü anabilim dalının** uygun görüşü üzerine **enstitü** veya **fakülte yönetim kurulu tarafından karar verilir.***

# ETİK BEYAN

Bu çalışmadaki bütün bilgi ve belgeleri akademik kurallar çerçevesinde elde ettiğimi, görsel, işitsel ve yazılı tüm bilgi ve sonuçları bilimsel ahlak kurallarına uygun olarak sunduğumu, kullandığım verilerde herhangi bir tahrifat yapmadığımı, yararlandığım kaynaklara bilimsel normlara uygun olarak atıfta bulunduğumu, tezimin kaynak gösterilen durumlar dışında özgün olduğunu, Doç. Dr. Tunca DOĞAN danışmanlığında tarafımdan üretildiğini ve Hacettepe Üniversitesi Sağlık Bilimleri Enstitüsü Tez Yazım Yönergesine göre yazıldığını beyan ederim.

Erva ULUSOY

# TEŞEKKÜR

Yüksek lisans eğitimim süresince değerli rehberliği ve desteği ile biyoinformatik alanındaki gelişimime büyük katkıda bulunan, daha iyi bir araştırmacı olmak için bana ilham kaynağı ve rol modeli olan danışman hocam Doç. Dr. Tunca Doğan'a,

Bilgi birikimi, deneyimi ve değerli fikirleri ile araştırmalarıma katkı sağlayan Biyoinformatik Anabilim Dalı'nın tüm öğretim üyelerine,

Bu tez çalışması için yaptıkları yapıcı eleştiri ve önerileri ile yeni bakış açıları sunan değerli hocalarım Dr. Öğr. Üyesi İdil Yet ve Dr.Öğr.Üyesi Aybar Can Acar'a,

Tüm eğitim sürecim boyunca sonsuz sabır, inanç ve teşvikleri ile beni her zaman bir adım sonrası için cesaretlendiren aileme,

Teşekkürlerimi sunarım.

Bu tez çalışması "TÜBİTAK - BİDEB 2210-A Genel Yurt İçi Yüksek Lisans Burs Programı" ve "TÜBİTAK - ARDEB 3501 - Kariyer Geliştirme Programı" tarafından desteklenmiştir.

# ÖZET

**ULUSOY E., Heterojen Biyomedikal Verinin Bilgi Çizgeleri ve Derin Öğrenme Tabanlı Analizi ile Protein Fonksiyonlarının Otomatik Tahmini, Hacettepe Üniversitesi; Sağlık Bilimleri Enstitüsü, Biyoinformatik Programı Yüksek Lisans Tezi, ANKARA, 2023.** Proteinlerin hücresel süreçlerdeki rollerinin belirlenmesi, kompleks biyolojik mekanizmaların tam olarak anlaşılması için büyük öneme sahiptir. Pahalı ve zaman alıcı deneysel yöntemlere alternatif olarak geliştirilen fonksiyon tahmini yöntemleri, biyolojik veritabanlarındaki herkese açık veri setlerinden yararlanmaktadır. Mevcut yöntemlerin genellikle tek bir veri türüne dayalı olması, proteinlerin çok yönlü fonksiyonel yapısını yakalama yeteneğini ve tahmin performansını sınırlamaktadır. Geometrik derin öğrenme yöntemlerindeki son gelişmeler, farklı kaynaklardaki çeşitli biyolojik bileşenleri ve ilişkilerini entegre eden heterojen çizgeleri kullanarak bu probleme çözüm olabilecek yeni algoritmalar sunmuştur. Bu tez çalışmasında heterojen çizge bazlı bir derin öğrenme yaklaşımı ve Gene Ontology (GO) tabanlı geniş çaplı protein fonksiyon tahminindeki uygulaması önerilmiştir. Bunun için öncelikle 14 farklı biyomedikal kaynaktan alınan veri kapsamlı bir heterojen bilgi çizgesi olarak entegre edilmiştir. Bu veri seti, çizge sinir ağları (heterojen çizge dönüştürücü mimarisi) ile tahmin modellerinin eğitiminde kullanılmıştır. Karşılaştırma veri setleri üzerinden yapılan performans değerlendirmesi, tüm GO kategorilerinde temel tahmin metodlarına kıyasla yüksek, son teknoloji tahmin modellerine kıyasla karşılaştırılabilir sonuçlara ulaşıldığını göstermiştir. Yüksek bilgi içerikli moleküler fonksiyon terimlerinin tahmininde önerilen model en başarılı üç yöntem arasında yer almıştır. Seçili proteinlere ait fonksiyon tahminlerinin biyolojik anlamlılığını araştıran literatür taramasında, hakkında kısıtlı bilgi bulunan yeni fonksiyonel ilişkilerin tahmin edilebildiği görülmüştür. Bu çalışma, son derece heterojen biyomedikal veri ile geometrik derin öğrenmenin protein fonksiyon tahmininde kullanımını araştırarak literatüre katkıda bulunmaktadır.

**Anahtar kelimeler:** protein fonksiyon tahmini, çizge tabanlı derin öğrenme, biyomedikal bilgi çizgeleri, gen ontolojisi, CAFA yarışması

## ABSTRACT

**ULUSOY E., Automated Prediction of Protein Functions with Knowledge Graph Representations and Deep Learning-Based Analysis of Heterogeneous Biomedical Data, Hacettepe University, Graduate School Health Sciences, Bioinformatics Program Master Degree Thesis, ANKARA, 2023.** Proteins are vital for cellular processes, and accurately determining their functions is crucial for understanding complex biological mechanisms. Computational approaches have emerged as alternatives to expensive and time-consuming experimental methods, leveraging publicly available data in biomedical databases to predict protein functions. However, existing methods often rely on a single data type, limiting their ability to capture the multifaceted functional complexity of proteins. Geometric deep learning offer new algorithms that can be utilized to address these issues by integrating diverse biological entities and relationships sourced from multiple databases using heterogeneous graphs. In this thesis study, we propose a heterogeneous graph learning approach and its implementation as a computational method for Gene Ontology (GO) based large-scale protein function prediction. For this, we first constructed a comprehensive biological knowledge graph by obtaining and integrating data from 14 different biomedical databases. Using this dataset, we trained function prediction models using graph neural networks, i.e., the heterogeneous graph transformer architecture. Performance evaluation on benchmark datasets indicated superior performance compared to baseline methods across all GO categories, while achieving comparable results to top predictors. Our model demonstrated excellent performance in predicting high-information-content molecular function terms, ranking among the top three models. To assess the biological relevance of predicted functional relationships, we conducted a use-case study for selected proteins, showcasing our approach's ability to identify unknown functions with limited available information. This study contributes to the existing literature by investigating protein function prediction using geometric deep learning on highly heterogeneous biomedical data.

**Keywords:** protein function prediction, graph-based deep learning, biomedical knowledge graphs, gene ontology, CAFA challenge

# İÇİNDEKİLER

# SİMGELER VE KISALTMALAR

| | |
|---|---|
| **AUPR** | Area Under the Precision-Recall Curve |
| **BPO** | Biological Process Ontology |
| **CAFA** | Critical Assessment of Functional Annotation |
| **CCO** | Cellular Component Ontology |
| **DAG** | Directed Acyclic Graph |
| **GAT** | Graph Attention Network |
| **GCN** | Graph Convolutional Network |
| **GNN** | Graph Neural Network |
| **GO** | Gene Ontology |
| **GOA** | Gene Ontology Annotation |
| **HGT** | Heterogeneous Graph Transformer |
| **MCC** | Matthews Correlation Coefficient |
| **MFO** | Molecular Function Ontology |

# ŞEKİLLER

# TABLOLAR

**Tablo** **Sayfa**

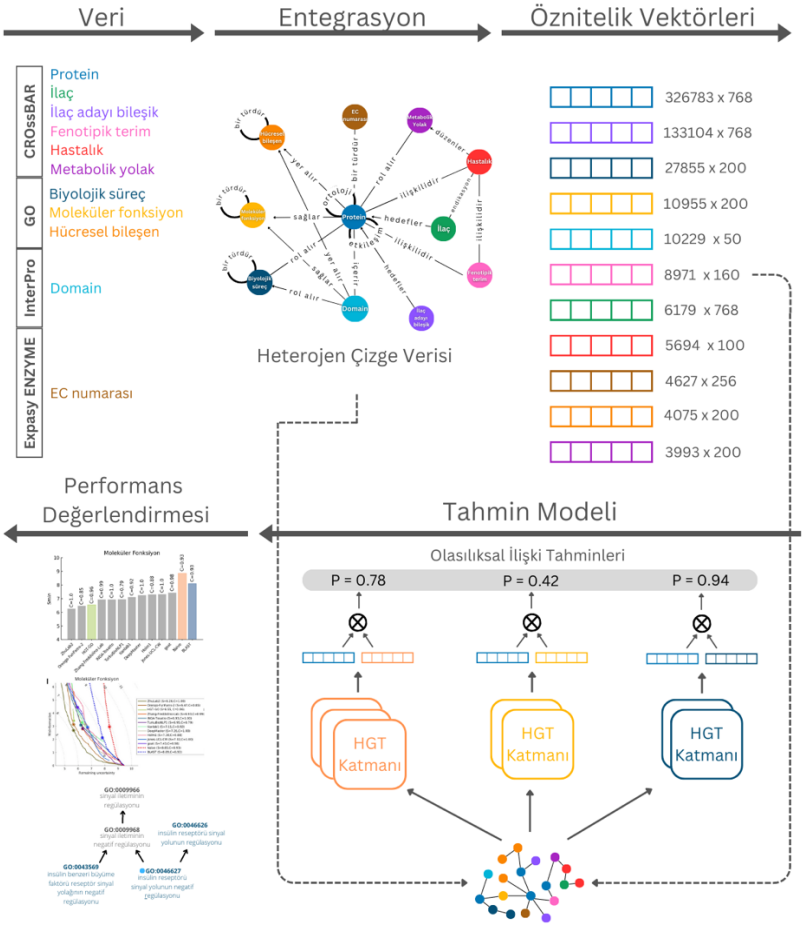
# 1. GİRİŞ

Proteinler, canlı organizmalarda yaşamsal fonksiyonların ardındaki birçok hücresel süreçte çeşitli görevler üstlenmektedir. Proteinlerin fonksiyonlarının tespit edilmesi, biyolojik sistemlerin karmaşık mekanizmalarının anlaşılmasına olanak sağlamakta ve ilaç keşfi, yeni tedavi yöntemleri ve biyoteknolojik uygulamaların geliştirilmesi gibi önemli bilimsel araştırmalara ışık tutmaktadır. Deneysel yöntemler, yüksek maliyetleri ve zaman alıcı olmaları nedeniyle geniş çaplı fonksiyon tespiti için yetersiz kalmaktadır. Bu durum araştırmacıları biyomedikal veri tabanlarında herkese açık olarak paylaşılan verilerle yeni bulgular üretebilen hesaplamalı yöntemler geliştirmeye yöneltmiştir. Proteinlerin kompleks organizmalarda fonksiyonlarını diğer biyokimyasal bileşenlerle etkileşim halinde gerçekleştirmesi, karmaşık hücresel mekanizmaları ve hastalıkların temel süreçlerini tam olarak anlamak için bu etkileşimlerin incelenmesini gerektirmektedir. Hesaplamalı yöntemlerde kullanılacak verinin eldesi için başvurulan biyomedikal veri tabanlarının çoğu spesifik bir veri türüne odaklanmıştır. Proteinlerin karmaşık fonksiyonel yapısı göz önüne alındığında, bu veri türlerinden yalnızca birini kullanarak yüksek doğrulukta fonksiyon tahmini yapmak mümkün değildir. Son yıllarda, birden fazla biyolojik bileşen ve ilişki tipini içeren heterojen verinin çizge yapısı kullanarak modellenmesi ve bu veriyi girdi olarak kullanan çizge tabanlı hesaplamalı yaklaşımların yaygınlaşması, çeşitli biyokimyasal bileşenler arasındaki yeni ilişkilerin keşfini hızlandırmıştır. Protein fonksiyon tahmini alanında geliştirilen çizge tabanlı yöntemlerde ise genellikle homojen ilişkisel veri setlerinden yararlanılmıştır. Heterojen çizge verisi kullanılarak geliştirilecek çizge tabanlı yeni yaklaşımlar, fonksiyonel bilginin daha bütünsel olarak yorumlanması ve bilinmeyen ilişkilerin ortaya çıkarılmasına katkı sağlayacaktır.

Bu ihtiyaçlar doğrultusunda bu tez çalışması kapsamında proteinler ve diğer biyokimyasal bileşenlerin birbirleri ile olan ilişkilerini kapsayan heterojen veri üzerinden çizge tabanlı derin öğrenme yaklaşımı ile yüksek performanslı ve geniş çaplı fonksiyon tahmini gerçekleştirmek amaçlanmıştır.

Bu amaca yönelik olarak çalışmada kamuya açık kapsamlı ve çok bilinen 14 farklı biyomedikal veri tabanından elde edilen heterojen veri birbiri ile ilişkilendirilip,

9 düğüm tipi ve 17 kenar tipinden oluşan kapsamlı bir bilgi çizgesi olarak modellenmiştir. Bu veri kullanılarak heterojen çizge dönüştürücü mimarisine dayalı protein-fonksiyon ilişki tahmini modelleri eğitilmiştir. Eğitilen modellerin performans sonuçları protein-fonksiyon tahmini alanında sıkça kullanılan karşılaştırma veri setleri üzerinden hesaplanıp diğer fonksiyon tahmini metodları ile karşılaştırılmıştır. Heterojen veride yer alan düğüm tipleri ve bunların içinde bulundukları kenar türlerinin öğrenmeye katkısının gözlemlenebilmesi için ablasyon modelleri eğitilip performans sonuçları araştırılmıştır. Modelin seçili proteinlerle ilişkilendirdiği fonksiyonel terim tahminlerinin biyolojik anlamlılığının incelenmesi için bir literatür taraması yürütülmüştür. Çalışma iş akışı Şekil 1.1.'de gösterilmektedir.




**Şekil 1.1.** Çalışma iş akışı.

Genel bilgiler bölümünde proteinlerin fonksiyonel anotasyonu, fonksiyon tahmini ve biyomedikal veri analizinde sıkça kullanılan çizge tabanlı sinir ağı mimarileri ile ilgili temel kavramlara değinilmiştir. Bunun yanı sıra bu kısımda

literatürdeki öznitelik ve çizge tabanlı fonksiyon tahmini çalışmalarına yer verilmiştir. Gereç ve yöntemler bölümünde çalışmada kullanılan veri setinin eldesi, entegrasyonu ve ön işlemesi; tahmin modelinin tasarımı, eğitimi ve optimizasyonu; performans karşılaştırma veri seti ve değerlendirme metrikleri ile ilgili bilgiler verilmiştir. Eğitilen modellerin performans sonuçları ve biyolojik anlamlılığının araştırılması için seçilen örnek tahminler bulgular kısmında verilmiştir. Bu kısımda verilen sonuçlar doğrultusunda performansın değerlendirilmesi, benzer yaklaşımlarla karşılaştırılması ve seçili tahminlerin biyolojik anlamlılığı ile ilgili detaylı incelemeler tartışma kısmında yer almaktadır. Sonuçlar ve öneriler bölümlerinde tez çalışmasının amaçları göz önünde bulundurularak elde edilen sonuçlar yorumlanmış ve bu yorumlar üzerinden fonksiyon tahmini alanında gerçekleştirilebilecek gelecek çalışmalar için önerilerde bulunulmuştur.

# 2. GENEL BİLGİLER

## 2.1. Proteinlerin Fonksiyonel Anotasyonu

Canlı organizmalardaki yaşamsal fonksiyonların yerine getirilmesinde rol oynayan temel bileşenler, bu organizmalarda sentezlenen gen ürünleridir. Yaşamın yapı taşı olarak tanımlanan proteinleri oluşturan hiyerarşik yapısal düzen, geniş yapısal ve fonksiyonel çeşitliliğe ulaşmalarını sağlar. Birincil yapıyı oluşturan doğrusal amino asit zincirleri, proteinlerin genel kimyasal özelliklerinin ve daha üst seviyedeki yapısal özelliklerinin belirlenmesinde etkilidir. İkincil yapıları alfa sarmalı ve beta yaprağı gibi lokal katlanma kalıplarını, üçüncül yapıları ise proteinlerin genel üç boyutlu düzenini ifade eder (1). Birden fazla protein ünitesinin bir araya gelip fonksiyonel bir kompleks oluşturması ile dördüncül yapı meydana gelir.

Proteinler biyokimyasal reaksiyonların katalizlenmesinden hücre ve doku yapılarının oluşturulmasına, moleküllerin taşınmasından gen ifadelerinin düzenlenmesine birçok hücresel süreçte görev alır. Proteinlerin fonksiyonel anotasyonu, deneysel veya hesaplamalı yöntemlerin kullanılması ile işlevlerinin belirlenmesi sürecidir. Fonksiyonel anotasyon sayesinde biyolojik sistemlerin altında yatan kompleks mekanizmaların incelenmesi mümkün hale getirilip, yeni ilaçların, tedavi yöntemlerinin ve biyoteknolojik uygulamaların geliştirilmesi gibi önemli bilimsel araştırmalara ışık tutulmaktadır. Kompleks organizmalarda biyolojik fonksiyonlar proteinler tarafından bağımsız olarak değil, diğer biyokimyasal bileşenlerle olan etkileşimleri sonucunda ortaya çıkmaktadır. Bu sebeple karmaşık biyokimyasal mekanizmaları ve kompleks hastalıkların ardında yatan süreçleri tam manasıyla anlayabilmek için bu etkileşimlerin de incelenmesi gerekmektedir.

Proteinlerin fonksiyonel anotasyonunda gen knockout deneyleri, hedefe yönelik mutasyonlar ve gen ekspresyonunun inhibisyonu gibi tek seferde bir protein ürününü hedef alan deneysel yöntemler uygulanmaktadır. Hassas ve güvenilir fonksiyonel bilgi sağlasa da, bu yöntemler yüksek maliyetleri ve zaman alıcı olmaları sebebi ile çok sayıda proteinin fonksiyonunun belirlenmesi için uygulanabilir değildir (2). Proteinlerin çok yönlü fonksiyonel yapısı ve diğer biyolojik bileşenlerle olan

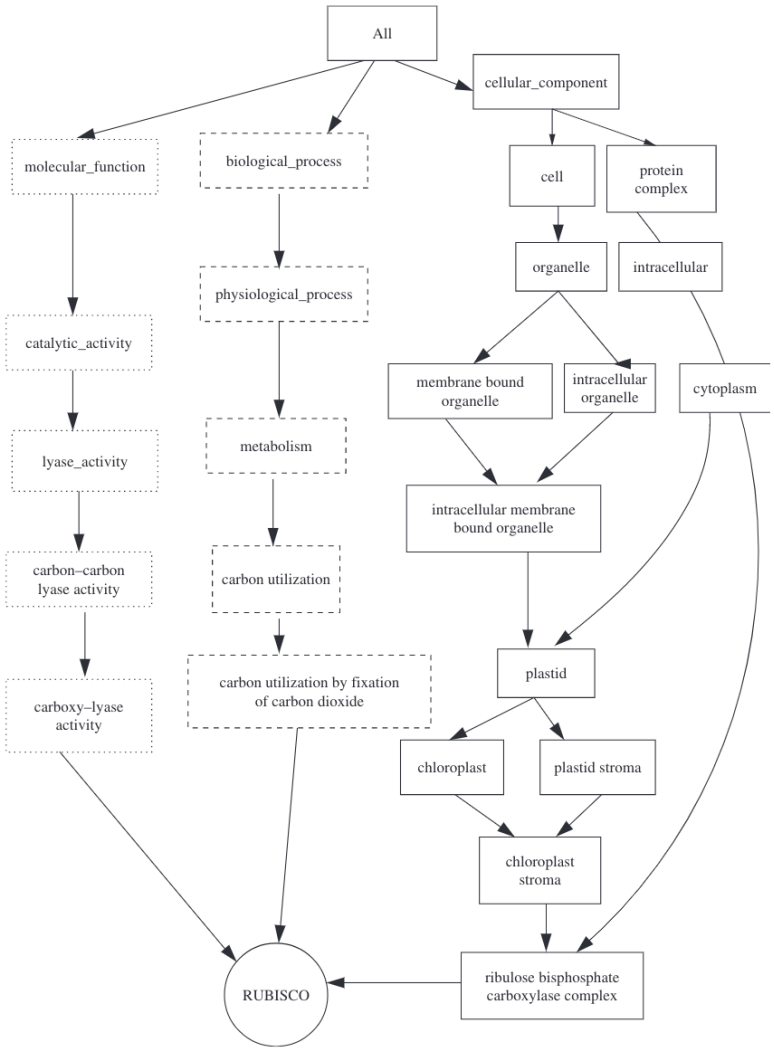
etkileşimleri sonucunda ortaya çıkan fonksiyonun büyük ölçüde değişikliklere uğraması da bu sürecin karmaşıklaşmasına yol açmaktadır. Özellikle yüksek-verimli sekanslama ve yapısal genomik yöntemlerindeki gelişmeler sayesinde sayısı hızla yükselen fonksiyonu bilinmeyen proteinler, geniş çaplı fonksiyonel anotasyon metodlarına olan ihtiyacı gün geçtikçe arttırmaktadır. Güncel olarak bilinen proteinlerin yalnızca yaklaşık %0.25'inin deneysel olarak anote edilmiş olması (3) bu ihtiyacın bir göstergesidir.

### 2.2. Protein Fonksiyon Tahmini

Araştırmacıların biyomedikal veri tabanları üzerinden herkese açık olarak paylaşılmış veriyi kullanarak geliştirdiği hesaplamalı yöntemler ile yeni bulgular üretmesi, fonksiyonu bilinmeyen proteinlerin hızlı ve geniş çaplı anotasyonu sürecini önemli ölçüde kolaylaştırmaktadır. Proteinlerin sekans, yapı, evrimsel ilişki ve diğer biyolojik bileşenlerle olan etkileşimleri gibi özellikleri üzerinden fonksiyonlarının tahmin edilmesini sağlayan bu algoritmalar, büyük genomik ve proteomik veri setlerinin yorumlanmasını mümkün kılarak yeni deneysel ve hesaplamalı yöntemlerin geliştirilmesine ve biyoteknoloji, farmakoloji ve kişiselleştirilmiş tıp gibi alanlardaki ilerlemelere katkıda bulunmaktadır.

Gene Ontology (GO), gen ve proteinlerin işlevlerini tanımlayan terimlerin standart ve hiyerarşik bir formatta saklanması ve araştırmacılara sunulması için yaygın olarak kullanılan bir kaynaktır (4). GO, fonksiyonel terimleri biyolojik süreç ontolojisi (*biological process ontology*, BPO), moleküler fonksiyon ontolojisi (*molecular function ontology*, MFO) ve hücresel bileşen ontolojisi (*cellular component ontology*, CCO) olmak üzere üç farklı alt kategori altında toplar. Moleküler fonksiyon ontolojisi, "kataliz" gibi moleküler düzeydeki aktiviteleri ifade eden terimleri içerir. Biyolojik süreç ontolojisi biyolojik yolaklar gibi birden fazla moleküler fonksiyonu içinde barındıran daha geniş işlevleri ifade eder. Hücresel bileşen ontolojisi ise proteinin hücre içinde fonksiyonunu gerçekleştirdiği lokasyon hakkında bilgi vermektedir. Bu alt ontolojiler içinde kategorize edilen terimler, araştırmacıların protein işlevlerini sistematik olarak anote etmelerine ve biyolojik süreçlerdeki rollerini tahmin etmelerine olanak tanır.

GO terimleri, bir yönlü asiklik çizge (*directed acyclic graph*, DAG) içerisinde birbiri ile hiyerarşik ilişkiler içerisinde organize edilmiş halde bulunmaktadır (Şekil 2.1.). DAG, üç alt GO kategorisini ifade eden üç ana daldan meydana gelmektedir. Her bir dal, proteinlerin moleküler işlevinin, dahil olduğu biyolojik süreçlerin veya hücre içindeki lokalizasyonunun belirli bir çeşidini ifade eden çok sayıda terim içerir. GO terimleri DAG içerisinde aralarında döngüsel bir bağlılık olmayan bir yapıda, genelden özele doğru sıralanmış halde, bir terimin birden fazla ebeveyn terimi olabilecek şekilde bulunurlar. Hesaplamalı yöntemler, birbiri ile ilişkili veya benzer proteinler arasında anotasyonların aktarılmasında GO terimleri arasındaki hiyerarşik ilişkiden de yararlanmaktadır.




**Şekil 2.1.** GO DAG yapısının ribulose-1,5-bisphosphatecarboxylase/oxygenase (RubisCO) enzimini anote eden GO

terimleri örneği ile gösterimi (5). Moleküler fonksiyon terimleri noktalı, biyolojik süreç terimleri çizgili, hücresel bileşen terimleri ise düz çizgi kenarlıklarla belirtilmiştir. DAG içerisinde yukarıdan aşağıya doğru indikçe terimlerin spesifikliği artmaktadır.

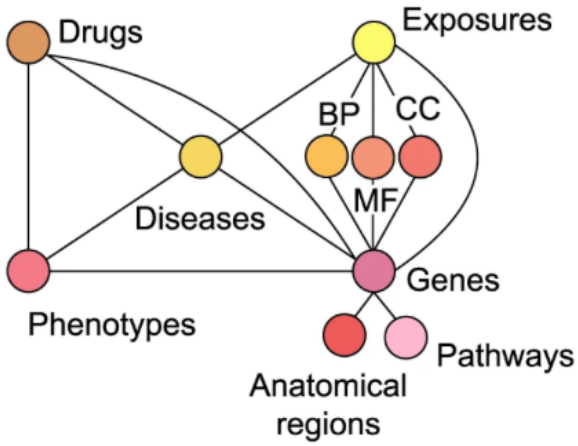
### 2.3. Biyolojik Veri Entegrasyonu ve Uygulamaları

Biyomedikal veri tabanlarında herkese açık olarak paylaşılan veri miktarı, gelişmekte olan teknoloji sayesinde gün geçtikçe artmaktadır. Bu kaynaklardan çoğunun odak ve kapsamı spesifik bir veri tipine yoğunlaşacak şekilde belirlenmiştir. Proteinlerin deneysel anotasyonları; proteinler, ilaçlar, ilaç adayı bileşikler, hastalıklar gibi diğer biyomedikal bileşenlerle olan ilişkileri; sekans ve yapı bilgileri, bu kaynaklarda saklanan ve protein fonksiyon tahmini için kullanılabilecek veri tiplerine örnek olarak verilebilir. Biyolojik sistemlerin kompleks yapısı göz önünde bulundurulduğunda bu veri tiplerinden tek birinden yararlanarak bu sistemlerin nasıl işlediğini tam olarak anlamak mümkün değildir. Araştırmacılar, biyolojik süreçleri daha bütünsel olarak yorumlayabilmek ve bilinmeyen ilişkileri ortaya çıkarabilmek için birden fazla veri kaynağının/tipinin birbiri ile entegre edilmesi ile elde edilen heterojen veriye ihtiyaç duymaktadır. Farklı veri kaynaklarında yer alan verinin üretilme, saklanma ve paylaşma şekillerinin çoğunlukla teknik olarak birbirinden farklı olması bu süreci zorlaştırmaktadır. Bu zorluklar, elde edilen heterojen ve yüksek boyutlu verinin işlenebilir hale getirilebilmesi için yüksek zaman ve iş gücü gerektiren veri entegrasyonu işlemlerini gerektirmektedir.

Bu iş yükünün azaltılması amacıyla gerçekleştirilen çalışmalardan bir kısmı farklı kaynaklarda mevcut aynı tip ve özellikteki homojen veriyi tek çatı altında toplamaktadır. Deneysel çalışmalar, literatür madenciliği, hesaplamalı tahminler ve diğer veri tabanları gibi farklı kaynaklardan toplanan biyolojik yolak verisini entegre ve kürate ederek kullanıcıya sunan Reactome (6) bu çalışmalara bir örnektir. STRING (7) ve BioGRID (8) veri tabanları da benzer şekilde farklı kaynaklardan elde ettikleri protein-protein etkileşim verisini entegre ederek, araştırmacılara bu etkileşim ağlarını analiz etmeleri için kapsamlı bir kaynak sunmaktadırlar. Bunlar dışında farklı biyolojik bileşenlere ait heterojen veriyi ilgili veri tabanlarından elde edip tek bir

kaynak altında toplamayı amaçlayan çalışmalar da mevcuttur. Heterojen ve kompleks yapıdaki biyolojik veriyi modelleyerek kullanıcıya anlaşılır bir görselleştirme ile sunma kapasitesine sahip olan bilgi çizgeleri, literatürde bu amaç için en çok tercih edilen yöntemlerden biridir (9–12).

Bilgi çizgeleri; heterojen veri noktalarını düğümler, bu veri noktaları arasındaki ilişkileri ise düğümleri birbirine bağlayan kenarlar şeklinde temsil edilmektedir. Biyolojik bilgi çizgelerinde düğümler biyolojik bileşenleri, kenarlar ise bu bileşenler arasındaki biyolojik ilişkileri ifade eder (Şekil 2.2.). Örneğin bir proteinin ilişkili olduğu fonksiyonel terim bilgisi, bilgi çizgesi içerisinde bu proteini ve fonksiyonel terimi ifade eden düğümler arasında bir kenar çizgisi olarak tanımlanır.



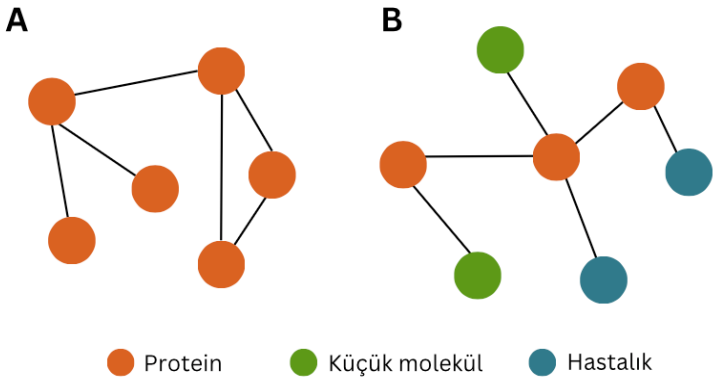

**Şekil 2.2.** 10 farklı düğüm tipi içeren bir biyolojik bilgi çizgesi örneği (10).

CROssBAR (13), entegre edilmiş heterojen biyolojik veriyi bilgi çizgesi yapısında sunan herkese açık sistemlerden biridir. Bu veri kaynağı, UniProt (proteinler) (3); Reactome ve KEGG (biyolojik yolaklar) (6,14); "DrugBank" (ilaçlar) (15); "ChEMBL" (bileşikler) (16); "OMIM", "Orphanet" ve "Experimental Factor Ontology - EFO" (hastalıklar) (17–19); ve "Human Phenotype Ontology - HPO" (fenotipler) (20) gibi çeşitli veri tabanlarından otomatik olarak toplanarak entegre edilmiş biyolojik bileşen koleksiyonlarını barındırmaktadır. Kullanıcılar, anlık sorgular ile ilgilendikleri biyolojik bileşen ve bu bileşenle yüksek ilişki seviyesine sahip diğer bileşenleri heterojen bir bilgi çizgesi yapısında çıktı alabilmektedir. Literatürdeki geniş çaplı biyolojik bilgi çizgeleri veri entegrasyonu ihtiyacını önemli ölçüde karşılasa da, bu koleksiyonlarda protein fonksiyon tahmini için kritik bilgiler

sağlayabilecek bazı biyolojik veri setleri henüz yer almamaktadır. Fonksiyonel terimler, protein domainleri ve enzimatik reaksiyonlar gibi bileşenlerin dahil edilmesi ile bu koleksiyonların zenginleştirilmesi, fonksiyon tahmini için yararlı bir kaynak oluşturmalarını sağlayacaktır.

### 2.4. Biyolojik Araştırmalarda Çizge Tabanlı Mimariler

Geometrik öğrenme; geometri, bilgisayar bilimi ve makine öğrenmesi konseptlerini birleştirerek çizgeler ve nokta bulutları gibi geometrik yapıların analizi ve anlaşılması problemlerini ele alan bir araştırma alanıdır. Son yıllarda biyomedikal görüntü işleme, protein yapı analizi ve ilaç tasarımı gibi alanlarda uygulamalarına sıkça rastlanmaktadır (21,22). Geometrik öğrenmenin temel amacı, geometrik verilerden anlamlı bilgiler çıkarmak ve sınıflandırma, kümeleme, regresyon ve boyut indirgeme gibi çeşitli görevler için kullanmaktır. Veri noktaları arasındaki doğal geometrik yapı ve ilişkilerin analiz edilmesi gibi geleneksel makine öğrenmesi yöntemlerinin yetersiz kaldığı durumlarda geometrik öğrenme algoritmalarının kullanılması büyük avantaj sağlamaktadır. Geometrik öğrenmede en yaygın tekniklerden biri çizge tabanlı algoritmalardır. Çizgeler veri noktalarının bir "düğüm", bu noktalar arasındaki ilişkilerin ise "kenarlar" olarak tanımlandığı bir ilişkisel veri gösterim türüdür. Proteinler arasındaki fonksiyonel veya fiziksel ilişkileri temsil eden protein-protein etkileşim ağları, sadece tek bir düğüm tipi barındıran "homojen" çizgelere bir örnektir. Protein, gen, ilaç, hastalık gibi bileşen türlerinden birkaç tanesinin birbirleri ile olan ilişkilerini barındıran "heterojen" çizge örnekleri de bulunmaktadır (Şekil 2.3.). Örneğin protein, küçük molekül ve hastalık düğümleri içeren heterojen bir çizgede kenarlar, belli bir proteini hedefleyen küçük molekül ve bu proteinin ilişkili olduğu belli bir hastalık ile arasındaki bağları temsil edebilir. Son yıllarda biyolojik veri analizi ile hesaplamalı tahmin alanlarında çizge tabanlı hesaplamalı yaklaşımların uygulanması, bu alandaki araştırmaların hızlanmasına katkıda bulunmuştur. Çeşitli biyokimyasal bileşenlerin birbiriyle olan ilişkilerini barındıran çizge verilerini yüksek performanslı algoritmalar için girdi olarak kullanmak proteinlerin kompleks ve çok yönlü fonksiyonel bilgisinin ortaya çıkartılmasını sağlamaktadır (23).




**Şekil 2.3.** Homojen ve heterojen biyolojik çizge örnekleri. **(A)** Protein-protein etkileşimlerini içeren bir homojen çizge örneği. **(B)** Proteinler, bunları hedefleyen küçük moleküller ve ilişkili olduğu hastalıkları içeren bir heterojen çizge örneği.

Çizgelerin karmaşık topolojik yapısı, tahmin görevleri için gerekli hesaplamaları gerçekleştirerek veri içinde bulunan önemli kalıpların tespit edilebilmesini zorlaştırmaktadır. Bu sorunla başa çıkabilmek için çizge yapısının düşük boyutlu gösterimlerini yakalayan gömme yöntemleri geliştirilmiştir. Çizge gömme yöntemlerinin literatürdeki ilk örnekleri matris faktörizasyon yaklaşımları (24) veya dönüşüm-tabanlı yaklaşımlar (25–27) kullanarak çizgede yer alan düğümlerin diğer düğümler ile olan ilişkileri üzerinden temsil vektörlerini oluşturmaktadırlar. Bu yöntemler ile elde edilen düğüm temsil vektörleri, ilaç-ilaç ilişki tahmini (28) ve ilaç-hedef protein ilişki tahmini (29) gibi çeşitli görevler için öznitelik vektörü olarak kullanılabilmektedir.

Çizge gömmesi için özel olarak geliştirilmiş bu yaklaşımlar dışında derin öğrenme algoritmalarından da düğümlerin çizge içindeki ilişkilerini temsil edecek vektörler elde etme amacıyla yararlanılmıştır (30–33). Bu algoritmalar arasında en çok tercih edilenlerden biri çizge sinir ağı (*graph neural network*, GNN) mimarileridir (Şekil 2.4.). GNN; protein-protein etkileşim ağları , moleküler çizgeler veya biyomedikal bilgi çizgeleri gibi çizge yapısındaki ilişkisel veri setleri üzerinde çalışmaya spesifik olarak geliştirilmiş bir makine öğrenmesi modeli sınıfıdır (34). Çizge yapısındaki girdi verisini işleyip tahmin üretmek için "ağırlık" adı verilen öğrenilebilir parametrelerden yararlanır. Çizgede yer alan her bir düğümün yerel

komşularının özniteliklerinden yararlanarak düğüm gömme vektörlerini elde etmeyi amaçlar. Her bir düğüm gömme vektörü, komşu düğümlerden elde edilen bilginin sinir ağları üzerinden birleştirilmesi ile elde edilir. Bu süreç "mesaj aktarımı" ve "birleştirme" olmak üzere iki temel adımda meydana gelmektedir. Mesaj aktarımı, girdi verisine uygulanan belirli filtre veya operasyon setleri ile çizge yapısı ve düğümlerin öznitelikleri kullanılarak önemli bilgilerin çıkarıldığı adımdır. Birleştirme adımında ise mesaj aktarımı adımının çıktısı olan filtrelenmiş veri bir araya getirilerek çizge evrişimi operasyonu tamamlanır (Formül 2.1.)

$$h_v^{(l+1)} = \sigma(\phi(AGG(MSG_{u \in N(v)}(h_u^{(l)})), MSG \quad (h_v^{(l)})))$$

(2.1.)

**l:** katman sayısı,

**N(v):** v düğümünün tüm komşu düğümleri

**$h_u^{(l)}$:** u düğümünün l katmanındaki gömme vektörü

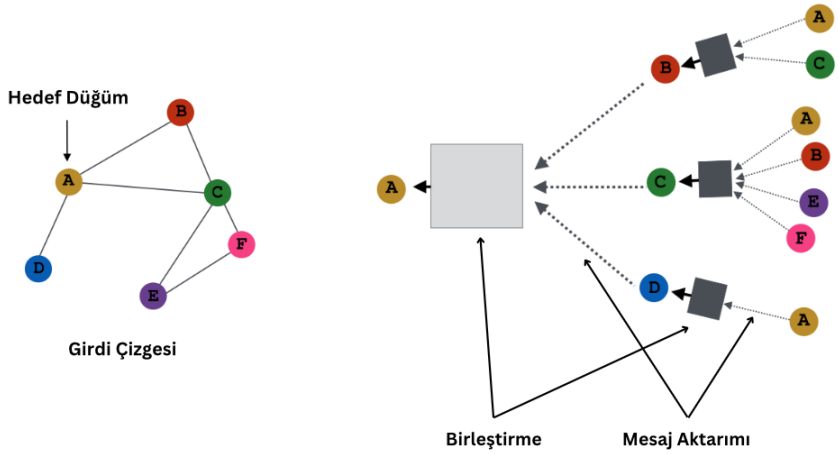
**MSG:** mesaj aktarımı

**AGG:** birleşim operasyonu

**$\phi$:** v düğümünün birleştirilmiş komşu mesajları arasında uygulanan operasyon

**$\sigma$:** aktivasyon fonksiyonu

Biyolojik ağların doğal ilişkisel yapısının bir çizge formunda olması, bu girdi verisi üzerinden tahmin üretmeyi amaçlayan çalışmalarda GNN mimarisinin başarıyla kullanımına olanak sağlamıştır (35–37).

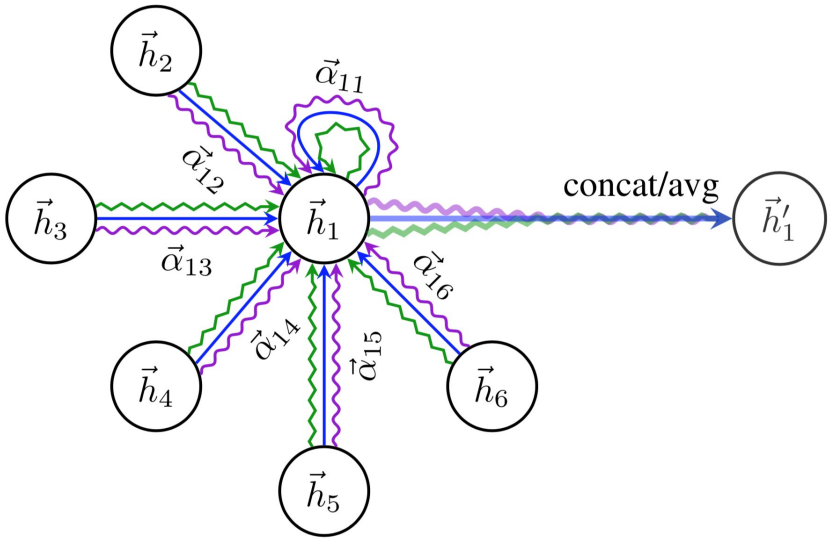



**Şekil 2.4.** Bir örnek çizge üzerinde A hedef düğümü için temel çizge sinir ağı operasyonlarının gösterimi (38).

Çizge evrişimsel sinir ağları (*graph convolutional neural networks*, GCN), literatürdeki çizge-bazlı biyolojik veri analizi çalışmalarında sıkça kullanılan popüler bir çizge sinir ağı sınıfıdır . GCN'de çizge evrişim katmanları üzerinden komşu düğüm bilgilerini birleştirilerek düğüm gömme vektörleri üretilir. Her bir evrişimsel katmanda düğümlerin öznitelik vektörleri komşu düğümlerin özniteliklerine bağlı olarak güncellenir (39). Bu sayede çizge içerisindeki lokal yapısal özellikler öğrenilir. Ağın çıktısı olarak elde edilen gömme vektörleri düğüm sınıflandırma ve bağlantı tahmini gibi çizge tabanlı tahmin görevlerinde kullanılabilir. GCN mimarileri, biyolojik çizgelerin anlaşılması amacıyla kullanılan önemli bir araç olsa da, çizge içerisinde sadece yakın komşulukları dikkate aldığı ve global çizge yapısının temsilinde yetersiz kaldığı için kompleks tahmin görevleri için uygun olmamaktadır.

Çizge sinir ağlarının bir sınıfı olan çizge dikkat ağları (*graph attention networks*, GAT), düğümler arasındaki daha kompleks ilişkileri yakalamak için mesaj aktarımı aşamasında dikkat mekanizmasını kullanır (Şekil 2.5.). GCN'in komşu düğüm özniteliklerini sabit ağırlıklarla birleştiren algoritmasına kıyasla bu mekanizma komşu düğümlerin önemine göre özniteliklerinin adaptif bir şekilde

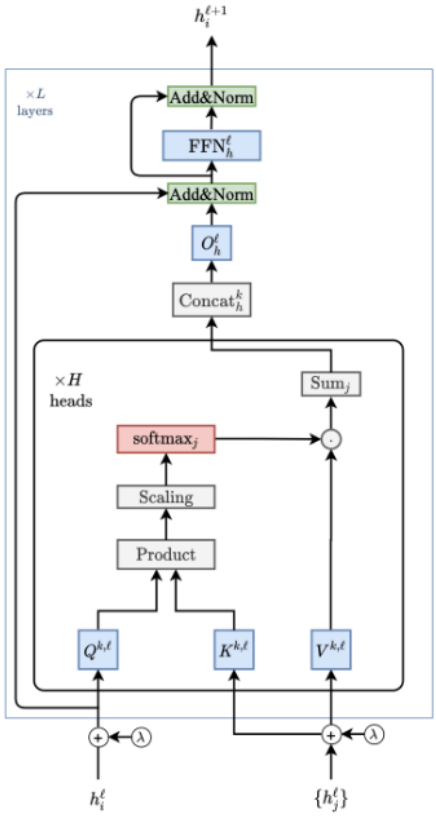
ağırlıklandırılmasını sağlar (40). Literatürde protein fonksiyon tahmini için bu mimariyi kullanan çalışmalar, GAT'ın dikkat mekanizması sayesinde artan ifade kapasitesinin protein fonksiyon tahmini için kullanıldığında GCN'e göre performansta artış sağladığını göstermiştir (41).




**Şekil 2.5.** Çizge dikkat ağlarında birleştirme adımının Düğüm 1 ve komşu düğümleri üzerinden gösterimi (40). Farklı çizgi renkleri farklı dikkat hesaplamalarını ifade etmektedir, her bir komşu düğüm için birden fazla "dikkat kafası" hesaplamak mümkündür. Bu örnekte 3 dikkat kafası mevcuttur. $\bar{a}$, komşu düğümlerin Düğüm 1 üzerindeki önemini ifade eden öğrenilebilir ağırlıklardır. Bunlarla ağırlıklandırılmış komşu öznitelik vektörlerinin birleştirilmesi (concat/avg) sonucu Düğüm 1'in güncellenmiş öznitelik vektörü oluşur.

Doğal dil işleme çalışmalarındaki başarısı ile dikkat çeken Transformer (42) modelinden ilham alan çizge dönüştürücü mimarisi (Şekil 2.6.), düğümler arasındaki kompleks etkileşimlerin ve uzak mesafedeki bağların yakalanmasını gerektiren biyolojik tahmin görevlerinde yüksek başarı göstermiştir (43–45). Çizge dönüştürücü mimarisinin ayırt edici özelliği olan öz-dikkat mekanizması, düğümler arasındaki uzak ilişkilerin etkili bir şekilde öğrenilmesini sağlamaktadır. Bu mekanizma, her bir düğümün diğer tüm düğümlere bir önem değeri ataması ile çalışır. Düğümlerin gösterimleri her bir katmanda diğer düğümlere atanan önem değeri ile ağırlıklandırılarak güncellenir. Literatürdeki çalışmalar, çizge dönüştürücü mimarisinin bu özelliği sayesinde büyük çaplı heterojen çizge verileri içindeki

karmaşık biyokimyasal ilişkilerin öğrenilmesini gerektiren görevler için oldukça uygun olduğunu göstermiştir (46).




**Şekil 2.6.** Temel çizge dönüştürücü mimarisi gösterimi (47). λ, düğüm özniteliklerine ilk katmana girmeden önce eklenen pozisyonel gömme vektörleridir. $h^l_i$ hedef düğümün, $\{h^l_j\}$ ise komşu düğümlerin l katmanındaki gömme vektörlerini ifade eder. Hedef düğüm üzerinde komşu düğümlerin önemi üzerinden hesaplanan tüm dikkat değerleri toplanır. Her bir dikkat kafasından ("*head*") çıkan değerin birbirine eklenmesinin ardından bu çıktı bir tam bağlı ağdan geçirilerek hedef düğümün güncellenmiş gömmesi ($h^{l+1}_i$) elde edilir.

## 2.5. Protein Fonksiyon Tahmini Çalışmaları

### 2.5.1. Öznitelik Tabanlı Yaklaşımlar

Öznitelik tabanlı fonksiyon tahmini yöntemleri, proteinlerin amino asit dizisi, üç boyutlu yapısı ve diğer proteinlerle/biyolojik bileşenlerle etkileşimleri gibi çeşitli karakteristik özelliklerinden yararlanarak fonksiyonel bilginin çıkarılmasına dayanır.

Bu yöntemlerde girdi verisi olarak çeşitli veri tabanları veya hesaplamalı yaklaşımlardan elde edilmiş fizikokimyasal özellikler, sekans motifleri, evrimsel ilişkiler, yapısal veya fonksiyonel anotasyonlar kullanılmaktadır (2). Hesaplamalı yaklaşımlar ile fonksiyonu bilinen bir protein seti üzerinden bu öznitelikler ve içlerinde barındırdıkları desenler analiz edilerek, bilinmeyen proteinler belirli fonksiyonel kategorilere sınıflandırılır. Proteinlerin sınıflandırılacağı kategorilerinin belirlenmesinde GO gibi fonksiyonel terimleri standardize bir formatta sunan anotasyon şemaları kullanılır.

Sekanslama teknolojilerindeki ilerlemeler sayesinde veri tabanlarında mevcut protein miktarının artsa da çoğunlukla bu proteinler için bilinen tek veri türü protein sekansıdır. Bu durum, sekans bilgisinin protein fonksiyon tahmini için en yaygın kullanılan öznitelik türlerinden biri olmasına neden olmuştur (23). Sekans verisinin kullanıldığı ilk çalışmalarda PSI-BLAST (48), PROSITE (49) ve PFAM (50) gibi sekans hizalama araçları kullanılarak protein dizileri karşılaştırılmış ve benzerlik seviyesine göre homolog oldukları tespit edilen proteinlerde bilinen fonksiyonlar birbirine aktarılmıştır. Homolojinin fonksiyonda korunmayı garanti etmediğini gösteren çalışmalar bu yaklaşımın hatalı tahminlere sebep olabildiğine işaret etmektedir (2). Bu sebeple protein sekans verisindeki gizli desenleri tespit edebilecek daha kompleks algoritmaları kullanan yaklaşımlar yaygınlaşmıştır. Bu yaklaşımlarda tahmin üretilecek protein setine ait sekans verisi alfabetik amino asit dizilerinden sayısal gösterimlere dönüştürülerek işlenebilir hale getirilir ve elde edilen sayısal öznitelik vektörleri belirlenen yüksek performanslı sınıflandırma algoritması üzerinden fonksiyon tahmini için kullanılır. Son yıllarda bu sınıflandırma görevi için makine öğrenmesi algoritmaları sıklıkla tercih edilmiştir (23). Bu çalışmalardan biri olan DeepGOPlus (51), sekans tabanlı fonksiyon tahmini yaklaşımı ve evrişimli sinir ağı mimarisini birleştirmiştir. DeepGOPlus'ın evrişimli sinir ağı modeli, sekans verisi üzerinde protein fonksiyonu için belirleyici olabilecek motifleri ortaya çıkartarak bu motifleri içeren proteinlerin fonksiyonları ile ilişkilendirerek tahmin üretir. Benzer metodlarla yapılan karşılaştırmalarda DeepGOPlus'ın özellikle CCO tahmininde en iyi tahmin metodları arasında olduğu görülmüştür.

Hesaplamalı protein fonksiyon tahmini için yararlanılabilecek bir diğer öznitelik türü de proteinlerin domain içeriğidir. Domainler, proteinleri oluşturan yapısal ve fonksiyonel üniteler olmaları dolayısıyla fonksiyonu belirleyen en önemli faktörlerden biridir. Literatürde proteinlerdeki benzer domain kompozisyonlarını tespit ederek bu benzerliği fonksiyon tahmini için kullanılan çalışmalar mevcuttur. Bunlardan bir bölümü sadece aynı domainlere sahip farklı proteinler üzerinden fonksiyon aktarımı yaparken (52), bazıları ise benzerlik tespiti için domainlerin protein dizisi içindeki sıralaması, pozisyonu ve tekrarlarından da yararlanmıştır (53). Domain içeriğinin de dahil olduğu birden fazla sekansla ilişkili özelliğin makine öğrenmesi algoritmalarıyla kullanımının protein fonksiyon tahmin performansını arttırdığı gözlemlenmiştir (54).

Hücre içindeki katlanmalar sonucu ulaştıkları üç boyutlu yapı, proteinlerin katalizledikleri biyokimyasal reaksiyonlar, katıldıkları sinyal mekanizmaları ve diğer moleküllere bağlanma seçicilikleri gibi önemli fonksiyonel karakteristikleri üzerinde belirleyicidir. Üç boyutlu yapının sekansa göre daha iyi korunması sebebiyle bazı proteinler arasında düşük sekans benzerliği olmasına rağmen yüksek yapısal benzerlik olduğu gözlemlenmiştir (55). Bu durum fonksiyon bilgisinin açığa çıkartılmasında yapı benzerliğinin sekans benzerliğine göre daha güvenilir olarak görülmesine yol açmıştır (55). Deneysel yaklaşımların yanında hesaplamalı yapısal biyoloji yöntemlerindeki gelişmeler (56–59), proteinlerin yapısal bilgisinin ortaya çıkartılması sürecini hızlandırmıştır. Bu gelişmeler ışığında gerçekleştirilen çalışmalarından biri olan COFACTOR (60), hem global hem de lokal yapısal karşılaştırma algoritmalarını birleştirerek fonksiyon tahmini yapmaktadır. Yapı-fonksiyonel anotasyon kaynaklarında bulunan benzer katlanmalara ve fonksiyonlara sahip proteinler şablon olarak belirlenip bu fonksiyonlar yine benzer katlanmalara sahip fonksiyonu bilinmeyen proteinlere aktarılmıştır. Sonuçlar lokal yapı karşılaştırmalarını da işin içine dahil etmenin farklı global katlanmalara ancak benzer bağlanma bölgelerine sahip olan fonksiyonel homologları tespit etmedeki avantajını göstermiştir.

Ancak, sadece yapısal homolojiye dayalı fonksiyon tahmininin birçok dezavantajı vardır. Mevcut yapı-fonksiyon anotasyonu veri tabanlarındaki eksiklikler ve işlevsel benzerliğin her zaman yapısal homolojinin bir sonucu olmaması (23)

bunlara bir örnektir. Fonksiyon tahmini yaparken proteinlerin tek bir özelliğinden yararlanmanın getirdiği bu ve benzeri kısıtlamalar, araştırmacıları birden fazla protein özelliğini bir arada kullanan yeni yaklaşımlar geliştirmeye itmiştir. Örneğin COFACTOR çalışmasının ilk kısmında sunulan yapısal homoloji tabanlı yaklaşım, sonrasında sisteme hem yapı ve sekans homolojilerinden hem de protein-protein etkileşim ağlarından elde edilen bilgilerden yararlanan hibrit modellerin dahil edilmesiyle geliştirilmiştir (61). Birden fazla protein özelliğinden yararlanan derin öğrenme bazlı bir model olan SDN2GO (62), proteinlerin sekans, domain içeriği ve protein-protein ağlarından işlevsel bilgiyi öğrenmek ve çıkarmak için evrişimli sinir ağlarını kullanır, ardından bu özellikleri ağırlıklı bir sınıflandırıcısıyla entegre ederek fonksiyon tahmini yapar.

Hibrit çalışmaların sonuçlarında da bir kez daha gözlemlenen performans gelişmeleri (63–65), tek bir özellik üzerinden tahmin üreten yaklaşımlarda proteinlerin çok yönlü fonksiyonel yapısına dair bilgilerin tam olarak yakalanamadığına işaret etmektedir. Yüksek tahmin performansı elde edilebilmesi için proteinlerin kendi özniteliklerini, diğer biyokimyasal bileşenler ile olan ilişkilerini ve diğer bileşen tiplerinin özniteliklerini de dikkate alan kapsamlı yaklaşımlara ihtiyaç duyulmaktadır.

### 2.5.2. Çizge Tabanlı Yaklaşımlar

Çizge tabanlı yöntemler, proteinlerin kompleks yapısını, etkileşim ağlarını ve evrimsel ilişkilerini temsil eden çizgeleri kullanarak protein fonksiyonlarının tahmininde önemli bir rol oynamaktadır. Bu tahmin görevi için çizge tabanlı yaklaşımlar ne kadar yeni keşfedilmeye başlansa da, literatürdeki sınırlı sayıda çalışmanın tahmin performansı sonuçları umut vericidir.

Bu çalışmalardan bir bölümünde girdi verisi olarak proteinlerin üç boyutlu yapısı içerisindeki amino asit etkileşimlerinin kullanımı tercih edilmiştir. DeepFRI (66), protein fonksiyonu ve fonksiyonel bölgeleri tespit etme amacıyla homojen amino asit etkileşim çizge verisi ile GCN mimarisini kullanmıştır. DeepFRI'nin öğrenme sürecinin birinci adımında önceden eğitilmiş bir protein dil modeli kullanılarak protein sekansının gömme vektörleri elde edilir. Sonrasında bu vektörleri girdi olarak kullanan GCN, her fonksiyonel terim için olasılıksal tahmin değerlerini çıktı olarak verir.

Sonuçlar, hesaplamalı olarak elde edilmiş protein yapı tahminleri bile girdi olarak kullanıldığında GCN'in gürültüyü yüksek ölçüde giderme yeteneğinin DeepFRI'nin yapıdaki tahmin hatalarına karşı dirençli olmasına katkıda bulunduğunu göstermiştir. Bu sayede deneysel olarak eldesi zor olan protein yapısı verisine ihtiyacı büyük ölçüde azaltmıştır.

GAT-GO (41) çalışmasında da benzer şekilde girdi olarak homojen protein içi etkileşim ağları kullanılmış, fakat öğrenme için temel GCN yerine GAT mimarisi tercih edilmiştir. GAT-GO, bir derin protein dil modeli kullanılarak üretilen düğüm (amino asit) öznitelik vektörleri ile tüm çizge (protein) seviyesinde bir gömme elde eder. Sonrasında elde edilen bu gömmeyi yine aynı derin protein dil modeli ile üretilen protein seviyesindeki öznitelik vektörü ile birleştirir. Bu birleşik vektörün tam bağlantılı bir sınıflandırıcıya verilmesi ile fonksiyonel terimler için olasılıksal terimler elde edilir. Sonuçlar, derin protein dil modeli ile üretilen öznitelik vektörlerinin kullanımının tahmin performansına büyük katkı sağladığını göstermiştir.

Çizge tabanlı yaklaşımlara bir diğer bakış açısı ise fonksiyonlar için önemli ipuçları içeren bir veri olan farklı proteinlerin birbiri ile olan ilişkilerinin kullanımıdır. Bu ilişkiler sekans benzerliği veya protein-protein etkileşimleri gibi farklı protein özellikleri ele alınarak tanımlanabilir. DeepGraphGO (36) bir GCN mimarisi ile hem protein-protein etkileşim çizgeleri hem de protein domain içeriği bilgisi üzerinden fonksiyon tahmini yapan bir çalışmadır. Çizge içerisinde yer alan düğümlerin (proteinler) öznitelik vektörleri domain kompozisyonlarını temsil etmektedir. Bu öznitelik vektörleri 2 katmanlı bir GCN ile güncellendikten sonra tam bağlı bir katmana verilerek fonksiyonel terimler için olasılıksal terimler elde edilir.

Girdi verisi olarak protein etkileşimleri dışında GO terimleri arasındaki hiyerarşik ilişkilerinin temsil edildiği çizgeleri kullanan çalışmalar da bulunmaktadır. Bu çalışmalardan biri olan PANDA2 (67), GO DAG yapısını homojen bir çizge olarak ele alıp GNN modeli için girdi olarak kullanır. Tüm sorgu proteinleri için girdi çizge topolojisi korunurken, düğüm ve çizge öznitelikleri her protein için dizi ve evrimsel özellikleri dikkate alınarak özel olarak tanımlanmıştır. 3 GNN katmanı boyunca güncellenen düğüm öznitelikleri fonksiyonel terimler için olasılıksal tahminlere dönüştürülür.

Çizge tabanlı yaklaşımlar, protein fonksiyon tahmini görevine getirdiği yenilikler ve sağladığı birçok fayda ile önemli bir ilerleme kaydetmiştir. Bu çalışmalar proteinlerin yapısal ve fonksiyonel özelliklerini bir arada ele alabilme yeteneği sunarak fonksiyon tahmininde daha güvenilir sonuçlar elde etmeyi sağlamıştır. Çizge tabanlı yöntemlerin esnekliği ve genişletilebilirliği, yeni veri kaynaklarının ve bilgi türlerinin entegrasyonunu kolaylaştırmaktadır. Bu sayede proteinlerin amino asit dizilimleri gibi sadece tek bir özelliği ile ilgili bilgilerle sınırlı kalmak yerine, protein yapıları, motifler, evrimsel koruma, işlevsel veya yapısal anotasyonlar gibi çeşitli özellikleri de öğrenmeye dahil ederek daha kapsamlı bir analiz yapabilir. Bununla birlikte, literatürdeki çalışmaların çoğunda proteinlerin sınırlı sayıda özelliğini ele alabilecek homojen çizge verileri ele alınmıştır. Proteinlerin çok yönlü fonksiyonel yapısını tam olarak yakalamak için diğer biyolojik bileşenlerle olan ilişkilerini de ele almak önemlidir. Bu amaç doğrultusunda çeşitli biyomedikal bileşenlerin birbiri ile olan ilişkilerini ve diğer bileşen türlerinin de özniteliklerini temsil edebilen heterojen çizge verileri kullanılarak daha kapsamlı ve bütüncül bir yaklaşım benimsenebilir.

### 2.5.3. Protein Fonksiyon Tahmini Metodlarının Değerlendirilmesi

Yeni geliştirilen protein fonksiyon tahmini yöntemlerinin mevcut fonksiyonel bilgi açığını kapatmadaki etkisinin anlaşılması için bu yöntemlerin tahmin doğruluğunu objektif olarak değerlendirip benzer yöntemlerle karşılaştırılmasının yapılması büyük önem taşımaktadır. Girdi verisinin kalitesi ve seçilen tahmin algoritması gibi etkenlerin tahminlerin güvenilirliği üzerinde önemli etkisi bulunmaktadır. Veri kaynaklarında yer alması muhtemel eksik, hatalı veya yanlı bilgi ve deneysel anotasyonların sınırlılığı, tahmin sürecinde belirsizliklerin oluşmasına sebep olabilmektedir. Aynı zamanda modellerin dayalı olduğu varsayım ve algoritmaların, proteinlerin çok yönlü fonksiyonel özelliklerini her yönden ele alabilecek şekilde seçilmesinin tahmin doğruluğuna olan etkisi göz önünde bulundurulmalıdır. Bu gibi etkenlerin tahmin doğruluğuna olan etkisinin değerlendirilmesi için geliştirilmiş kapsamlı karşılaştırma veri setleri ve yarışmalar bulunmaktadır. Fonksiyonel Anotasyonun Kritik Değerlendirmesi (*Critical Assessment of Functional Annotation*, CAFA), geniş çaplı fonksiyon tahmini yöntemlerinin performans değerlendirmesi ve karşılaştırması için geliştirilmiş, alanda

en benimsenmiş yarışmadır. Yarışmanın ilk aşamasında katılımcılara fonksiyonları bilinmeyen bir hedef protein seti sunar. Katılımcılar, kendi belirledikleri bir girdi verisi ve algoritma türü üzerinden bu proteinleri GO terimleri ile ilişkilendirerek tahmin üretirler. Tahminlerin bildirilmesi için belirlenen son tarihi, birkaç ay süren "anotasyon birikme" süreci izler. Bu süreç boyunca hedef proteinlerin bir kısmı için üretilen deneysel anotasyonlar toplanır. Son olarak bu deneysel anotasyonlar yarışmacılar tarafından üretilen tahminlerin doğruluğunu ölçmek için karşılaştırma veri seti olarak kullanılır. CAFA ve benzeri bu tür yarışmalar sayesinde mevcut yaklaşımların güçlü ve zayıf yönleri tespit edilerek daha güvenilir tahmin yöntemlerinin geliştirilmesine yön verilmektedir.

# 3. GEREÇ VE YÖNTEM

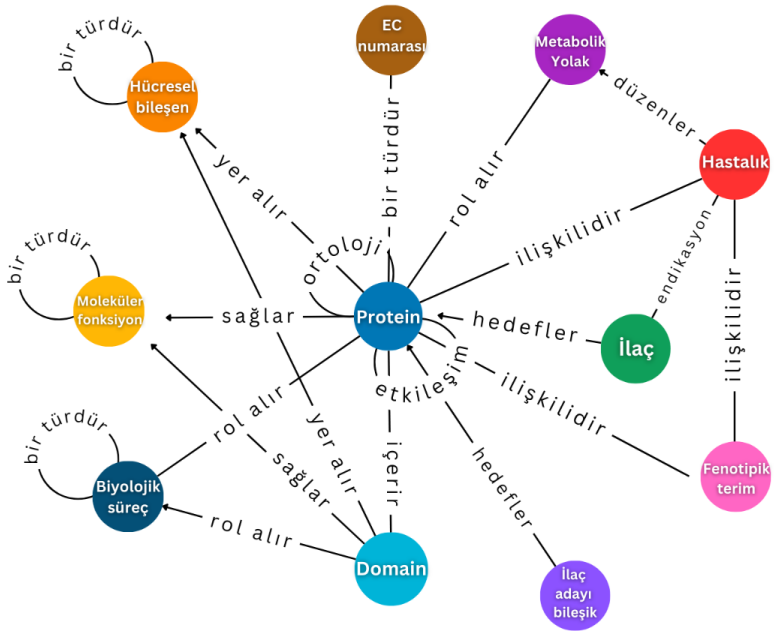
Bu tez çalışması kapsamında izlenen genel akış Şekil 1.1.'de gösterilmiştir. İlk olarak farklı biyomedikal veri tabanlarından elde edilen heterojen veri birbiri ile entegre edilip bilgi çizgeleri şeklinde modellenmiştir. Temsil yeteneği yüksek gömme yöntemleri kullanılarak bilgi çizgesinde yer alan bileşenlere ait öznitelik vektörleri oluşturulmuştur. Ardından çizge verisini girdi olarak kullanan, heterojen çizge dönüştürücü mimarisine dayalı tahmin modellerinin tasarımı, eğitimi ve optimizasyonu gerçekleştirilmiştir.

## 3.1. Veri

Farklı biyolojik/biyomedikal bileşenlere ait heterojen veri ilgili veri tabanlarından elde edilmiş ve bileşenler arasındaki ilişkiler bilgi çizgeleri şeklinde modellenmiştir. Veriyi oluşturan bileşen ve ilişkilerin büyük kısmının eldesi ve entegrasyonu için CROssBAR sistemi (13) kullanılmıştır. Bu kaynakta yer alan tüm insan proteinleri sorgu edilerek diğer bileşenler arasındaki ilişkilerini temsil edilen bilgi çizgeleri elde edilmiştir. Bu çizgeler birleştirilerek tek bir ana bilgi çizgesi haline getirilmiştir. Literatürde sıkça ele alınmış ve protein fonksiyon tahmini çalışmalarının performans değerlendirilmesi için oluşturulmuş yaygın kullanılan veri setlerinde yer verilmiş (68) önemli organizmalara ait protein bileşenleri ve aralarındaki ikili ortoloji ilişkileri, Orthologous MAtrix (OMA) veri tabanı (69) elde edilerek veri zenginleştirilmiştir. Bununla beraber veride yer alan protein koleksiyonunun kapsadığı organizma sayısı toplamda 29'a ulaşmıştır. Bu koleksiyonda yer alan proteinlerin tamamı, uzmanlar tarafından kürate edilmiş UniProtKB/Swiss-Prot proteinlerinden oluşmaktadır. UniProtKB/TrEMBL proteinleri kendileri ile ilgili çalışma ve verinin azlığı sebebi ile çizgeye dahil edilmemiştir.

Ardından veriye protein domainleri ve ontoloji bazlı biyomoleküler fonksiyonel terim (GO terimleri ve EC numaraları) bileşenlerinin eklenmesi gerçekleştirilmiştir. Domain terimleri ve protein-domain InterPro (70) veri tabanından, domain-fonksiyonel terim ilişkileri ise InterPro2GO veri setinden elde edilmiştir. GO terimleri ve birbirleri arasındaki hiyerarşik ilişkiler Gene Ontology (4) kaynağından

alınmıştır. Protein-GO terimi ilişkileri ise Gene Ontology Annotation (GOA) (71) veri tabanından elde edilmiştir. GOA, deneysel çalışma sonuçları, hesaplamalı tahminler ve literatür taramaları gibi çeşitli kaynaklardan toplanmış güvenilir ve güncel fonksiyonel anotasyon verisini kullanıcıya sunmaktadır. Elde edilen protein-GO terimi ilişkileri arasında "Elektronik Anotasyonla Tahmin Edildi" (*Inferred from Electronic Annotation*, IEA) kanıt koduna sahip olanlar dışındaki tümü veriye dahil edilmiştir. Bu kanıt koduna sahip ilişkiler güvenilirliği sorgulanabilir olduğundan ele alınmamıştır. Son olarak EC numarası bileşenleri ve bunların protein bileşenleri ile aralarındaki ilişkiler İsviçre Biyoinformatik Enstitüsü'nün sunduğu Expasy ENZYME (72) veri tabanından alınmıştır. Bu eklemelerin ardından heterojen çizge verisinde yer alan tüm bileşenler ve içinde yer aldığı kenar tipleri Şekil 3.1.'de verilmiştir.




**Şekil 3.1.** Heterojen çizge verisinde yer alan düğüm ve kenar tipleri.

Dizi benzerliği yüksek olan proteinlerden yalnızca birini veri setinde bırakmak için UniProtKB'de sunulan UniProt Referans Kümeleri (*UniProt Reference Clusters*, UniRef) veri setlerinden UniRef90 kullanılmıştır (73). Bu sayede öğrenme sırasında eğitim ve test setleri arasında veri sızması olması engellenmiştir.

### 3.2. Öznitelik Vektörleri

Veride yer alan her düğüm tipi için temsil yeteneği yüksek son teknoloji gömme metodlarından uygun bir tanesi seçilerek öznitelik vektörlerinin (sayısal gösterimler) üretilmesinde kullanılmıştır. Protein ve biyoteknolojik ilaç düğümlerine ait öznitelik vektörleri, dönüştürücü mimarisi-tabanlı bir dil modeline (74) protein sekanslarının girdi olarak verilmesi ile elde edilmiş 768 boyutlu gömmelerdir. Fonksiyonel terim düğümleri için anc2vec yöntemi (75) ile GO veri tabanı 2020-10-06 tarihli versiyonu üzerinden önceden hesaplanmış 200 boyutlu vektör seti kullanılmıştır. Anc2Vec yöntemi, fonksiyonel terimler için GO veri tabanının; terimlerin ontolojik özgünlüğü, atasal ilişkileri ve hangi alt ontolojiye ait olduğu olmak üzere üç ana yapısal özelliği üzerinden sayısal gösterimler üretmek üzere tasarlanmıştır. Domain düğümlerinin öznitelik vektörleri; protein sekanslarını cümleler, domainleri ise bu cümleleri oluşturan kelimeler olarak ele alarak word2vec yöntemi (76) ile oluşturulmuş 50 boyutlu dom2Vec (77) gömme setinden alınmıştır. EC numarası düğümlerinin 256 boyutlu öznitelik vektörleri, her numarayla ilişkili reaksiyonun SMILES gösterimleri (78) üzerinden Python RXNFP kütüphanesi (79) kullanılarak elde edilmiştir.

Hastalık düğümlerinin öznitelik vektörleri, PrimeKG veri setinde (10) yer alan hastalık tanımlarının doc2vec metodu (80) ile 100 boyutlu sayısal gösterimlere dönüştürülmesi ile elde edilmiştir. Eğer hastalığın tanımı PrimeKG kaynağında yoksa kaynak veri tabanındaki tanımı, bunun da bulunmaması halinde hastalığın ismi doc2vec modeline girdi olarak verilmiştir. Biyolojik yolak düğümlerinin 200 boyutlu vektörlerinin üretilmesinde, gen-biyolojik yolak etkileşim ağları üzerinden TransE yöntemini (25) kullanarak gömmeler üreten Python BioKEEN kütüphanesinden (81) yararlanılmıştır. Fenotipik terim düğümleri için gen-fenotip ilişki ağları üzerinden node2vec yöntemi (82) ile gömmeler üreten CADA aracı (83) kullanılarak 160 boyutlu vektörler üretilmiştir. İlaç adayı bileşik düğümlerinin öznitelik vektörleri, bileşiklerin SELFIES gösterimleri (84) üzerinden dönüştürücü-bazlı bir dil işleme modeli (45) kullanılarak elde edilmiş 768 boyutlu gömmelerdir.

### 3.3. Heterojen Çizge Dönüştürücü Mimarisi

Heterojen çizge dönüştürücü (*Heterogeneous graph transformer*, HGT) mimarisi, heterojen bilge çizgeleri üzerinde yer alan farklı tiplerdeki kenar ve düğümleri kendine özgü gösterimlerle modellemek üzere tasarlanmıştır (85). HGT mimarisinin amacı, bir düğümü hedef alan kaynak düğümlerin barındırdığı bilgiyi toplayıp bir araya getirerek bu hedef düğüm için anlamlı bir temsil elde etmektir. Bu işlem bir HGT bloğunu oluşturan üç alt modül üzerinden gerçekleşmektedir (Şekil 3.2.).

Bunlardan ilki olan heterojen ortak dikkat modülü (Şekil 3.2. (1)), bir kenarı oluşturan *s* kaynak (*source*) düğümü ile *t* hedef (*target*) düğümü arasındaki dikkat skorunu (Formül 3.1.) hesaplamaktadır. Bu adımda Transformer (42) mimarisinin tasarımına benzer şekilde kaynak ve hedef düğümleri ifade eden vektörler doğrusal birer izdüşüm aracılığıyla sırasıyla Anahtar (*Key*) ve Sorgu (*Query*) vektörlerine haritalanmaktadır.

Her düğümün kendi tipine özgü bağımsız bir doğrusal izdüşüme girdi olarak verilmesi ile heterojen çizge içerisindeki tüm düğüm tiplerinin özgün bir şekilde modellenmesi sağlanmıştır. Elde edilen Sorgu ve Anahtar vektörleri arasındaki benzerliğin hesaplanması da ele alınan kenarın tipine ($\phi(e)$) özgü bir matris ($W_{\phi(e)}^{ATT}$) üzerinden gerçekleştirilmektedir. Böylelikle aynı iki düğüm arasında birden fazla tipte kenar bulunması halinde modelin bu kenarlar arasındaki anlamsal farklılıkları yakalaması sağlanmıştır. Buna ek olarak her ilişkinin hedef düğüme olan katkısı eşit olmayacağından, bu adıma her düğüm-kenar-düğüm üçlü tipinin genel önemini belirten bir μ tensörü eklenmiştir. Her düğüm çifti için hesaplanan her dikkat kafasının birbiri ardına eklenmesi ile o çiftin dikkat vektörü elde edilir.

Dikkat skorunun hesaplanmasına paralel olarak kaynak düğümlerden hedef düğümlere mesaj geçişi gerçekleştirilmektedir (Şekil 3.2. (2)). Dikkat işlemine benzer şekilde, kaynak düğümün vektörü doğrusal bir izdüşüm mesaj vektörüne aktarılmakta ve ele alınan kaynak-hedef düğümü arasındaki kenarın tipine özgü bir matristen ($W_{\phi(e)}^{MSG}$) geçmektedir (Formül 3.2.).

$$\mathbf{Attention}_{HGT}(s,e,t) = \underset{\forall s \epsilon N(t)}{\text{Softmax}}\left(\underset{i\epsilon[1,h]}{\Vert} ATT-head^i(s,e,t)\right)$$

$$ATT-head^i(s,e,t) = \left(K^i(s)W^{ATT}_{\phi(e)}Q^i(t)^T\right)\cdot\frac{\mu\langle\tau(s),\phi(e),\tau(t)\rangle}{\sqrt{d}}$$

(3.1.)

$$K^i(s) = K-Linear^i_{\tau(s)}\left(H^{(l-1)}[s]\right)$$

$$Q^i(t) = Q-Linear^i_{\tau(t)}\left(H^{(l-1)}[t]\right)$$

**s** kaynak düğüm

**t** hedef düğüm

**e** kaynak ve hedef düğümler arasındaki kenar

**N(t)** hedef düğümün tüm komşu düğümleri

**h** dikkat kafası sayısı

**$K^i(s)$** kaynak düğümün i dikkat kafasındaki anahtar vektörü

**$Q^i(t)$** hedef düğümün i dikkat kafasındaki sorgu vektörü

**K-Linear$_{\tau(s)}^i$** kaynak düğüm tipi için i dikkat kafasındaki anahtar lineer izdüşümü

**Q-Linear$_{\tau(t)}^i$** kaynak düğüm tipi için i dikkat kafasındaki sorgu lineer izdüşümü

**$W^{ATT}_{\phi(e)}$** kaynak ve hedef düğümler arasındaki kenar tipinin dikkat matrisi

**d** lineer izdüşüm vektör boyutu

**μ** kaynak düğüm tipi-kenar tipi-hedef düğüm tipi üçlüsünün önem tensoru

**$H^{(l-1)}[s]$** kaynak düğümün önceki katmandaki gömme vektörü

**$H^{(l-1)}[t]$** hedef düğümün önceki katmandaki gömme vektörü

$$\mathbf{Message}_{HGT}(s,e,t) = \underset{i\epsilon[1,h]}{\Vert} MSG-head^i(s,e,t)$$

(3.2.)

$$MSG-head^i(s,e,t) = M-Linear^i_{\tau(s)}\left(H^{(l-1)}[s]\right)W^{MSG}_{\phi(e)}$$

**s** kaynak düğüm

**t** hedef düğüm

**e** kaynak ve hedef düğümler arasındaki kenar

**h** mesaj kafası sayısı

**M-Linear$_{\tau(s)}^{i}$** kaynak düğüm tipi için i dikkat kafasındaki mesaj lineer izdüşümü

$\boldsymbol{W_{\phi(e)}^{MSG}}$ kaynak ve hedef düğümler arasındaki kenar tipinin mesaj matrisi

**H$^{(l-1)}$[s]** kaynak düğümün önceki katmandaki gömme vektörü

Son olarak *t* düğümünü hedefleyen tüm kaynak düğümlerin mesajları, dikkatleri ile ağırlıklandırılmaktadır (Formül 3.3.). Bu şekilde tüm komşuların özellik bilgilerinin *t* düğümü üzerinde toplanması sağlanmaktadır. t düğümünün güncellenmiş vektörü ($H\sim^{(l)}[t]$) son olarak bir doğrusal olmayan aktivasyon ve rezidüel bağlanma katmanının ardından bir doğrusal izdüşüm ile kendi düğüm tipine özgü dağılıma haritalanır (Formül 3.3.) Böylece *t* düğümünün bu bloktaki ($l$) güncellenmiş vektörü ($H^{(l)}[t]$) elde edilmektedir.

$$\widetilde{H}^{(l)}[t] = \bigoplus_{\forall s \epsilon N(t)} (\mathbf{Attention}_{HGT}(s,e,t) \cdot \mathbf{Message}_{HGT}(s,e,t)) \tag{3.3.}$$

$$H^{(l)}[t] = \sigma\left(A - Linear_{\tau(t)}\widetilde{H}^{(l)}[s]\right) + H^{(l-1)}[t]$$

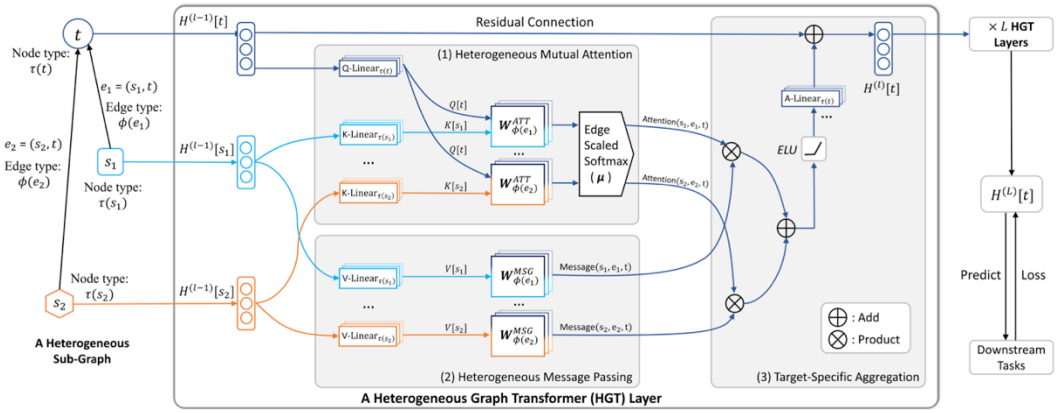
**s** kaynak düğüm

**t** hedef düğüm

**e** kaynak ve hedef düğümler arasındaki kenar

Birden fazla HGT bloğunun bir sinir ağını oluşturan katmanlar olarak kullanılması ile her bir düğüme çizgede yer alan düğümlerin büyük bir kısmından bilgi aktarımı yapılması sağlanabilecektir. Bu bilgi aktarımı sayesinde her bir düğüm için diğer düğümlerle olan bağlantılarını ve çizge içerisindeki yer aldığı lokal yapıyı anlamlı şekilde ifade eden gösterimler elde edilebilmektedir. Bu gösterimler düğüm

sınıflandırılması veya ilişki tahmini gibi çeşitli heterojen ağ görevlerinde girdi olarak kullanılmaya uygun olacaktır.




**Şekil 3.2.** Heterojen Çizge Dönüştürücü (HGT) mimarisi (85). HGT bloğu üç modülden oluşmaktadır: (1) İlişki-duyarlı heterojen ortak dikkat, (2) kaynak düğümlerden heterojen mesaj aktarımı ve (3) hedefe-özgü heterojen mesaj birleşimi modülü.

179 milyon düğüm ve 2 milyar kenar içeren bir heterojen çizge veri seti üzerinde yapılan deneyler, çeşitli tahmin görevlerinde HGT modelinin diğer çizge sinir ağı modellerine göre daha iyi performans elde ettiğini göstermiştir (85). Bu bulgular ışığında tez çalışmasında sunulan tahmin modeli, HGT katmanlarından oluşan bir sinir ağı şeklinde tasarlanmıştır (Şekil 1.1.). Her bir olası protein-fonksiyonel terim düğüm çifti için, bu düğümlerin son HGT katmanından çıkan gömme vektörlerinin iç çarpımı alınmaktadır. Bu çıktının sigmoid aktivasyon fonksiyonuna tabi tutulması ile protein-fonksiyonel terim arasındaki ilişkinin varlığına dair olasılık değeri elde edilmektedir.

### 3.4. Model Eğitimi

Model PyTorch Geometric kütüphanesinin (86) HGT mimarisi implementasyonu olan HGTConv evrişim katmanları kullanılarak inşa edilmiştir. GO alt ontolojilerinin kapsam, bilgi içeriği ve yapı olarak farklılık göstermesi sebebiyle her bir alt ontoloji için (moleküler fonksiyon, biyolojik süreç ve hücresel bileşen) ayrı bir model kurulmuş ve bu üç model model bağımsız olarak eğitilip değerlendirilmiştir. Farklı oranlarda (1:20 ve 1:100) negatif örnekleme deneyleri yapılarak gerçeğe en uygun tahmin senaryosu yaratılıp modelin anlamlı bağlantılar yerine rastgele bağlantılar tahmin edilmesinin önüne geçilmiştir. Model parametreleri uçtan uca

öğrenme yaklaşımına uygun olarak toplu şekilde ikili çapraz entropi kayıp fonksiyonu (Formül 3.4.) üzerinden optimize edilmiştir. Negatif örnekleme sonucu dengesiz bir veri seti ile öğrenme gerçekleştirildiğinden kayıp fonksiyonunda pozitif tahminlere negatif örnekleme oranı miktarınca ağırlıklandırma uygulanmıştır. Hiperparametre optimizasyonu ve performans testleri, bilinen protein-fonksiyon ilişkilerinin sırasıyla %80, %10, %10 olarak ayrılan eğitim, validasyon ve test setleri üzerinden gerçekleştirilmiştir. Rastgele arama yöntemi kullanılarak saklı kanal sayısı (16, 32, 64, 128, 256), öğrenme oranı (0.0001, 0.001, 0.005, 0.01), kafa sayısı (2, 4, 8, 16), evrişim katmanı sayısı (1, 2, 3, 4), eğitim seti küme boyutu (512, 1024, 2048, 4096), eğitim seti küme komşu sayısı (32, 128, 256) ve ağırlık sönümü (0, 1e-5, 5e-5, 0.0001, 0.001) hiperparametrelerinin model performansına etkisi incelenip optimal hiperparametreler belirlenmiştir. Denenen hiperparametre setleri ile 25 döngü boyunca eğitilen modeller arasından en yüksek validasyon performansı elde eden belirlenip kapsamlı protein-fonksiyon tahmini için kullanılmıştır.

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i}^{N} y_i \log(\hat{y}_i) \tag{3.4.}$$

**N** Gözlem sayısı

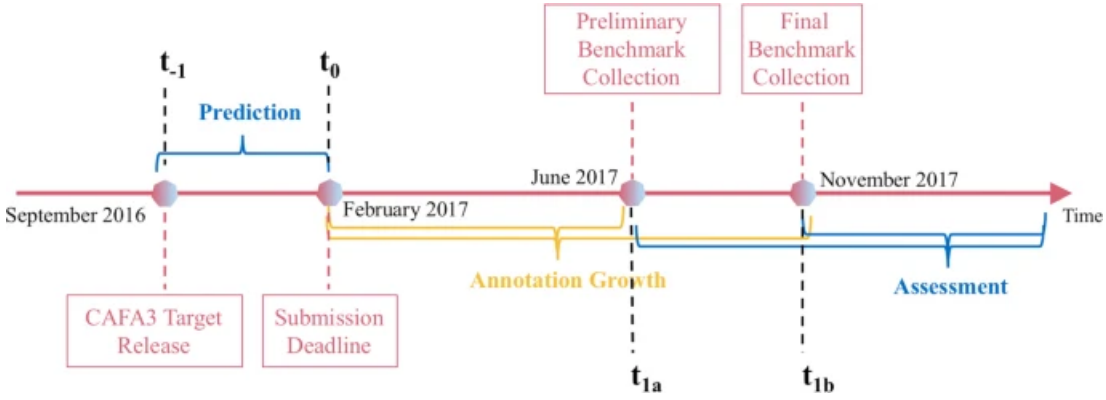
$\boldsymbol{y_i}$ *i* gözleminin gerçek etiketi

$\boldsymbol{\hat{y}_i}$ *i* gözlemi için üretilen olasılıksal tahmin değeri

### 3.5. Performans Değerlendirmesi

Çalışma kapsamında denenen modellerin kendi içlerinde değerlendirilmesi başta ayrılan %10 test veri seti üzerinden gerçekleştirilmiştir. Geliştirilen modelin tahmin performansı ise bundan bağımsız olarak CAFA yarışmasının sonuçları yayınlanmış olan en yeni versiyon test veri seti (CAFA3) (87) üzerinden değerlendirilip diğer protein fonksiyon tahmini modelleri ile karşılaştırılmıştır. Yarışmanın bu versiyonunda fonksiyonel anotasyon tahmini üretilmesi beklenen hedef protein listesi Eylül 2016'da duyurulmuştur. Yarışmacı metodlar, veri tabanlarının bu tarihe kadarki en güncel versiyonunda yer alan veri setleri kullanılarak geliştirilip anotasyon toplama süresi başlangıcına kadar tahmin üretmişlerdir. Şubat-Kasım 2017

tarihleri arasındaki toplanan deneysel anotasyonlar model performanslarının değerlendirilmesinde kullanılan karşılaştırma veri setini oluşturmuştur (Şekil 3.3.).




**Şekil 3.3.** CAFA3 yarışma zaman çizelgesi (87). Eylül 2016: Hedef protein listesinin duyurulması. Eylül 2016-Şubat 2017 aralığı: tahmin üretme. Şubat-Kasım 2017 aralığı: Deneysel anotasyonların toplanması. Kasım 2017 ve sonrası: Tahminlerin değerlendirilmesi.

CAFA3 deneylerine katılan diğer metodlar ile adil bir karşılaştırma yapılabilmesi için yarışma takvimine uygun veri setlerinin entegrasyonu ile ayrı bir eğitim çizgesi hazırlanmıştır. Bu veri setinde CROssBAR veri tabanından elde edilen iskelet çizgeye (Bkz. Bölüm 3.1) olduğu gibi yer verilmiştir. Bu iskelet üzerine sonradan eklenen ve fonksiyon tahmini ile daha ilişkili olduğu düşünülen fonksiyonel terim ve domain düğümleri ile bu düğümlerin yer aldığı kenarlar ise CAFA3 takvimine uygun olacak şekilde ilgili veri tabanlarının Eylül 2016 öncesi son versiyonundan elde edilmiştir. Hazırlanan veri seti ile eğitilen modellerde de Bölüm 3.6'da belirtilen hiperparametre değer aralıkları ile hiperparametre optimizasyonu yapılmış ve en yüksek validasyon performansına ulaşan modeller belirlenmiştir. Bu modellerin, karşılaştırma veri setinde yer alan protein ve fonksiyonel terimlerin tüm olası kombinasyonları için ürettiği olasılıksal ilişki tahmin performansı değerlendirilmiştir.

### 3.5.1. Değerlendirme Metrikleri

Modellerin kendi içlerinde değerlendirilmesinde sınıflandırma ve ilişki tahmini çalışmalarının değerlendirilmesi için sıkça kullanılan 3 metrik ele alınmıştır. F1-skoru tahminlerin kesinlik (*precision*, pr) (Formül 3.5.) ve doğruluk (*recall*, rc) (Formül 3.6.) değerlerinin harmonik ortalaması ile hesaplanır (Formül 3.7.). Kesinlik-doğruluk eğrisi altında kalan alan (*area under the precision-recall curve*, AUPR) değeri bir

eksende kesinlik, diğerinde doğruluk değerlerinin yer aldığı eğrinin altında kalan alanı ifade eder. Matthews korelasyon katsayısı (*Matthews correlation coefficient*, MCC), tahmin performansını değerlendirirken veri setindeki dengesizlik durumunu da hesaba katan bir metriktir (Formül 3.8.). Bu sebeple bu çalışmada olduğu gibi negatif veri noktası sayısının pozitif veri noktası göre oldukça yüksek olduğu veri setlerinde değerlendirmeye dahil edilmesi anlamlı olacaktır.

$$pr = \frac{TP}{TP + FN} \tag{3.5.}$$

$$rc = \frac{TP}{TP + FP} \tag{3.6.}$$

$$F1 = 2 * \frac{pr * rc}{pr + rc} \tag{3.7.}$$

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.8.}$$

***TP, TN, FP ve FN*** Doğru pozitif (*true positive*), doğru negatif (*true negative*) yanlış pozitif (*false positive*) ve yanlış negatif (*false negative*) tahmin sayısı

Model performansının diğer metodlarla karşılaştırılması CAFA yarışmasında sunulan protein-merkezli değerlendirme modunda (87) yapılmıştır. Bu değerlendirme şeklinde modelin bir protein için ürettiği GO terimi tahminlerinin doğruluğu ölçülmektedir. Ölçümler iki temel değerlendirme metriği üzerinden yapılmaktadır. Bu metrikler modelin ürettiği olasılıksal tahminleri belirli eşik değerlerinde; eşik değerin altında kalan tahminleri negatif, üstündekileri ise pozitif olarak ele alarak ölçümlemektedir.

Metriklerden ilki olan Fmax değeri (Formül 3.9.) her eşik değerindeki kesinlik (Formül 3.5.) ve doğruluk (Formül 3.6.) değerleri kullanılarak hesaplanan F-skorlarının en yükseğidir. Smin metriği her eşik değerinde gerçek anotasyonlar ile tahminler arasındaki anlamsal uzaklığın minimum değerini ifade eder. Kalan belirsizlik (*remaining uncertainty*, ru) ve yanlış bilgi (*misinformation*, mi) metrikleri üzerinden hesaplanır (Formül 3.10.). ru ve mi değerleri, GO terimlerinin bilgisel içerik miktarını da hesaba katarak tahmin performansını değerlendirir. Daha az bilgi verici terimler için yapılan tahminlerin performansa katkısı, daha bilgi verici terimlere göre düşük olarak değerlendirilir (88).

$$Fmax = max_{\tau}\left\{\frac{2 * pr(\tau) * rc(\tau)}{pr(\tau) + rc(\tau)}\right\} \tag{3.9.}$$

$$Smin = min_{\tau}\left\{\sqrt{ru(\tau)^2 + mi(\tau)^2}\right\} \tag{3.10.}$$

$\boldsymbol{pr}$ Kesinlik değeri

$\boldsymbol{rc}$ Doğruluk değeri

$\boldsymbol{ru}$ Kalan belirsizlik değeri

$\boldsymbol{mi}$ Yanlış bilgi değeri

$\boldsymbol{\tau}$ Olasılıksal tahmin eşik değeri

### 3.6. Uygulama Detayları ve Kullanılan Araçlar

Çalışma Python programlama dili kullanılarak gerçekleştirilmiştir. Veri setinin hazırlanmasında Pandas (89) ve NumPy (90) kütüphaneleri kullanılmıştır. Öznitelik vektörlerinin oluşturulmasında Biopython (91), gensim (80), Natural Language Toolkit (92), BioKEEN (81), TAPE (74), anc2vec (75), RXNFP (79) kütüphanelerinden yararlanılmıştır. Verilerin görselleştirilmesi Scikit-learn (93) ve Matplotlib (94), modelin programlanması ise PyTorch (95) kütüphaneleri kullanılarak gerçekleştirilmiştir. Bilgi çizgelerinin ve öznitelik vektörlerinin oluşturulması aşamaları kişisel bilgisayar; model eğitimi, validasyonu, optimizasyonu ve büyük çaplı

tahmin üretimi aşamaları ise GPU'lu sunucu kullanılarak gerçekleştirilmiştir. GPU'lu sunucu özellikleri; HP Z8 G4 Workstation, 2 x HP Intel Xeon Gold 5215 2.50GHz CPU -40 cores-, HP 128GB bellek, 2 x NVIDIA GeForce RTX2080, HP Z Turbo Drive G2 512GB SSD + HP 1TB 7200rpm SATA HDD.

# 4. BULGULAR

## 4.1. Veri Araştırması

### 4.1.1. Eğitim Veri Seti İstatistikleri

Bilgi çizgesinde yer alan kenar tipi istatistikleri Tablo 4.1.'de verilmiştir. Veride yer alan en kapsamlı kenar tipi toplam 2,599,935 kenar sayısıyla çizgenin %68'ini oluşturan protein-fonksiyonel terim kenarlarıdır. Bu kenarların %38'ini proteinlerin biyolojik süreç terimleri ile, %36'sını moleküler fonksiyon terimleri ile ve kalan %26'sını hücresel bileşen terimleri ile olan ilişkileri oluşturmaktadır. Bu veri, çalışma kapsamında denenen modellerin kendi içlerinde değerlendirilmesi ve geniş çaplı fonksiyon tahmini üretimi için kullanılan final modellerin geliştirilmesinde kullanılmıştır.

**Tablo 4.1.** Eğitim çizge verisinde yer alan ilişki tipleri ve sayıları.

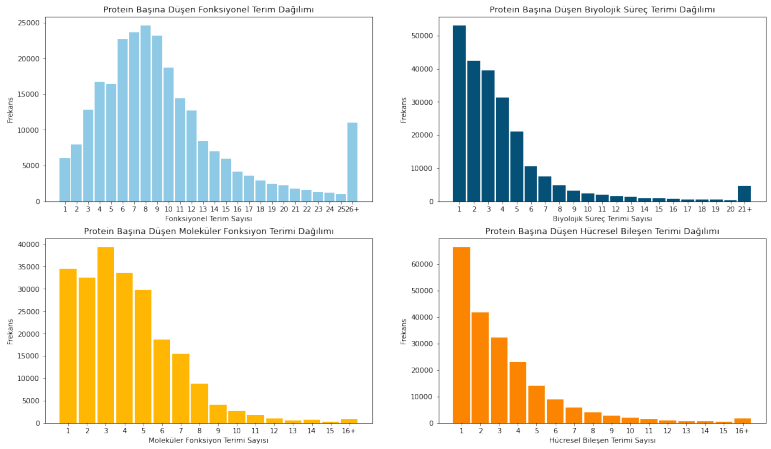
| İlişki Tipi | Sayısı |
|---|---|
| **Protein-Biyolojik süreç** | 1,000,952 |
| **Protein-Moleküler fonksiyon** | 926,844 |
| **Protein-Hücresel bileşen** | 672,139 |
| **Protein-Domain** | 461,668 |
| **Protein-EC numarası** | 128,242 |
| **Protein-İlaç adayı bileşik** | 153,764 |
| **Ortoloji** | 125,372 |
| **Fonksiyonel terim-Fonksiyonel terim** | 83,831 |
| **Protein-protein etkileşimi** | 82,524 |
| **Protein-Biyolojik yolak** | 51,692 |
| **Protein-Fenotipik terim** | 38,239 |
| **Hastalık-Fenotipik terim** | 24,259 |
| **Protein-İlaç** | 22,718 |
| **Protein-Hastalık** | 12,817 |

**Tablo 4.1.** Eğitim çizge verisinde yer alan ilişki tipleri ve sayıları (Devamı).

| | |
|---|---|
| **Domain-Fonksiyonel terim** | 5,791 |
| **Hastalık-Biyolojik yolak** | 1,682 |
| **Hastalık-İlaç** | 298 |
| **Toplam** | **3,792,832** |

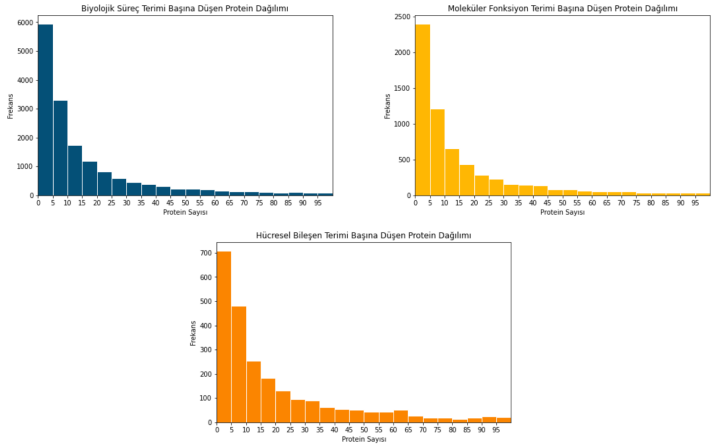
"Fonksiyonel terim" şeklinde belirtilen düğüm tipi "biyolojik süreç", "moleküler fonksiyon" ve "hücresel bileşen" düğüm tiplerinin tümünü kapsamaktadır.

Veride yer alan protein düğümü başına düşen fonksiyonel terim komşu sayısı dağılımları Şekil 4.1.'de gösterilmiştir. Protein düğümü başına düşen ortalama biyolojik süreç ve moleküler fonksiyon komşu düğüm sayısı yaklaşık 4 iken bu sayı hücresel bileşen komşuları için yaklaşık 3 olarak gözlemlenmiştir. Protein düğümlerinin alt kategori fark etmeksizin ortalama yaklaşık 10 fonksiyonel terim komşusu bulunmaktadır.




**Şekil 4.1.** Eğitim verisinde yer alan protein düğümleri başına düşen fonksiyonel terim düğümü dağılımları.

Veride yer alan fonksiyonel terim düğümü başına düşen protein komşu sayısı dağılımları ise Şekil 4.2.'de gösterilmiştir. Biyolojik süreç düğümleri ile ilişkili ortalama yaklaşık 56 protein düğümü bulunurken bu sayı moleküler fonksiyon düğümleri için ortalama yaklaşık 131, hücresel bileşen düğümleri içinse ortalama yaklaşık 238'dir.




**Şekil 4.2.** Eğitim verisinde yer alan fonksiyonel terim düğümleri başına düşen protein düğümü dağılımları.

### 4.1.2. Öznitelik Vektörleri

Düğüm öznitelik vektörlerinin elde edilmesi aşamasında girdi verisindeki eksiklikler veya önceden hesaplanmış gömme setlerindeki veri tabanı versiyon uyumsuzlukları gibi sebeplerle düğümlerin belli bir kısmı için vektör elde edilebilmiştir. Öznitelik vektörleri elde edilemeyen düğümler girdi çizge verisinden çıkartılmıştır. Son durumda çizgenin %60'lık en büyük kısmını protein düğümleri oluşturmaktadır. Fonksiyonel terim düğümleri ise çizgede yer alan en kapsamlı üçüncü düğüm tipi olup verinin %8'ini oluşturmaktadır. Fonksiyonel terim düğümlerinin %65'i (27.855 düğüm) biyolojik süreç, %26'sı (10.955 düğüm) moleküler fonksiyon ve kalan %9'u (4.075 düğüm) hücresel bileşen terimi düğümlerini kapsamaktadır. Her bir düğüm tipi için öznitelik vektörü elde edilebilen ve çizgenin son halinde yer alan düğüm sayısı, bu düğüm sayısının başlangıç verisindeki düğüm sayısına oranı, vektörün oluşturulmasında kullanılan yöntem, girdi verisi tipi, ve vektör boyutu Tablo 4.2.'de gösterilmiştir.

**Tablo 4.2.** Bilgi çizgesinde yer alan düğüm tipleri için oluşturulan öznitelik vektörlerinin genel özellikleri.

| Düğüm Tipi | Gömme Yöntemi | Girdi Verisi | Vektör Boyutu | Gömmesi Üretilebilen Düğüm Sayısı |
|---|---|---|---|---|
| **Protein** | TAPE (74) | Protein sekansı | 768 | 326,783 (%99) |
| **İlaç adayı bileşik** | SELformer (96) | SELFIES | 768 | 133,104 (%99) |
| **Fonksiyonel terim** | anc2Vec (75) | Ontoloji Yönlü Döngüsüz Çizgesi | 200 | 42,885 (%98) |
| **Domain** | dom2vec (77) | Protein sekansı | 50 | 10,229 (%100) |
| **Fenotipik terim** | CADA (83) | Gen-fenotip ilişki ağı | 160 | 8,971 (%97) |
| **İlaç** | TAPE (74) | Protein sekansı | 768 | 6,179 (%93) |
| **Hastalık** | doc2vec (97) | Hastalık tanımları | 100 | 5,694 (%100) |
| **EC numarası** | RXNFP (79) | Reaksiyon SMILES | 256 | 4,627 (%83) |
| **Biyolojik yolak** | BioKEEN (81) | Gen-biyolojik yolak etkileşim ağları | 200 | 3,993 (%99) |
| **Toplam** | | | | **542,465 (%99)** |

Gömmesi üretilebilen düğüm sayısının çizgede yer alan bu tipteki düğümlerin yüzde kaçını kapsadığı aynı sütunda parantez içinde verilmiştir.

### 4.1.3. CAFA3 Karşılaştırma Verisi İstatistikleri

Geliştirilen modelin tahmin performansının değerlendirilip diğer protein fonksiyon tahmini modelleri ile kıyaslanmasında CAFA3 karşılaştırma veri seti kullanılmıştır. Yarışmaya katılan diğer yöntemler tahminlerini Eylül 2016 tarihine kadar veri tabanlarında mevcut olan veri üzerinden üretip, Kasım 2017'de paylaşılan karşılaştırma veri seti üzerinden değerlendirmiştir. Bu yöntemlerle uygun bir karşılaştırma yapılabilmesi için aynı zaman çizelgesine uygun olacak şekilde eski tarihli veri tabanı versiyonlarından elde edilen verilerle ayrı bir çizge verisi oluşturulmuştur. Performans karşılaştırması modellerinin eğitim ve testinde kullanılan CAFA3 yarışma takvimine uygun veri setinde yer alan düğüm ve kenar tipi istatistikleri Tablo 4.3. ve Tablo 4.4.'de verilmiştir.

**Tablo 4.3.** CAFA3 yarışma takvimine uygun performans karşılaştırması modellerinin eğitim verisinde yer alan düğüm tipleri ve sayıları.

| Düğüm Tipi | Sayısı |
|---|---|
| **Protein** | 352,772 |
| **İlaç adayı bileşik** | 133,104 |
| **Fonksiyonel terim (Biyolojik süreç)** | 28,985 |
| **Fonksiyonel terim (Moleküler fonksiyon)** | 10,290 |
| **Fenotipik terim** | 8,971 |
| **Domain** | 6,755 |
| **İlaç** | 6,179 |
| **Hastalık** | 5,694 |
| **EC numarası** | 4,627 |
| **Fonksiyonel terim (Hücresel bileşen)** | 4,009 |
| **Biyolojik yolak** | 3,993 |
| **Toplam** | **565,379** |

"Fonksiyonel terim" şeklinde belirtilen düğüm tipi "biyolojik süreç", "moleküler fonksiyon" ve "hücresel bileşen" düğüm tiplerinin tümünü kapsamaktadır.

Bu veri, fonksiyon tahmini için geliştirilen asıl modellerin eğitiminde kullanılan güncel veri setinden (Bkz. Bölüm 4.1.1) farklı olup, yalnızca performans karşılaştırmasında kullanılacak bağımsız modellerin eğitim ve testi için hazırlanmıştır. Bahsedilen modellerin mimarisi birebir aynı olup sadece girdi olarak kullanılan veri setleri farklılık göstermektedir.

**Tablo 4.4.** CAFA3 yarışma takvimine uygun performans karşılaştırması modellerinin eğitim verisinde yer alan ilişki tipleri ve sayıları.

| İlişki Tipi | Sayısı |
|---|---|
| **Protein-Biyolojik süreç** | 552,109 |
| **Protein-Hücresel bileşen** | 366,132 |
| **Protein-Moleküler fonksiyon** | 337,869 |
| **Protein-Domain** | 218,687 |
| **Protein-EC numarası** | 214,227 |
| **Protein-İlaç adayı bileşik** | 160,499 |
| **Ortoloji** | 133,052 |
| **Protein-protein etkileşimi** | 91,453 |
| **Fonksiyonel terim-Fonksiyonel terim** | 89,330 |
| **Protein-Biyolojik yolak** | 55,695 |
| **Protein-Fenotipik terim** | 39,384 |
| **Hastalık-Fenotipik terim** | 24,259 |
| **Protein-İlaç** | 23,908 |
| **Protein-Hastalık** | 13,208 |
| **Domain-Fonksiyonel terim** | 6,119 |
| **Hastalık-Biyolojik yolak** | 1,682 |
| **Hastalık-İlaç** | 298 |
| **Toplam** | **2,327,911** |

"Fonksiyonel terim" şeklinde belirtilen düğüm tipi "biyolojik süreç", "moleküler fonksiyon" ve "hücresel bileşen" düğüm tiplerinin tümünü kapsamaktadır.

## 4.2. Ön Analiz Sonuçları

Tahmin modelinin tasarımı aşamasında farklı veri bölüm yöntemi, negatif örnekleme oranı ve kayıp fonksiyonu seçeneklerinin tahmin performansına etkisi araştırılmıştır. Seçimlerin tahmin performansına olan etkisinin adil bir şekilde karşılaştırılabilmesi için tüm denemeler belirli bir hiperparametre seti (öğrenme oranı= 0.0001, ağırlık sönümü=1e-5, saklı kanal sayısı=32, kafa sayısı=16, evrişim katmanı

sayısı=3, eğitim seti küme boyutu=256, eğitim seti küme komşu sayısı=32, döngü sayısı=50) kullanılarak yapılmıştır.

### 4.2.1. Negatif Örnekleme Oranı Seçimi

İdeal bir protein-fonksiyon tahmini senaryosunda bir proteinin tüm olası fonksiyonel terimler ile ilişkisi üzerine tahmin üretilmelidir. Bu çalışma kapsamında ele alınan protein setinin kapsamlı ve büyük bir veri olması sebebiyle tüm fonksiyonel terimler ile kombinasyonlarının incelenmesi kaynak açısından mümkün olmamaktadır. Bu sebeple ideale yakın bir tahmin senaryosu oluşturmak için veri setinde yer alan protein düğümleri %80/20 oranında eğitim ve test setlerine ayrılmış, her bir sette yer alan proteinlerin bilinen fonksiyonel terim ilişkileri pozitif olarak etiketlenmiş, ardından her bir pozitif ilişki için yüksek oranda negatif örnekleme yapılmıştır. Negatif örnekler, aralarındaki ilişki bilinmeyen protein-fonksiyonel terim çiftlerinden seçilmiştir. 1:20 ve 1:100 oranındaki negatif örnekleme senaryolarının tahmin performansına etkisi Tablo 4.5.'de gösterilmiştir. Daha yüksek olan negatif örnekleme oranı gerçek tahmin senaryosuna daha uygun olacağından model tasarımı ile ilgili denemelere 1:100 oranıyla devam edilmesine karar verilmiştir.

**Tablo 4.5.** Farklı negatif örnekleme oranı seçeneklerinin performansa etkisi.

| Fonksiyonel Terim Kategorisi | Negatif Örnekleme Oranı | F1 | AUPR | MCC |
|---|---|---|---|---|
| **Moleküler Fonksiyon** | 1:20 | **0.806** | **0.883** | **0.800** |
| | 1:100 | 0.317 | 0.616 | 0.416 |
| **Biyolojik Süreç** | 1:20 | **0.738** | **0.813** | **0.725** |
| | 1:100 | 0.243 | 0.532 | 0.341 |
| **Hücresel Bileşen** | 1:20 | **0.843** | **0.906** | **0.837** |
| | 1:100 | 0.657 | 0.718 | 0.666 |

Her GO kategorisi için en yüksek performans değerleri kalın metin ile belirginleştirilmiştir.

### 4.2.2. Kayıp Fonksiyonu Seçimi

Yüksek oranda negatif örnekleme yapıldığında veride oluşacak sınıf dengesizliği modelin baskın olan negatif sınıf yönde taraflı tahmin etmesine sebep olabilir. Bu durumun önüne geçmek için kayıp fonksiyonunda pozitif tahminlere daha

yüksek ağırlık vererek modelin pozitif örnekleri doğru tahmin etmeye yöneltilmesi düşünülmüştür. Ağırlıksız kayıp fonksiyonu ve negatif örnekleme oranınca (100) ağırlıklandırılmış kayıp fonksiyonu kullanılarak eğitilmiş model performansları Tablo 4.6.'de karşılaştırılmıştır. Moleküler fonksiyon ve biyolojik süreç kategorilerinde gözlemlenen performans artışı ve yüksek negatif örnekleme oranı kullanımı durumunda ağırlıklı kayıp fonksiyonu kullanmanın daha uygun olması sebebiyle model geliştirme sürecine ağırlıklı kayıp fonksiyonu kullanılarak devam edilmesine karar verilmiştir.

**Tablo 4.6.** Farklı kayıp fonksiyonu seçeneklerinin performansa etkisi.

| Fonksiyonel Terim Kategorisi | Kayıp Fonksiyonu | F1 | AUPR | MCC |
|---|---|---|---|---|
| **Moleküler Fonksiyon** | Ağırlıksız Kayıp Fonksiyonu | 0.317 | 0.616 | 0.416 |
| | Ağırlıklı Kayıp Fonksiyonu | **0.486** | **0.751** | **0.550** |
| **Biyolojik Süreç** | Ağırlıksız Kayıp Fonksiyonu | 0.243 | 0.532 | 0.341 |
| | Ağırlıklı Kayıp Fonksiyonu | **0.310** | **0.596** | **0.412** |
| **Hücresel Bileşen** | Ağırlıksız Kayıp Fonksiyonu | **0.657** | **0.718** | **0.666** |
| | Ağırlıklı Kayıp Fonksiyonu | 0.407 | 0.682 | 0.486 |

Her GO kategorisi için en yüksek performans değerleri kalın metin ile belirginleştirilmiştir.

### 4.2.3. Veri Bölme Yöntemi Seçimi

Kayıp fonksiyonu seçiminin ardından son olarak verinin eğitim ve test setlerine bölünmesinde kullanılabilecek iki farklı yöntemin performansa olan etkisinin incelenmiştir. Daha önceki deneyler için geliştirilen modellerin eğitiminde kullanılan protein-bazlı bölme yönteminde çizgede yer alan protein düğümleri istenilen oranında eğitim ve test setlerine bölünüp, her bir sette yalnızca bu sete atanan proteinlerin fonksiyonel terim ilişkilerinin bulunması sağlanır. Bu yöntemde test aşamasında tahmin üretilen proteinlerin hiçbir fonksiyonel terim ilişkisi eğitim aşamasında görülmez. Diğer bir seçenek olan kenar-bazlı bölme yönteminde ise protein-fonksiyonel terim ilişkileri istenilen oranlarda rastgele olarak eğitim ve test setlerine dağıtılır. Bu yöntemde yalnızca aynı kenarların eğitim ve test setinde yer almaması dikkate alınmış olup, eğitim setinde yer alan bir protein aynı zamanda test setinde yer

alabilir. Bu iki bölüm yöntemine göre eğitilmiş modellerin performansları Tablo 4.7.'de karşılaştırılmıştır. Protein fonksiyon tahmini yöntemleri özellikle yeni keşfedilen ve fonksiyonu daha önce hiç bilinmeyen proteinlerin fonksiyonlarını tahmin etmek üzere geliştirildiğinden model eğitimine bu senaryoya daha uygun olan protein-bazlı bölüm yöntemi ile devam edilmesine karar verilmiştir.

**Tablo 4.7.** Farklı veri bölme yöntemlerinin performansa etkisi.

| Fonksiyonel Terim Kategorisi | Veri Bölme Yöntemi | F1 | AUPR | MCC |
|---|---|---|---|---|
| **Moleküler Fonksiyon** | Kenar-bazlı Bölüm | 0.470 | 0.692 | 0.531 |
| | Protein-bazlı Bölüm | **0.486** | **0.751** | **0.550** |
| **Biyolojik Süreç** | Kenar-bazlı Bölüm | 0.239 | 0.589 | 0.350 |
| | Protein-bazlı Bölüm | **0.310** | **0.596** | **0.412** |
| **Hücresel Bileşen** | Kenar-bazlı Bölüm | 0.323 | 0.656 | 0.416 |
| | Protein-bazlı Bölüm | **0.407** | **0.682** | **0.486** |

Her GO kategorisi için en yüksek performans değerleri kalın metin ile belirginleştirilmiştir.

### 4.3. Hiperparametre Optimizasyonu

Geliştirilen geniş çaplı fonksiyon tahmini modeli için farklı tasarım seçeneklerinin değerlendirilip en uygun olanların belirlenmesinin ardından hiperparametre optimizasyonu yapılmıştır. Bu işlem heterojen çizge verisinin %80, %10 ve %10 olarak eğitim, validasyon ve test setlerine bölünmesi ile gerçekleştirilmiştir. Saklı kanal sayısı, öğrenme oranı, kafa sayısı, evrişim katmanı sayısı, eğitim seti küme boyutu, eğitim seti küme komşu sayısı ve ağırlık sönümü hiperparametreleri için belirlenen değer setleri üzerinden 25 döngü boyunca (Bkz. Bölüm 3.4) rastgele arama yöntemi uygulanmıştır. Validasyon seti kullanılarak hesaplanmış performans metriklerinden MCC değeri en yüksek olan modeller geniş çaplı tahmin üretimi için seçilip 100 döngü boyunca eğitilmiştir. Bu modellerin hiperparametre değerleri ve test seti üzerinden hesaplanan tahmin performansları Tablo 4.8.'de verilmiştir.

**Tablo 4.8.** Modeller için belirlenen optimal hiperparametre setleri ve bu hiperparametrelerle elde edilen performans değerleri.

| | **Moleküler Fonksiyon** | **Biyolojik Süreç** | **Hücresel Bileşen** |
|---|---|---|---|
| **Saklı kanal sayısı** | 32 | 128 | 32 |
| **Öğrenme oranı** | 0.005 | 0.001 | 0.005 |
| **Ağırlık sönümü** | 0 | 1e-05 | 0 |
| **Kafa sayısı** | 2 | 8 | 16 |
| **Evrişim katmanı sayısı** | 2 | 2 | 3 |
| **Eğitim seti küme boyutu** | 512 | 1024 | 2048 |
| **Eğitim seti küme komşu sayısı** | 32 | 256 | 256 |
| **F1** | 0.7844 | 0.6940 | 0.7443 |
| **AUPR** | 0.9073 | 0.8352 | 0.8429 |
| **MCC** | 0.7973 | 0.7154 | 0.7575 |

## 4.4. Ablasyon Analizi

Heterojen çizge verisindeki belirli düğümler ve bu düğümlere ait kenarların model öğrenmesi üzerindeki etkisinin gözlemlenebilmesi amacıyla bir ablasyon analizi gerçekleştirilmiştir.

### 4.4.1. Tek Düğüm Tipi Çıkartma Analizleri

Veriden belirli bir düğüm tipi ve bu düğüm tipine ait tüm kenarlar çıkartılarak kalan veri üzerinden modeller eğitilmiştir. Sonuçların birbiri ile karşılaştırılabilir olması için tüm modellerin eğitiminde, model tasarımı için olası seçeneklerin denenmesinde de kullanılmış olan sabit hiperparametre seti (öğrenme oranı= 0.0001, ağırlık sönümü=1e-5, saklı kanal sayısı=32, kafa sayısı=16, evrişim katmanı sayısı=3, eğitim seti küme boyutu=256, eğitim seti küme komşu sayısı=32, döngü sayısı=50) kullanılmıştır. Moleküler fonksiyon, biyolojik süreç ve hücresel bileşen modelleri için gerçekleştirilmiş ablasyon analizlerinin sonuçları sırasıyla Tablo 4.9., Tablo 4.10. ve Tablo 4.11.'de verilmiştir.

**Tablo 4.9.** Moleküler fonksiyon tahmini için tek düğüm tipi çıkartma ablasyon analizi sonuçları.

| Çıkartılan Düğüm Tipi | Çıkartılan Düğüm Sayısı | Çıkartılan Kenar Sayısı | F1 | AUPR | MCC |
|---|---|---|---|---|---|
| **Hastalık** | 5,694 | 39,056 | 0.5076 | 0.7083 | 0.5620 |
| **Fenotipik terim** | 8,971 | 62,498 | 0.5026 | 0.7061 | 0.5559 |
| **İlaç** | 6,179 | 23,016 | 0.4949 | 0.7372 | 0.5537 |
| **İlaç adayı bileşik** | 133,104 | 153,764 | **0.5558** | 0.6066 | **0.5690** |
| **Domain** | 10,229 | 467,459 | 0.4547 | 0.7326 | 0.5243 |
| **Biyolojik yolak** | 3,748 | 33,215 | 0.4524 | 0.7296 | 0.5204 |
| **Biyolojik yolak (KEGG)** | 245 | 20,159 | 0.5134 | 0.6910 | 0.5636 |
| **EC numarası** | 4,627 | 128,242 | 0.4336 | 0.7424 | 0.5088 |
| **Hücresel bileşen** | 4,075 | 678,951 | 0.4524 | 0.7290 | 0.5215 |
| **Biyolojik süreç** | 27,855 | 1,067,259 | 0.4038 | 0.7259 | 0.4864 |
| **Hiçbiri** | 0 | 0 | 0.4860 | **0.7510** | 0.5500 |

Biyolojik yolak düğümlerinden kaynağı Reactome (6) veri tabanı olanlar "Biyolojik yolak", KEGG (14) veri tabanı olanlar ise "Biyolojik yolak (KEGG)" olarak isimlendirilmiştir. En yüksek performans değerleri kalın metin ile belirginleştirilmiştir.

**Tablo 4.10.** Biyolojik süreç tahmini için tek düğüm tipi çıkartma ablasyon analizi sonuçları.

| Çıkartılan Düğüm Tipi | Çıkartılan Düğüm Sayısı | Çıkartılan Kenar Sayısı | F1 | AUPR | MCC |
|---|---|---|---|---|---|
| **Hastalık** | 5,694 | 39,056 | 0.3215 | 0.5916 | 0.4197 |
| **Fenotipik terim** | 8,971 | 62,498 | 0.2830 | 0.5778 | 0.3888 |
| **İlaç** | 6,179 | 23,016 | 0.3302 | 0.6146 | **0.4290** |
| **İlaç adayı bileşik** | 133,104 | 153,764 | 0.3209 | 0.5965 | 0.4190 |

**Tablo 4.10.** Biyolojik süreç tahmini için tek düğüm tipi çıkartma ablasyon analizi sonuçları (Devamı).

| Çıkartılan Düğüm Tipi | Çıkartılan Düğüm Sayısı | Çıkartılan Kenar Sayısı | F1 | AUPR | MCC |
|---|---|---|---|---|---|
| **Domain** | 10,229 | 467,459 | 0.3113 | 0.6059 | 0.4143 |
| **Biyolojik yolak** | 3,748 | 33,215 | 0.2920 | **0.6310** | 0.4000 |
| **Biyolojik yolak (KEGG)** | 245 | 20,159 | 0.3308 | 0.6087 | 0.4281 |
| **EC numarası** | 4,627 | 128,242 | 0.3065 | 0.5947 | 0.4093 |
| **Hücresel bileşen** | 4,075 | 678,951 | **0.3340** | 0.6012 | 0.4267 |
| **Moleküler fonksiyon** | 10,955 | 943,347 | 0.3051 | 0.6086 | 0.4081 |
| **Hiçbiri** | 0 | 0 | 0.3100 | 0.5960 | 0.4120 |

Biyolojik yolak düğümlerinden kaynağı Reactome (6) veri tabanı olanlar "Biyolojik yolak", KEGG (14) veri tabanı olanlar ise "Biyolojik yolak (KEGG)" olarak isimlendirilmiştir. En yüksek performans değerleri kalın metin ile belirginleştirilmiştir.

**Tablo 4.11.** Hücresel bileşen tahmini için tek düğüm tipi çıkartma ablasyon analizi sonuçları.

| Çıkartılan Düğüm Tipi | Çıkartılan Düğüm Sayısı | Çıkartılan Kenar Sayısı | F1 | AUPR | MCC |
|---|---|---|---|---|---|
| **Hastalık** | 5,694 | 39,056 | 0.4887 | 0.6930 | 0.5434 |
| **Fenotipik terim** | 8,971 | 62,498 | 0.4780 | 0.7126 | 0.5403 |
| **İlaç** | 6,179 | 23,016 | 0.4149 | 0.7215 | 0.4948 |
| **İlaç adayı bileşik** | 133,104 | 153,764 | 0.3271 | 0.7196 | 0.4278 |
| **Domain** | 10,229 | 467,459 | 0.4546 | 0.6932 | 0.5199 |
| **Biyolojik yolak** | 3,748 | 33,215 | 0.4728 | 0.6970 | 0.5316 |
| **Biyolojik yolak (KEGG)** | 245 | 20,159 | 0.4064 | 0.6889 | 0.4866 |
| **EC numarası** | 4,627 | 128,242 | 0.3413 | 0.6624 | 0.4355 |

**Tablo 4.11.** Hücresel bileşen tahmini için tek düğüm tipi çıkartma ablasyon analizi sonuçları (Devamı).

| Çıkartılan Düğüm Tipi | Çıkartılan Düğüm Sayısı | Çıkartılan Kenar Sayısı | F1 | AUPR | MCC |
|---|---|---|---|---|---|
| **Biyolojik süreç** | 27,855 | 1,067,259 | 0.4487 | **0.7385** | 0.5215 |
| **Moleküler fonksiyon** | 10,955 | 943,347 | **0.4943** | 0.7265 | **0.5545** |
| **Hiçbiri** | 0 | 0 | 0.4070 | 0.6820 | 0.4860 |

Biyolojik yolak düğümlerinden kaynağı Reactome (6) veri tabanı olanlar "Biyolojik yolak", KEGG (14) veri tabanı olanlar ise "Biyolojik yolak (KEGG)" olarak isimlendirilmiştir. En yüksek performans değerleri kalın metin ile belirginleştirilmiştir.

Veriden çıkartıldığında model performansında artış sağlayan düğüm tiplerinde bu durumun belirlenen hiperparametre setinin bu veriye daha uygun olmasından kaynaklanıp kaynaklanmadığının anlaşılabilmesi için hiperparametre optimizasyonu uygulanmıştır. Her bir GO kategorisi için sabit hiperparametre seti ile en yüksek MCC değerine ulaşan 2 ablasyon modeli optimize edilmiştir. Sonuçlar Tablo 4.12.'de verilmiştir.

**Tablo 4.12.** Optimize edilmiş ablasyon modellerinin performans sonuçları.

| Tahmin Modeli | Çıkartılan Düğüm Tipi | F1 | AUPR | MCC |
|---|---|---|---|---|
| **Moleküler Fonksiyon** | **İlaç adayı bileşik** | 0.6313 | 0.7381 | 0.6500 |
| | **Biyolojik yolak (KEGG)** | 0.6300 | 0.8213 | 0.6675 |
| | **Hiçbiri** | **0.7844** | **0.9073** | **0.7973** |
| **Biyolojik Süreç** | **İlaç** | 0.3911 | 0.6489 | 0.4686 |
| | **Biyolojik yolak (KEGG)** | 0.4683 | 0.8181 | 0.5454 |
| | **Hiçbiri** | **0.6940** | **0.8352** | **0.7154** |
| **Hücresel Bileşen** | **Moleküler fonksiyon** | 0.5802 | 0.8414 | 0.6324 |
| | **Hastalık** | 0.4983 | 0.8282 | 0.5668 |
| | **Hiçbiri** | **0.7443** | **0.8429** | **0.7575** |

Her GO kategorisi için en yüksek performans değerleri kalın metin ile belirginleştirilmiştir.

### 4.4.2. Tek Düğüm Tipi Kullanma Analizleri

Veride yalnızca belirli bir düğüm tipi ve bu düğüm tipine ait tüm kenarlar kalacak şekilde geri kalan tüm düğüm ve kenar tiplerinin çıkartılması ile elde edilen veri üzerinden modeller eğitilmiştir. Bölüm 4.4.1.'de gerçekleştirilen analizlerde sabit hiperparametre seti üzerinden kıyaslama yapmanın yanıltıcı sonuçlara sebep olabileceği gözlemlendiğinden her ablasyon modeli için hiperparametre optimizasyonu yapılmıştır. Model performans sonuçları her üç kategori için Tablo 4.13, Tablo 4.14 ve Tablo 4.15'de verilmiştir.

**Tablo 4.13.** Moleküler fonksiyon tahmini için tek düğüm tipi kullanma ablasyon analizi sonuçları.

| Kullanılan Düğüm Tipi | Kullanılan Düğüm Sayısı | Kullanılan Kenar Sayısı | F1 | AUPR | MCC |
|---|---|---|---|---|---|
| **Hastalık** | 358,827 | 965,900 | 0.5692 | 0.8493 | 0.6253 |
| **Fenotipik terim** | 352,403 | 989,342 | 0.5865 | 0.8104 | 0.6337 |
| **İlaç** | 349,611 | 949,860 | 0.5867 | 0.8599 | 0.6395 |
| **İlaç adayı bileşik** | 470,842 | 1,080,608 | 0.4913 | 0.8247 | 0.5615 |
| **Domain** | 379,897 | 1,394,303 | 0.6568 | 0.8898 | 0.6931 |
| **Biyolojik yolak** | 341,486 | 960,059 | 0.5876 | 0.8523 | 0.6375 |
| **Biyolojik yolak (KEGG)** | 343,677 | 947,003 | 0.5707 | 0.8590 | 0.6352 |
| **EC numarası** | 342,365 | 1,055,086 | 0.5962 | 0.9001 | 0.6475 |
| **Hücresel bileşen** | 352,042 | 1,605,795 | 0.6026 | 0.8609 | 0.6503 |
| **Biyolojik süreç** | 375,822 | 1,994,103 | 0.7306 | 0.9015 | 0.7513 |
| **Moleküler fonksiyon** | 347,967 | 943,347 | 0.5772 | 0.8855 | 0.6320 |
| **Tüm düğüm tipleri** | 542,465 | 3,792,832 | **0.7844** | **0.9073** | **0.7973** |

Biyolojik yolak düğümlerinden kaynağı Reactome (6) veri tabanı olanlar "Biyolojik yolak", KEGG (14) veri tabanı olanlar ise "Biyolojik yolak (KEGG)" olarak isimlendirilmiştir. En yüksek performans değerleri kalın metin ile belirginleştirilmiştir.

**Tablo 4.14.** Biyolojik süreç tahmini için tek düğüm tipi kullanma ablasyon analizi sonuçları.

| Kullanılan Düğüm Tipi | Kullanılan Düğüm Sayısı | Kullanılan Kenar Sayısı | F1 | AUPR | MCC |
|---|---|---|---|---|---|
| **Hastalık** | 375,727 | 1,040,008 | 0.3621 | 0.7718 | 0.4660 |
| **Fenotipik terim** | 369,303 | 1,063,450 | 0.3986 | 0.8079 | 0.4961 |
| **İlaç** | 366,511 | 1,023,968 | 0.4312 | 0.8244 | 0.5223 |
| **İlaç adayı bileşik** | 487,742 | 1,154,716 | 0.4625 | 0.7898 | 0.5469 |
| **Domain** | 379,897 | 1,468,411 | 0.3577 | 0.7932 | 0.4623 |
| **Biyolojik yolak** | 358,386 | 1,034,167 | 0.4294 | 0.7474 | 0.5192 |
| **Biyolojik yolak (KEGG)** | 360,577 | 1,021,111 | 0.4716 | 0.7813 | 0.5517 |
| **EC numarası** | 359,265 | 1,129,194 | 0.4667 | **0.8688** | 0.5506 |
| **Hücresel bileşen** | 368,942 | 1,679,903 | 0.4012 | 0.7032 | 0.4917 |
| **Biyolojik süreç** | 364,867 | 1,067,259 | 0.4168 | 0.7499 | 0.5171 |
| **Moleküler fonksiyon** | 375,822 | 1,944,299 | 0.4145 | 0.7211 | 0.5041 |
| **Tüm düğüm tipleri** | 542,465 | 3,792,832 | **0.6940** | 0.8352 | **0.7154** |

Biyolojik yolak düğümlerinden kaynağı Reactome (6) veri tabanı olanlar "Biyolojik yolak", KEGG (14) veri tabanı olanlar ise "Biyolojik yolak (KEGG)" olarak isimlendirilmiştir. En yüksek performans değerleri kalın metin ile belirginleştirilmiştir.

**Tablo 4.15.** Hücresel bileşen tahmini için tek düğüm tipi kullanma ablasyon analizi sonuçları.

| Kullanılan Düğüm Tipi | Kullanılan Düğüm Sayısı | Kullanılan Kenar Sayısı | F1 | AUPR | MCC |
|---|---|---|---|---|---|
| **Hastalık** | 351,947 | 711,195 | 0.4576 | 0.7849 | 0.5342 |
| **Fenotipik terim** | 345,523 | 734,637 | 0.4337 | 0.8134 | 0.5146 |
| **İlaç** | 342,731 | 695,155 | 0.4502 | 0.7514 | 0.5248 |
| **İlaç adayı bileşik** | 463,962 | 825,903 | 0.4291 | 0.8060 | 0.5119 |

**Tablo 4.15.** Hücresel bileşen tahmini için tek düğüm tipi kullanma ablasyon analizi sonuçları (Devamı).

| | | | | | |
|---|---|---|---|---|---|
| **Domain** | **379,897** | **1,139,598** | 0.4812 | 0.8379 | 0.5536 |
| **Biyolojik yolak** | 334,606 | 705,354 | 0.4258 | 0.8205 | 0.5098 |
| **Biyolojik yolak (KEGG)** | 336,797 | 692,298 | 0.4710 | 0.8096 | 0.5439 |
| **EC numarası** | 335,485 | 800,381 | 0.4632 | 0.7976 | 0.5387 |
| **Hücresel bileşen** | 341,087 | 678,951 | 0.4984 | 0.7893 | 0.5649 |
| **Biyolojik süreç** | 368,942 | 1,739,398 | 0.4789 | 0.7868 | 0.5491 |
| **Moleküler fonksiyon** | 352,042 | 1,615,486 | 0.4517 | 0.8161 | 0.5307 |
| **Tüm düğüm tipleri** | 542,465 | 3,792,832 | **0.7443** | **0.8429** | **0.7575** |

Biyolojik yolak düğümlerinden kaynağı Reactome 6) veri tabanı olanlar "Biyolojik yolak", KEGG (14) veri tabanı olanlar ise "Biyolojik yolak (KEGG)" olarak isimlendirilmiştir. En yüksek performans değerleri kalın metin ile belirginleştirilmiştir.

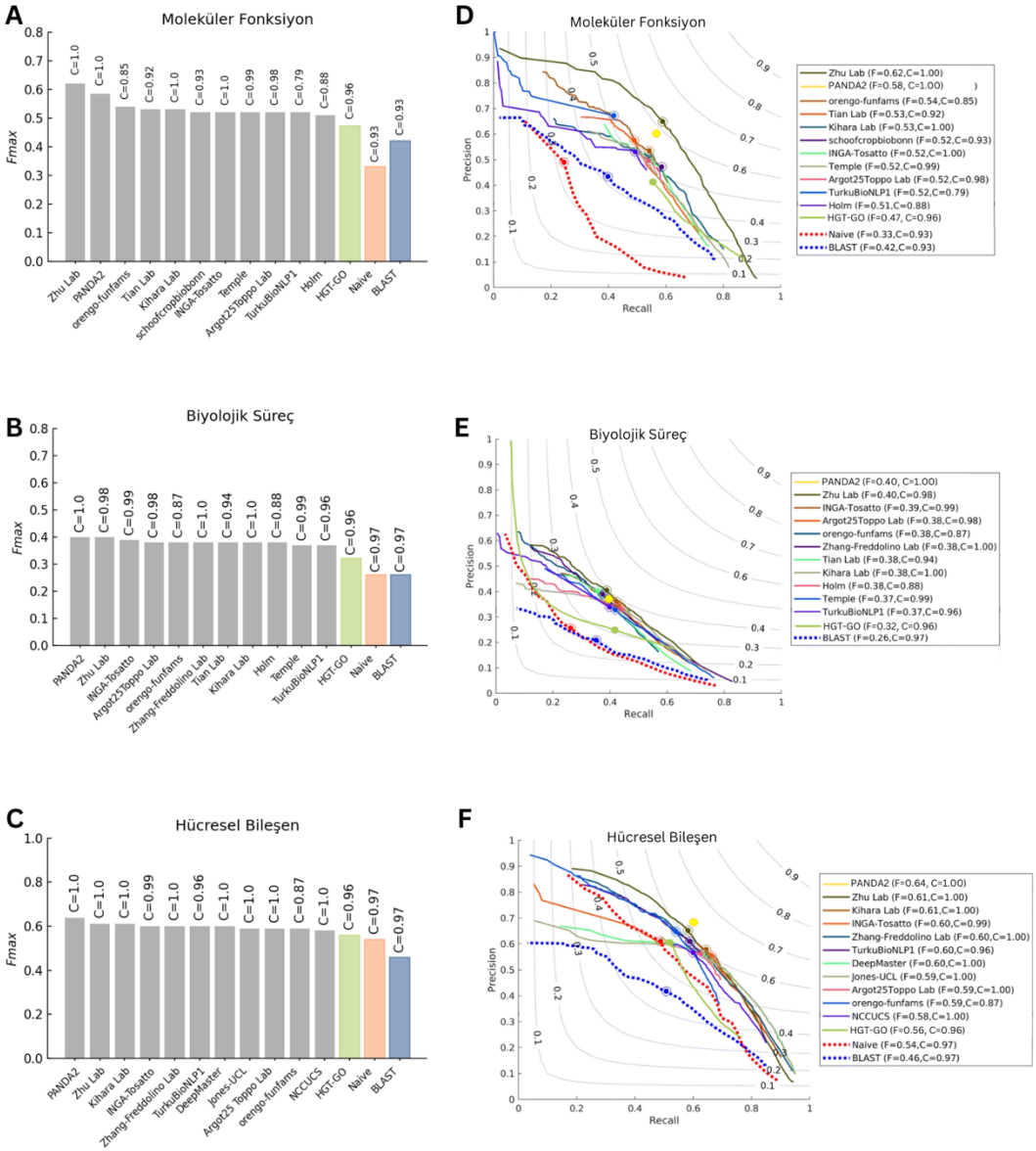
### 4.5. Benzer Metodlar ile Tahmin Performansı Karşılaştırması

Geliştirilen modelin tahmin performansı CAFA yarışmasının 3. versiyon karşılaştırma veri seti (87) üzerinden değerlendirilmiş ve benzer protein fonksiyon tahmini yöntemleri ile karşılaştırılmıştır. Karşılaştırma veri seti, daha önce deneysel anotasyonu bulunmayıp CAFA3 zaman çizelgesi içerisinde deneysel anotasyon edinmiş olan proteinleri içeren katı bir veri setidir. Performans değerlendirmesi tüm karşılaştırma veri seti üzerinden, yani "tam mod"da (*full mode*) gerçekleştirilmiştir.

Moleküler fonksiyon, biyolojik süreç ve hücresel bileşen kategorilerinde ulaşılan tahmin performansı Fmax skoru üzerinden sırasıyla 0.473, 0.321 ve 0.56'dır. Tahmin performansları Smin skoru üzerinden incelendiğinde ise 6.545, 15.953 ve 6.500 değerlerine ulaşılmıştır.

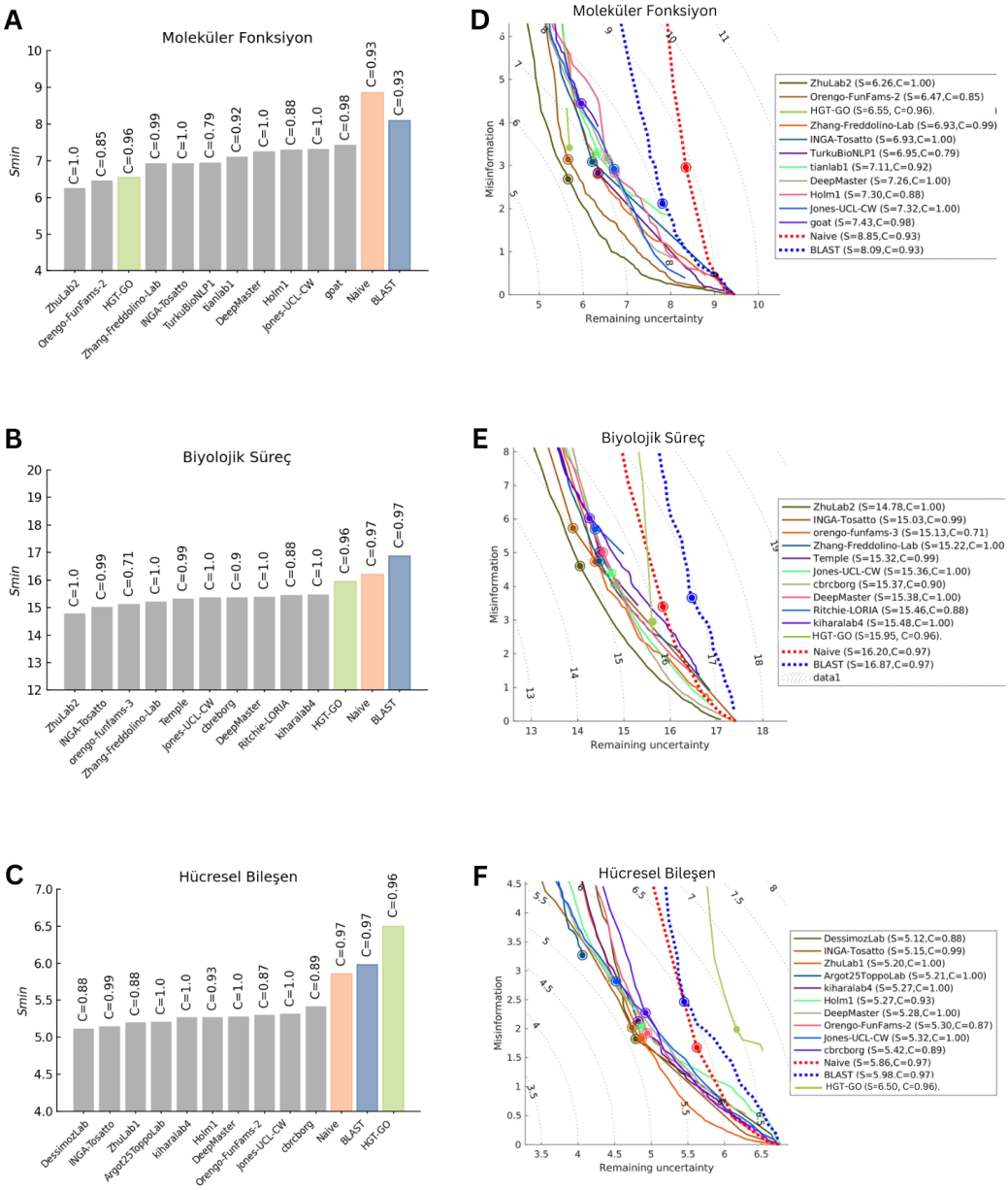
Elde edilen performans değerleri, CAFA3 yarışmasına katılan yöntemler arasında her fonksiyonel terim kategorisi için en yüksek performansa ulaşan 10 model, çizge tabanlı bir tahmin modeli ve 2 temel tahmin modeli ile karşılaştırılmıştır. Fmax

skoru ve kesinlik-doğruluk (*precision-recall*, PR) eğrisi üzerinden karşılaştırma sonuçları Şekil 4.3.'te; Smin skoru ve kalan belirsizlik-yanlış bilgi (*remaining uncertainty-misinformation*, RU-MI) eğrisi üzerinden karşılaştırma sonuçları ise Şekil 4.4.'te verilmiştir. Modelimiz her iki şekilde de "HGT-GO" ismi ile tanımlanmıştır.




**Şekil 4.3.** 3 fonksiyonel terim kategorisi için CAFA3 karşılaştırma veri seti üzerinden fonksiyon tahmini Fmax skoru ve PR eğrisi sonuçları. **A-C:** Tüm modellerin fonksiyon tahmini Fmax skorlarını göstermektedir. Daha yüksek Fmax skoru, daha yüksek performansa işaret eder. Modellerin kapsama oranı (karşılaştırma veri setinde yer alan proteinlerden yüzde

kaçı için tahmin üretilebildiği) sütunların üzerinde verilmiştir. **D-F:** Tüm modellerin fonksiyon tahmini PR eğrilerini göstermektedir. Modellere ait eğrilerin üzerindeki nokta, Fmax skoruna ulaşılan noktaya işaret etmektedir. İdeal bir tahmin metodu grafiğin sağ üst köşesinde yer alan Fmax=1 değerine ulaşmalıdır.




**Şekil 4.4.** 3 fonksiyonel terim kategorisi için CAFA3 karşılaştırma veri seti üzerinden fonksiyon tahmini Smin skoru ve RU-MI eğrisi sonuçları. **A-C:** Tüm modellerin fonksiyon tahmini Smin skorlarını göstermektedir. Daha düşük Smin skoru, daha yüksek performansa işaret eder. Modellerin kapsama oranı (karşılaştırma veri setinde yer alan proteinlerden yüzde kaçı için tahmin üretilebildiği) sütunların üzerinde verilmiştir. **D-F:** Tüm

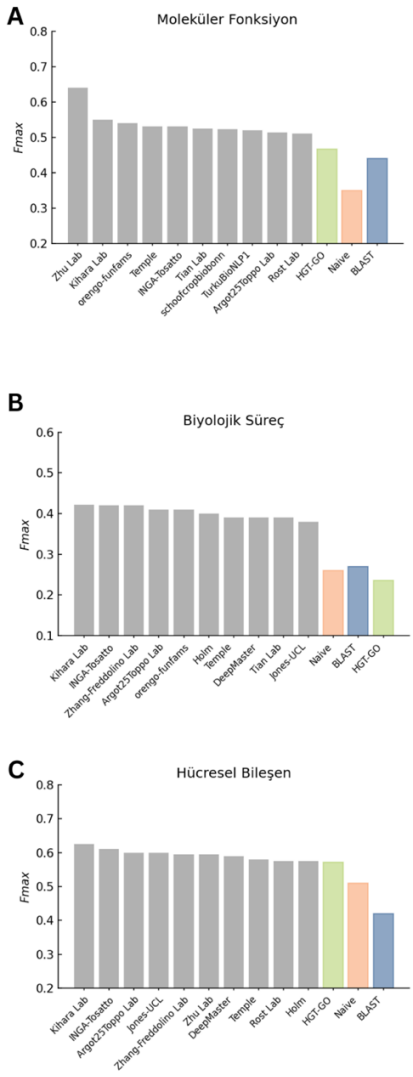
modellerin fonksiyon tahmini RU-MI eğrilerini göstermektedir. Modellere ait eğrilerin üzerindeki nokta, Smin skoruna ulaşılan noktaya işaret etmektedir. İdeal bir tahmin metodu grafiğin sol alt köşesinde yer alan Smin=0 değerine ulaşmalıdır.

Ardından modellerin organizmalara özgü protein fonksiyon tahmin performansı ayrı ayrı değerlendirilmiştir. Karşılaştırma veri setinde bulunan farklı organizmalara ait proteinler için elde edilen performans değerleri Tablo 4.16'de gösterilmiştir.

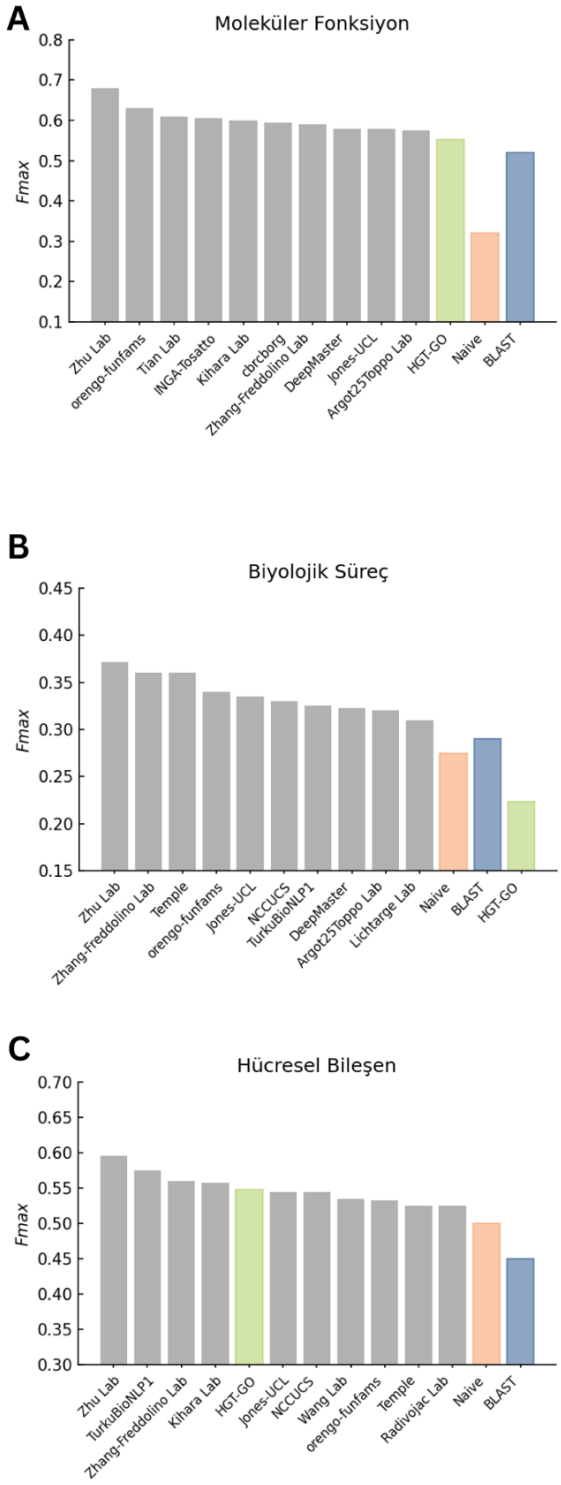
**Tablo 4.16.** 3 fonksiyonel terim kategorisi için CAFA3 karşılaştırma veri seti üzerinden organizmaya özgü fonksiyon tahmini performans değerleri.

| Tahmin Modeli | Organizma | Veri setindeki protein sayısı | Fmax | Smin |
|---|---|---|---|---|
| **Moleküler Fonksiyon** | *Pseudomonas putida* | 5 | 0.979 | 0.599 |
| | *Mycoplasmoides genitalium* | 3 | 0.843 | 2.501 |
| | *Methanocaldococcus jannaschii* | 4 | 0.703 | 4.943 |
| | *Mus musculus* | 59 | 0.553 | 4.642 |
| | *Xenopus laevis* | 7 | 0.504 | 7.817 |
| | *Rattus norvegicus* | 15 | 0.481 | 10.879 |
| | *Homo sapiens* | 120 | 0.467 | 5.490 |
| **Biyolojik Süreç** | *Pseudomonas putida* | 5 | 0.919 | 0.487 |
| | *Schizosaccharomyces pombe* | 5 | 0.488 | 9.333 |
| | *Xenopus laevis* | 7 | 0.451 | 14.186 |
| | *Rattus norvegicus* | 15 | 0.246 | 15.608 |
| | *Homo sapiens* | 120 | 0.236 | 15.917 |
| | *Mus musculus* | 59 | 0.223 | 19.279 |
| **Hücresel Bileşen** | *Methanocaldococcus jannaschii* | 4 | 0.934 | 0.693 |
| | *Xenopus laevis* | 7 | 0.795 | 4.892 |
| | *Homo sapiens* | 120 | 0.571 | 5.471 |
| | *Mus musculus* | 59 | 0.548 | 7.663 |
| | *Rattus norvegicus* | 15 | 0.445 | 10.134 |

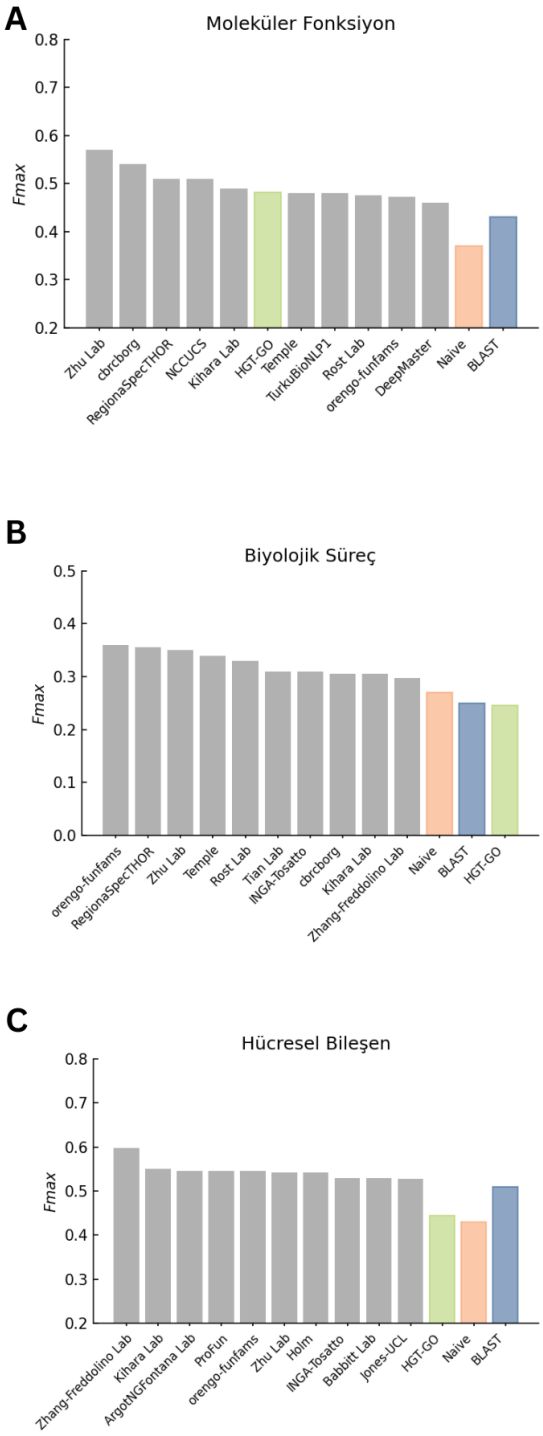
Organizmaya özgü performans sonuçları CAFA3 yarışmasına katılan yöntemler arasında her fonksiyonel terim kategorisi için en yüksek performansa ulaşan 10 model ve 2 temel tahmin modeli ile karşılaştırılmıştır. Duyurulan yarışma sonuçları arasında yalnızca karşılaştırma veri setinde en az 15 proteini bulunan organizmalara yer verildiğinden (87) tüm organizmalar için karşılaştırılma yapılamamıştır. Buna göre *Homo sapiens*, *Mus musculus* ve *Rattus norvegicus* türleri için Fmax skoru üzerinden performans karşılaştırması sırasıyla Şekil 4.5, Şekil 4.6 ve Şekil 4.7'te gösterilmiştir.




**Şekil 4.5.** 3 fonksiyonel terim kategorisi için CAFA3 karşılaştırma veri setinde yer alan *Homo sapiens* türü proteinleri özelinde fonksiyon tahmini Fmax skoru sonuçları. Daha yüksek Fmax skoru, daha yüksek performansa işaret eder.




**Şekil 4.6.** 3 fonksiyonel terim kategorisi için CAFA3 karşılaştırma veri setinde yer alan *Mus musculus* türü proteinleri özelinde fonksiyon tahmini Fmaxskoru sonuçları. Daha yüksek Fmax skoru, daha yüksek performansa işaret eder.




**Şekil 4.7.** 3 fonksiyonel terim kategorisi için CAFA3 karşılaştırma veri setinde yer alan *Rattus norvegicus* türü proteinleri özelinde fonksiyon tahmini Fmax skoru sonuçları. Daha yüksek Fmax skoru, daha yüksek performansa işaret eder.

### 4.6. Seçili Tahminlerin Biyolojik Anlamlılığı

Bu tez kapsamında hakkında üretilen tahminlerin biyolojik anlamlılığının incelenmesi için literatürde çok sayıda araştırmaya konu olmuş iki protein seçilmiştir. Seçili proteinlerden ilki olan Tensin-2 (UniProt Tanımlayıcısı: Q63HR2) için modelimizin ürettiği 873 tahminin 816'sı doğru-negatif, 34'ü yanlış-pozitif, 22'si doğru-pozitif ve 1'i yanlış negatif olarak değerlendirilmiştir. Tensin-2 ile ilişkili olduğu tahmin edilen fonksiyonel terimler arasından literatür taraması için seçilenler Tablo 4.17.'te verilmiştir. Tahmin olasılıklarına göre sıralandığında ilk beş tahminden dördünün manuel anotasyonunun da bulunduğu görülmüştür. Bu fonksiyonel terimler arasındaki anlamsal hiyerarşik ilişkiler ise Şekil 4.8.A'da gösterilmiştir.

**Tablo 4.17.** Modelin Tensin-2 proteini ile ilişkilendirdiği seçili fonksiyonel terim tahminleri.

| GO Terimi | Terim Kategorisi | Olasılık Tahmini | Manuel Anotasyon Durumu |
|---|---|---|---|
| insülin reseptörü sinyal yolunun negatif düzenlenmesi (GO:0046627) | B | 0.998 | ✓ |
| endozom membranı (GO:0010008) | H | 0.998 | |
| plazma membranı (GO:0005886) | H | 0.996 | ✓ |
| fosfoprotein fostataz aktivitesi (GO:0004721) | M | 0.995 | ✓ |
| peptidil-tirozin defosforilasyonu (GO:0035335) | B | 0.991 | ✓ |
| insülin reseptör sinyal yolunun regülasyonu (GO:0046626) | B | 0.987 | |
| protein defosforilasyonunun pozitif regülasyonu (GO:0035307) | B | 0.979 | |
| protein tirozin fosfataz aktivitesi (GO:0004725) | M | 0.932 | ✓ |
| insülin benzeri büyüme faktörü reseptör sinyal yolağının negatif regülasyonu (GO:0043569) | B | 0.698 | |
| büyüme faktörü bağlanması (GO:0019838) | M | 0.679 | |

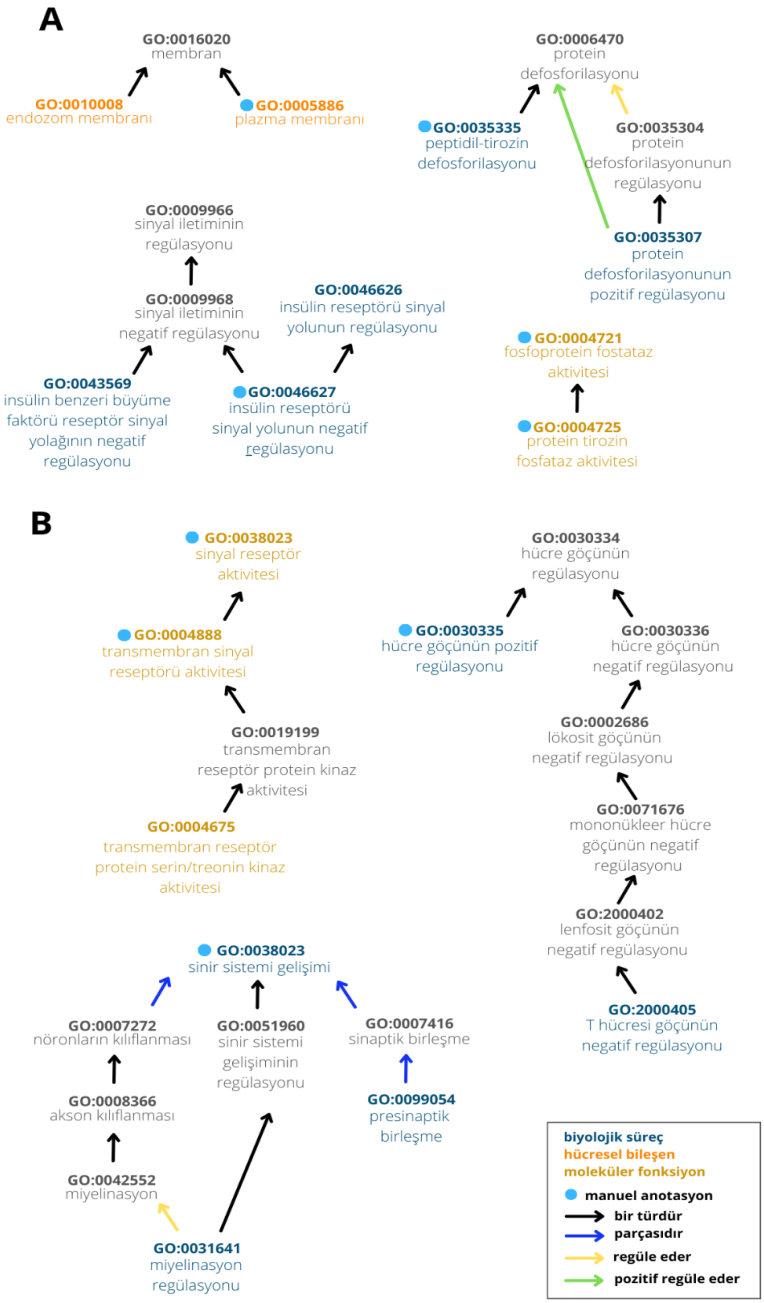
B: biyolojik süreç, M: moleküler fonksiyon, H: hücresel bileşen.

Araştırma için seçilen bir diğer protein Semaphorin-4D (Uniprot Erişim ID: O09126) için modelin ürettiği 827 tahminin 769'sı doğru-negatif, 40'ı doğru-pozitif, 17'si yanlış-pozitif ve 1'i yanlış-negatif olarak değerlendirilmiştir. Semaphorin-4D ile ilişkili olduğu tahmin edilen fonksiyonel terimler arasından literatür taraması için seçilenler Tablo 4.18.'te verilmiştir. Bu tahminler olasılıklarına göre sıralandığında ilk beş tahmin protein-fonksiyonel terim ilişkisinin manuel anotasyonunun bulunduğu görülmüştür. Bu fonksiyonel terimler arasındaki anlamsal hiyerarşik ilişkiler ise Şekil 4.8.B'de gösterilmiştir.

**Tablo 4.18.** Modelin Semaphorin-4D proteini ile ilişkilendirdiği seçili fonksiyonel terim tahminleri.

| GO Terimi | Terim Kategorisi | Olasılık Tahmini | Manuel Anotasyon Durumu |
|---|---|---|---|
| transmembran sinyal reseptörü aktivitesi (GO:0004888) | M | 0.999 | ✓ |
| hücre göçünün pozitif regülasyonu (GO:0030335) | B | 0.999 | ✓ |
| sinir sistemi gelişimi (GO:0007399) | B | 0.999 | ✓ |
| hücre dışı alan (GO:0005615) | H | 0.999 | ✓ |
| sinyal reseptörü aktivitesi (GO:0038023) | M | 0.997 | ✓ |
| transmembran reseptör protein serin/treonin kinaz aktivitesi (GO:0004675) | M | 0.905 | |
| miyelinasyon regülasyonu (GO:0031641) | B | 0.810 | |
| T hücresi göçünün negatif regülasyonu (GO:2000405) | B | 0.707 | |
| nükleer zarf (GO:0005635) | H | 0.672 | |
| presinaptik birleşme (GO:0099054) | B | 0.539 | |

B: biyolojik süreç, M: moleküler fonksiyon, H: hücresel bileşen.




**Şekil 4.8.** Seçili proteinlerle ilişkilendirilmiş fonksiyonel terim tahminleri arasındaki anlamsal hiyerarşik ilişkiler. (A) Tensin-2 proteini ile ilişkilendirilmiş fonksiyonel terimler. (B) Semaphorin-4D proteini ile ilişkilendirilmiş fonksiyonel terimler. Tahminler arasında yer almayan fakat anlamsal bağlılığı ifade etmek için figürde verilmesi gereken terimler gri yazı ile belirtilmiştir. Tahminler arasında yer alan terimler ise kategorisine göre sağ alt köşedeki açıklamalarda belirtilen renkle işaretlenmiştir.

# 5. TARTIŞMA

Çalışmada bilinmeyen proteinlerin fonksiyon tahmini için heterojen çizge dönüştürücü mimarisi tabanlı tahmin modelleri önerilmiştir. Modellerin eğitimi için 14 farklı biyomedikal veri tabanından elde ve entegre edilen, 9 düğüm ve 17 kenar tipinden oluşan heterojen çizge verisi kullanılmıştır. Moleküler fonksiyon, biyolojik süreç ve hücresel bileşen fonksiyonel terim kategorilerinin her biri için ayrı bir tahmin modeli eğitilmiştir. Eğitilen modellere ait performans sonuçları ve karşılaştırmaları Bulgular bölümünde gösterilmiştir.

Olası model tasarımı seçenekleri arasında karar verebilmek için farklı negatif öğrenme oranları, kayıp fonksiyonu ve veri bölüm yöntemi seçenekleri kullanıldığında elde edilen performanslar birbiri ile karşılaştırılmıştır. 1:20 ve 1:100 oranında negatif örnekleme yapılan modeller üzerinden yapılan performans karşılaştırmalarında tüm fonksiyonel terim kategorilerinde negatif örnekleme oranı arttırıldığında performansta düşüş gözlemlenmiştir (Tablo 4.5.). Negatif örnekleme oranının artması tahmin görevini zorlaştırdığından bu durum beklenen bir sonuçtur. 1:20 negatif örnekleme ile eğitilen model sonuçlarında tüm kategorilerde F1-skor değerlerinin MCC değerlerine göre genellikle daha yüksek olduğu gözlemlenmektedir. Negatif örnekleme oranı arttırıldığında ise MCC değerlerinin F1-skor değerlerine göre daha yüksek olduğu görülmüştür. Bu durum, MCC'nin F1-skorun aksine doğru negatif tahmin sayısını da hesaba katmasından kaynaklanmaktadır. Negatif örnekleme oranı 1:20'den 1:100'e çıkartıldığında veri setindeki negatif örnek sayısı, dolayısıyla doğru ve yanlış negatif tahmin sayısı önemli ölçüde artmaktadır. Örneğin moleküler fonksiyon kategorisinde 1:20 negatif örnekleme yapılan modelde doğru pozitif, yanlış pozitif, doğru negatif ve yanlış negatif tahmin sayıları 137180, 17896, 3689764 ve 48203 iken 1:100 negatif örnekleme uygulanan modelde bu değerler 35502, 3430, 18534870 ve 149881'dir. Bunlardan sadece yanlış negatif tahmin sayısı dikkate alınarak hesaplanan (Formül 3.7.) F1-skor değerlerinde, 1:100 negatif örneklemede yanlış negatif tahmin sayısının artışı sebebiyle büyük bir düşüş gözlemlenmiştir. MCC değerleri ise doğru negatif tahminleri de hesaba kattığından (Formül 3.8.) F1-skordakine göre daha az düşüş göstermiştir. MCC, hata matrisindeki tüm tahmin kategorilerindeki sayıları göz önünde bulundurarak sınıflandırma performansına dair daha bütüncül bir

değerlendirme sunmaktadır. Bu sebeple çalışmadaki tüm performans değerlendirmelerinde birincil metrik olarak MCC ele alınmıştır.

Pozitif tahminlerinin eğitime katkısını negatif örnekleme oranınca ağırlıklandırmanın performansa katkısını gözlemlemek için ağırlıklı ve ağırlıksız kayıp fonksiyonları ile eğitilen modeller karşılaştırılmıştır. Pozitif tahminlerin ağırlıklandırıldığı bir kayıp fonksiyonu kullanıldığında model pozitif örnekleri doğru sınıflandırmaya odaklanmaktadır. Dolayısıyla tüm fonksiyonel terim kategorilerinde doğru-negatif ve yanlış-negatif tahminlerin sayısında düşüş, doğru-pozitif ve yanlış-pozitif tahminlerin sayısında ise artış gözlemlenmiştir (Tablo 4.6.). Bu değişimlerin oranına bağlı olarak moleküler fonksiyon ve biyolojik süreç kategorilerinde ağırlıklı kayıp fonksiyonu kullanıldığında performansta artış görülürken, hücresel bileşen kategorisinde ise performans düşmüştür.

Verinin eğitim ve test setlerine bölümünde protein-bazlı ve kenar-bazlı ayırma yapmanın performansa etkisi incelenmiştir. Protein-bazlı bölünmüş eğitim ve test setleri ile eğitilen modellerde performansın kenar-bazlı bölünmüş eğitim ve test setleri ile eğitilen modellere kıyasla daha yüksek olması ilginç bir gözlemdir (Tablo 4.7.). Protein-bazlı bölümde test setinde bulunan proteinlerle ilgili hiçbir ilişki bilgisi eğitim sırasında model tarafından görülmediğinden performans sonuçlarının kenar-bazlı bölüm modellerine göre daha düşük olması beklenmektedir. Tam tersi olan durumun gözlenme sebebi, karşılaştırılan modellerde farklı bölüm yöntemleri uygulanırken karşılıklı eğitim ve test setlerinde aynı dağılımın yakalanmasının mümkün olmaması olabilir.

Model tasarımı için yapılan denemelerin ardından 1:100 negatif oranı, ağırlıklı kayıp fonksiyonu ve protein-bazlı veri bölümü seçenekleri belirlenmiş ve modelin hiperparametre optimizasyonuna geçilmiştir. Saklı kanal sayısı, öğrenme oranı, kafa sayısı, evrişim katmanı sayısı, eğitim seti küme boyutu, eğitim seti küme komşu sayısı ve ağırlık sönümü hiperparametreleri üzerinden, her 3 tahmin modeli için paralel şekilde rastgele arama yöntemi ile yapılan optimizasyon, 2 NVIDIA GeForce RTX2080 GPU kullanımında yaklaşık 10 gün sürmüştür. Hiperparametre optimizasyonu, model tasarımı denemeleri için belirlenen sabit hiperparametre seti ile son durumda elde edilen MCC performansının (Tablo 4.7.) moleküler fonksiyon

tahmininde yaklaşık %45, biyolojik süreç tahmininde yaklaşık %73, hücresel bileşen tahmininde yaklaşık %56 artışını sağlamıştır (Tablo 4.8.).

Veride yer alan düğüm tipleri ve bu düğüm tiplerini barındıran kenar tiplerinin eğitime olan katkısının incelenmesi için iki farklı ablasyon analizi gerçekleştirilmiştir. Öncelikle belirli bir düğüm tipi ve bu düğüm tipine ait tüm kenarların çıkartıldığı veri üzerinden modeller eğitilerek bu düğüm tipinin performansa etkisi incelenmiştir. Bu analiz sonuçlarının birbiri ile karşılaştırılabilir olması için eğitim sırasında model tasarımı seçeneklerini değerlendirirken kullanılan sabit hiperparametre seti kullanılmıştır. Ablasyon modelleri kendi aralarında ve aynı hiperparametre seti ile eğitilen tüm veri modeli ile karşılaştırılmıştır. Moleküler fonksiyon tahmininde ilaç adayı bileşik ablasyon modeli 0.5690 değeri ile en yüksek MCC performansına performansına ulaşmıştır. Biyolojik süreç tahmininde en yüksek MCC performansına ulaşan ablasyon modeli 0.4290 değeri ile ilaçken, hücresel bileşen tahmininde 0.5545 değeri ile moleküler fonksiyondur. Her üç fonksiyonel terim kategorisi için de hastalık ve ilaç ablasyon modelleri tüm veri modellerine göre daha iyi performansa ulaşmıştır (Tablo 4.9.-Tablo 4.11.). EC numarası düğüm tipinin çıkartılması ise üç fonksiyonel terim kategorisi için de performansta düşüşe sebep olmuştur. Bu durum EC numaralarının proteinlerin enzimatik aktivitesi ve dahil oldukları biyokimyasal reaksiyonlar hakkında bilgiler sağlayarak protein fonksiyon tahminine önemli katkıda bulunduğunu göstermektedir. Biyolojik yolak düğüm tipinin çıkartılması moleküler fonksiyon ve hücresel bileşen tahminini olumsuz etkilemiştir. Bunun dışında moleküler fonksiyon tahmininde domain, hücresel bileşen ve biyolojik süreç düğümlerinin çıkartılması performansta düşüşe sebep olmuştur. Biyolojik süreç tahmininde ise fenotipik terim ve moleküler fonksiyon düğümlerinin çıkartılması ile performans düşmüştür. Hücresel bileşen tahmininde ilaç adayı bileşik düğümlerinin çıkartılması performansı oldukça önemli oranda (%12) olumsuz etkilemiştir. Bu düşüşteki yüksek orana ilaç adayı bileşiklerin çizgede en çok kenara sahip düğüm tiplerinden biri olmasının sebep olduğu düşünülmektedir. Hücresel bileşen tahmininde gözlemlenen bir diğer ilginç durum da moleküler fonksiyon düğümleri çıkartıldığında en yüksek performansa ulaşılmasıdır. Bilinen fonksiyonel ilişkilerin bilinmeyen ilişkileri tahmin etmeye yardımcı olacağı varsayıldığından bu durum araştırılmak istenmiştir. Her üç fonksiyonel terim kategorisinde de en yüksek MCC performansına

ulaşan iki ablasyon modeli için hiperparametre optimizasyonu yapılıp elde edilen sonuçlar optimize edilmiş tüm veri modelleri ile karşılaştırılmıştır. Bu karşılaştırmaya göre optimal hiperparametreler ile eğitilmiş tüm veri modellerinin ablasyon modellerine göre daha iyi performans gösterdiği gözlemlenmiştir (Tablo 4.12.).

Düğüm tiplerinin eğitime tek başına olan katkısının incelenmesi için ikinci bir ablasyon analizi yapılmıştır. Bu analizde belirli bir düğüm tipi ve bu düğüm tipine ait kenarlar dışındaki tüm düğüm ve kenar tiplerinin çıkartıldığı veri setleri ile modeller eğitilmiştir. İlk ablasyon analizinde sabit hiperparametre kullanımının yanıltıcı sonuçlara ulaştırabileceği gözlemlendiğinden bu analizde her model için hiperparametre optimizasyonu yapılıp sonuçlar optimize edilmiş tüm veri modelleri ile karşılaştırılmıştır. Bu karşılaştırmada da tüm veri modellerinin her üç kategoride ablasyon modellerinden daha iyi performansa ulaştığı görülmüştür (Tablo 4.13.-Tablo 4.15.). Moleküler fonksiyon tahmininde biyolojik süreç düğüm tipinin tek başına veride yer alması 0.7513 değeri ile ablasyon modelleri arasında en yüksek MCC performansına ulaşmayı sağlamıştır. Bu sonucun sebebinin çizgeyi oluşturan kenarların büyük çoğunluğunda biyolojik süreç düğümünün yer alması olduğu düşünülmektedir (Tablo 4.13.). Bu durum biyolojik süreç ablasyon modelinin tek düğüm tipi kullanma ablasyon modelleri arasında en kapsamlı girdi çizgesinden yararlanmasını sağlamıştır. Tek düğüm tipi çıkarma ablasyon analizlerinde biyolojik süreç düğüm tipinin çıkartılmasının MCC performansında önemli ölçüde düşüşe sebep olması da (Tablo 4.9.) bu durumun bir göstergesidir.

Moleküler fonksiyon tahmini performansında biyolojik süreç ablasyon modelini 0.6931 MCC değeri ile domain ablasyon modeli takip etmektedir. Domain düğüm tipi hücresel bileşen tahmininde de 0.5536 değeri ile en yüksek MCC performansı sağlayan düğüm tiplerinden biri olmuştur**.** Domainlerin proteinlerin moleküler fonksiyonlarını, etkileşimlerini ve hücre içindeki yerleşimlerini belirlemedeki önemli rolünün bu sonuçlar üzerinde etkili olduğu düşünülmektedir. Diğer düğüm tipleri girdi verisinden çıkartıldığında, gürültünün azalması ve öğrenme görevinin basitleşmesi ile modelin domainlerin temsil ettiği önemli karakteristiklere daha fazla odaklanması mümkün olmuştur. Sadece hücresel bileşen düğüm tipinin yer aldığı veri ile eğitilen ablasyon modeli hücresel bileşen tahmininde 0.5649 değeri ile

ablasyon modelleri arasında en yüksek MCC değerine ulaşmıştır. Bu sonuç, özellikle diğer ablasyon modellerinin hiçbirinde yer almayan hücresel bileşen düğümlerinin birbirleri arasındaki hiyerarşik ilişkilerinin tahmin performansına olan önemli etkisine işaret etmektedir. Biyolojik süreç tahmininde ablasyon modelleri arasında en yüksek performansı 0.5517 MCC değeri ile KEGG biyolojik yolak düğüm tipi elde etmiştir.

Ardından, geliştirilen modellerin tahmin performansının daha objektif ve geniş çaplı değerlendirmesi için CAFA3 karşılaştırma veri seti için tahmin üretilmiş ve elde edilen performans diğer protein fonksiyon tahmini metodları ile karşılaştırılmıştır. Bu karşılaştırma için CAFA3 yarışma takvimine uygun olacak şekilde, veri tabanlarının Eylül 2016 öncesi son versiyonlarından elde edilmiş veri setleri ile yeni bir eğitim çizgesi oluşturulmuştur. CAFA3 eğitim çizge verisi (Tablo 4.3, Tablo 4.4) ile güncel çizge verisi (Tablo 4.1) arasında çeşitli sebeplerden kaynaklı kompozisyon farkları bulunmaktadır. Bu sebeplerden ilki, çizgeyi oluştururken entegre edilen veri setlerinin tarih/versiyon farklılığıdır. Bu durum çizgede yer alan protein, domain ve fonksiyonel terim düğüm sayısında ve dolayısıyla bu düğümlerin yer aldığı kenar sayılarında farklılığa sebep olmuştur. CAFA3 eğitim çizgesinde yer alan fonksiyonel terim düğümlerinden 399 tanesi Eylül 2016'dan bu yana benzer girdilerin birleştirilmesi gibi sebeplerle veri tabanından kaldırıldığından güncel eğitim çizgesinde yer almamaktadır. Buna rağmen bu süreçte GOA veri tabanına eklenen yeni fonksiyonel anotasyonlar sebebiyle güncel çizgede CAFA3 eğitim çizgesine kıyasla ortalama 2 kat daha fazla protein-fonksiyonel terim kenarı yer almaktadır. GO terimlerinin kendi aralarındaki hiyerarşik ilişkilerinin sayısı ise CAFA3 eğitim çizgesindeki fonksiyonel terimlerin fazlalığı sebebiyle bu veri setinde daha fazladır. İki eğitim çizgesindeki protein düğümlerinin kompozisyonundaki farklılığın bir diğer sebebi de güncel veri setinde yalnızca literatürde sıkça ele alınmış 29 organizmaya ait proteinlere yer verilirken, CAFA3 eğitim çizgesinde bunlara ek olarak yarışma için duyurulmuş eğitim protein listesindeki tüm proteinlere yer verilmesidir. Bu listede bulunan proteinlerin 4883 tanesi belirlenen 29 organizma dışındaki türlere ait olduğundan güncel eğitim çizgesinde yer almamaktadır. Bunun yanı sıra iki çizgeyi oluşturan UniProt ve proteinlerin filtrelenmesinde kullanılan UniRef veri setlerindeki versiyon farklılıkları, CAFA3 eğitim çizgesinde yer alan protein düğümlerinin ve dolayısıyla protein düğümlerinin yer aldığı kenar tiplerinin sayısında fazlalığa sebep olmuştur. Bu

fazlalığın veri içerisinde gürültüye sebep olarak öğrenme sürecine olumsuz etkileme ihtimali bulunmaktadır.

CAFA3 karşılaştırma veri seti üzerinden performans hesaplaması protein-merkezli değerlendirme yaklaşımı ile yapılmıştır. Bu yaklaşım bir proteinle ilişkilendirilen fonksiyonel terimlerin doğruluğunu ölçmektedir. Karşılaştırma veri setinde yer alan proteinlerden 8 tanesinin öznitelik vektörleri elde edilemediği için çizge verisine dahil edilememiş ve dolayısıyla bu proteinler için tahmin üretilememiştir. Bu durum modelimizin veri setindeki protein kapsama oranını tüm fonksiyonel terim kategorileri için 0.96'ya düşürmüştür. Fonksiyonel terimlere bakıldığında ise öznitelik vektörü elde edilemeyen 2 terim sebebiyle 0.999 kapsama oranı elde edildiği görülmüştür. Modelimiz karşılaştırma setinde yer alan 220 protein ve 1791 fonksiyonel terimin tüm olası eşleşmeleri için toplamda 394020 tahmin skoru üretmiştir. Gömme vektörlerindeki eksiklikler sebebiyle tahmin skoru üretilemeyen 14784 olası protein-fonksiyon tahmini bulunmaktadır. Bu çiftlerden 83'ü CAFA3 zaman çizelgesinde deneysel anotasyon kazanarak karşılaştırma setine dahil edilmiş olduğundan değerlendirmeye yanlış negatif olarak girecektir.

Karşılaştırma yapılan yöntemler arasında CAFA3 yarışmasında en yüksek performansa ulaşan 10 model, çizge tabanlı 1 model ve 2 temel fonksiyon tahmini metodu bulunmaktadır. Temel karşılaştırma metodlarından Naive, tüm fonksiyonel terimler için ilgili deneysel anotasyon veri tabanındaki göreceli frekanslarını belirler. Göreceli frekans, bir fonksiyonel terimin anotasyon veri tabanında ilişkilendirildiği protein sayısının bu veri tabanındaki tüm anote edilmiş protein sayısına oranıdır. Naive karşılaştırma veri setindeki tüm proteinler ile tüm olası fonksiyonel terimleri, tahmin skoru göreceli frekansları olacak şekilde eşleştirerek tahmin üretir (88). Bir diğer temel karşılaştırma metodu olan BLAST'ta ise tahminler deneysel olarak anote edilmiş proteinleri içeren bir veri tabanına olan BLAST (48) eşleşmeleri üzerinden üretilir. Bir hedef protein-fonksiyonel terim çifti için tahmin skoru, hedef protein ve bu terimle anote edilmiş herhangi bir protein arasındaki maksimum dizi benzerliği oranıdır (88).

Fmax değerleri üzerinden yapılan performans karşılaştırmasında (Şekil 4.3.) modelimizin tüm fonksiyonel terim kategorilerinde Naive ve BLAST modellerinden daha iyi performansa ulaştığı görülmüştür. BLAST yönteminin moleküler fonksiyon

tahmininde yüksek performansa ulaşmasının sebebi olarak sekans benzerliğine dayalı metodların enzimatik aktivite gibi temel biyokimyasal anotasyonları aktarmada başarılı olması görülmektedir (87). Biyolojik süreç kategorisinin ele aldığı biyolojik yolaklardaki roller gibi sekans benzerliği ile korunamayan fonksiyonların tahmininde BLAST'ın bu avantajını kaybederek performansta önemli bir düşüş yaşaması bu düşünceyi desteklemektedir. Sekans tabanlı protein öznitelik vektörlerine ek olarak diğer biyokimyasal bileşenlerle olan ilişkilerden de yararlanmanın avantajı, modelimizin tüm tahmin kategorilerinde tutarlı, yüksek performans gösteren 11 modelle karşılaştırılabilir performans elde etmesinde gözlemlenebilmektedir. Çizge tabanlı bir yöntem olan PANDA2 (67), moleküler fonksiyon ve biyolojik süreç kategorilerinde CAFA3'ün en başarılı tahmin yöntemi olan Zhu Lab modeli (54) ile oldukça yakın sonuçlara ulaşmış, hücresel bileşen kategorisinde ise en yüksek performans gösteren yöntem olmuştur. PANDA2, GO terimleri arasındaki hiyerarşik ilişkilerin yanı sıra evrimsel düzeyde bir dil modeli kullanılarak kodlanan protein sekans bilgisi ve homolojiden de yararlanmaktadır. Bu bilgilerin çizge sinir ağları kullanılarak entegre edilmesi ile PANDA2'nin CAFA3 veri seti üzerinde yüksek tahmin performansına ulaşması mümkün olmuştur. Yöntemlerin farklı kategorilerde farklı performans düzeylerine ulaşmasının sebebi ontolojilerin yapı ve kompleksitesindeki farklılıklarla beraber, veri tabanlarının bu kategorilerdeki farklı anotasyon oranlarına da bağlanmaktadır (98,99).

Özellikle moleküler fonksiyon tahmininde yakaladığı Fmax performansıyla geri kalan yöntemlere göre öne çıkan Zhu Lab modeli (54) farklı tiplerdeki sekans verilerinden yararlanan bir hibrit makine öğrenmesi metodudur. Smin değerleri üzerinden yapılan performans karşılaştırmasında (Şekil 4.4.) Zhu Lab modelindeki bu performans üstünlüğünün oranında azalma görülmüştür. Smin metriği, fonksiyonel terimleri koşullu bilgi içeriği ile ağırlıklandırarak daha bilgilendirici terimlerin tahmin edilmesinin performansa etkisinin, daha az bilgilendirici ve genel terimlere göre daha yüksek olmasını sağlar. Modelimizin moleküler fonksiyon tahmininde Smin performansının en iyi üç model arasında olması, bu kategorideki spesifik ve bilgilendirici terimleri tahmin etmedeki başarısına işaret etmektedir. Tüm kategorilerde Fmax metriği üzerinden yapılan değerlendirmelerde Zhu Lab modeli ile eşit veya daha yüksek performansa ulaşan çizge tabanlı PANDA2 için Smin metriği

sonuçları sunulmadığı için bu yöntemle karşılaştırma yapılamamıştır. Biyolojik süreç kategorisinde modelimiz Naive ve BLAST metodlarına göre daha iyi, yüksek performanslı modellerle ise karşılaştırılabilir Smin performansı elde etmiştir. Hücresel bileşen kategorisinde ise diğer tüm yöntemlerden daha düşük performans göstermiştir. Yüksek bilgi içeriğine sahip hücresel bileşen terimleri için modelimizin düşük tahmin yeteneğinin arkasındaki sebep, proteinlerin spesifik hücresel bölümlere yerleşiminde özellikle etkisi olan domain içeriği gibi veri tiplerinde deneysel anotasyonların azlığı ve dolayısıyla girdi verimizdeki seyrekliği olabilir.

Modellerin CAFA3 veri seti üzerinde genel tahmin performansının değerlendirilmesinin ardından bu veri setinde yer alan organizmalara özgü performanslar ayrı ayrı incelenmiştir. Tüm fonksiyonel terim kategorilerinde tek hücreli organizmalara ait proteinler için üretilen tahmin performansının çok hücreli organizmalara kıyasla daha yüksek olduğu gözlemlenmiştir (Tablo 4.16). Fakat bu proteinler karşılaştırma veri setinin küçük bir kısmını oluşturduğu için bu proteinlerin model performansına olan katkısı büyük oranda maskelenmiştir. Elde edilen sonuçlar CAFA3 yarışmacıları arasında her kategoride en yüksek performansa ulaşan 10 model ve 2 temel tahmin modelinin *Homo sapiens*, *Mus musculus* ve *Rattus norvegicus* türlerine spesifik sonuçları ile karşılaştırılmıştır (Şekil 4.5.-Şekil 4.7). Moleküler fonksiyon kategorisinde modelimiz *Homo sapiens* ve *Mus musculus* türlerinde temel tahmin modellerinden daha yüksek Fmax performansına ulaşırken, *Rattus norvegicus* türünde en iyi 6 tahmin modeli arasında yer almıştır. Hücresel bileşen kategorisinde *Mus musculus* türünde modelimiz en iyi 5 tahmin modeli arasına girerken; *Homo sapiens* türünde her iki temel tahmin modelinden, *Rattus norvegicus* türünde ise Naive'den daha yüksek performansa ulaşmıştır. Biyolojik süreç tahmininde ise model performansı her üç tür için diğer tüm tahmin modellerinin altında kalmıştır. Buna rağmen bu üç organizma dışında kalan organizmalara özgü performans değerlerinin bu organizmalar için elde edilen Fmax skorlarına kıyasla en az 2 kat fazla olmasının (Tablo 4.16) genel biyolojik süreç tahmini performansını olumlu etkilediği görülmüştür.

Modelin Tensin-2 ve Semaphorin-4D proteinleri ile ilişkilendirdiği fonksiyonel terimlerin biyolojik anlamlılığını incelemek için bir literatür taraması

yapılmıştır. Tensin-2 (UniProt Tanımlayıcısı: Q63HR2), hücre hareketliliği, proliferasyon ve insüline kas tepkisini regüle eden bir tirozin-protein fosfatazdır (100). Tensin-2'nin fosfataz aktivitesini tanımlayan deneysel anotasyonlarından "peptidil-tirozin defosforilasyonu" ("*peptidyl-tyrosine dephosphorylation*", GO:0035335), "fosfoprotein fostataz aktivitesi" ("*phosphoprotein phosphatase activity*", GO:0004721) ve protein tirozin fosfataz aktivitesi ("*protein tyrosine phosphatase activity*", GO:0004725) gibi moleküler fonksiyon ve biyolojik süreç terimleri modelimiz tarafından yüksek olasılık tahmini (> 0.90) ile doğru-pozitif olarak tahmin edilmiştir. Bunların yanında anlamsal olarak yine fosfataz aktivitesi ile ilişkili olabilecek "protein defosforilasyonunun pozitif regülasyonu" ("*positive regulation of protein dephosphorylation*", GO:0035307) tahmini, bu terim için deneysel bir anotasyon bulunmadığından yanlış-pozitif olarak değerlendirilmiştir. GO:0035307'nin, Tensin-2 anotasyonu GO:0035335 ile aynı ata terim üzerinden ("protein defosforilasyonu", GO:0006470) hiyerarşik olarak ilişki içinde olması (Şekil 4.8.A), bu tahminin biyolojik anlamlılığına işaret etmektedir.

Tensin-2 için aralarındaki ilişkinin varlığı pozitif olarak tahmin edilen fonksiyonel terimlerden bir tanesi de biyolojik süreç kategorisinin içindeki "insülin reseptör sinyal yolunun regülasyonu" (*"regulation of insulin receptor signaling pathway"*, GO:0046626)'dur. Bu fonksiyonel terim GO veri tabanında "İnsülin reseptörü sinyalizasyonunun sıklığını, hızını veya yayılımını düzenleyen herhangi bir süreç." olarak tanımlanmıştır (101). Hakkında deneysel anotasyon bulunmadığından bu protein-fonksiyonel terim ilişki tahmini yanlış pozitif olarak değerlendirilmiştir. Tensin-2 için yapılan literatür taramasında GO:0046626 ile ilişkisini destekleyen bulgulara ulaşılmıştır. Bu protein kaslarda katabolik koşullar altında insülin reseptörü substratı-1 (IRS1)'i defosforile edip degradasyonunu sağlayarak insülin reseptörü sinyal yolunun regülasyonunda görev almaktadır (100,102,103). Buna ek olarak tensin-2 proteini GO:0046626'nın kapsadığı "insülin reseptörü sinyal yolunun negatif düzenlenmesi" ("*negative regulation of insulin receptor signaling pathway*", GO:0046627) terimi ile manuel olarak anote edilmiştir. Modelimiz GO:0046627'den daha spesifik olan bu terimle de tensin-2'yi doğru pozitif olarak ilişkilendirmiştir (Şekil 4.8.A).

Modelin girdisi olarak kullanılan çizgi verisi, Tensin-2 ile Fibroblast büyüme faktörü 21 (UniProt Tanımlayıcısı: Q9NSA1) gibi proteinler arasında etkileşimler barındırmaktadır. Bu etkileşimler üzerinden büyüme faktörleriyle ilgili fonksiyonel anotasyonların Tensin-2'ye aktarıldığı gözlemlenmiştir. Bu aktarımlarla üretildiği düşünülen tahminlerden ikisi "büyüme faktörü bağlanması" ("*growth factor binding*", GO:0019838) ve "insülin benzeri büyüme faktörü reseptör sinyal yolağının negatif regülasyonu" ("*negative regulation of insulin-like growth factor receptor signaling pathway*", GO:0043569) terimleridir. Tensin-2'nin büyüme faktörü mekanizmaları ile ilgili deneysel olarak kanıtlanmış bir görevi bulunmadığından bu tahminler yanlış-pozitif olarak değerlendirilmiştir.

Semaphorin-4D (Uniprot Erişim ID: O09126), sinirsel gelişim, immün düzenleme ve kanser gelişimi gibi birçok süreçte önemli role sahip olan bir proteinidir (104). Hem bir ligand hem de reseptör olarak çift yönlü sinyal etkileşimlerine katılır. Deneysel anotasyon eksikliği nedeniyle yanlış-pozitif olarak değerlendirilen tahminlerden biri "transmembran reseptör protein serin/treonin kinaz aktivitesi" ("*transmembrane receptor protein serine/threonine kinase activity*", GO:0004675) moleküler fonksiyon terimidir. Bu tahminin biyolojik anlamlılığını araştırmak için yapılan literatür taramasında Semaphorin-4D'nin reseptörü Plexin-B1 aracılığı ile hücre içi sinyal yollarını etkinleştirerek hücre iskeletinin yeniden düzenlenmesi, hücre göçü ve adezyonda rol aldığı görülmüştür (105). Plexin-B1 özgül bir kinaz özelliğine sahip olmamakla birlikte, diğer kinazların aktivitesi ile etkileşim halinde hücresel süreçlerde yer almaktadır (106). Bu dolaylı etkileşim Semaphorin-4D'nin GO:0004675 fonksiyonel terimi ile ilişkili olabileceğine işaret etmektedir. Bunun yanı sıra semaphorin sinyalizasyonunda Cdk5, GSK3, MAPK ve LIMK gibi serin/treonin kinaz türlerinin rollerini inceleyen çalışmalar bulunmaktadır (107). Bu çalışmalarda ulaşılan bazı sonuçlar semaphorinler aracılığıyla tetiklenen farklı fonksiyonel etkilerde serin kinaz yolaklarının yer alabileceğine işaret etmiştir (108). Semaphorin-4D'nin deneysel anotasyonları incelendiğinde GO:0004675'yi hiyerarşik ve anlamsal olarak kapsayan "transmembran sinyal reseptörü aktivitesi" ("*transmembrane signaling receptor activity*", GO:0004888) ve "sinyal reseptörü aktivitesi" ("*signaling receptor activity*", GO:0038023) terimlerine rastlanmıştır (Şekil 4.8.B). Modelimiz bu iki terimi de yüksek olasılıkla (> 0.99) Semaphorin-4D ile ilişkilendirmeyi başarmıştır.

GO:0004675'e göre daha geniş kapsamlı olan bu işlevlerin alt türlerinin deneysel çalışmalar aracılığı olarak irdelenmesi, serin/treonin kinaz aktivitesi gibi daha spesifik hücresel süreçlerde Semaphorin-4D'nin rolünün ortaya çıkartılmasını sağlayabilir.

Biyolojik süreç kategorisinde bu protein için üretilen tahminler incelendiğinde yine deneysel anotasyon eksikliği nedeni ile yanlış-pozitif olarak tahmin edilen fakat Semaphorin-4D'nin bilinen fonksiyonları ile ilişkili olabilecek 3 terime rastlanmıştır. Bunlardan ilki "T hücresi göçünün negatif regülasyonu" ("*negative regulation of T cell migration*", GO:2000405) terimidir. T hücresi aktivasyonu, göçü ve bağışıklık hücresi etkileşimlerinin düzenlenmesi, semaphorinlerin bağışıklık sistemindeki bu terimle ilişkili olabilecek görevlerindendir (109). Semaphorinlerin hücresel göç ve immün hücre cevaplarındaki fonksiyonunu inceleyen çalışmalarda, immün hücre göçünün pozitif veya negatif yönde regülasyonunda etkileri olabileceğine dair sonuçlara ulaşılmıştır (110). Buna ek olarak tahmin edilen GO:2000405 teriminin, Semaphorin-4D'nin deneysel anotasyonlarından biri olan "hücre göçünün pozitif regülasyonu" ("*positive regulation of cell migration*", GO:0030335) ile hiyerarşik anlam ilişkisi bulunmaktadır (Şekil 4.8.B).

İncelenen diğer iki yanlış-pozitif tahmin "presinaptik birleşme" ("*presynapse assembly*", GO:0099054) ve miyelinasyon regülasyonu ("*regulation of myelination*", GO:0031641) terimleridir. Sinir sisteminde akson yönlendirmesi ve sinaps oluşumu semaphorinlerin bilinen görevleri arasındadır. Bu protein ailesinin uyarıcı ve inhibe edici sinapsları meydana getirecek elementlerin birleşimindeki rolünü açıklayan çalışmalar bulunmaktadır (111). Bazı çalışmalarda semaphorinlerin oligodendrosit farklılaşmasını kontrol ederek miyelinasyonun düzenlenmesinde rol alabileceğine değinilmiştir (112). Bunlara ek olarak Semaphorin-4D'nin deneysel anotasyonlarından biri olan "sinir sistemi gelişimi" ("*nervous system development*", GO:0007399) terimi, tahmin edilen GO:0031641 ve GO:0099054'ü hiyerarşik ve anlamsal olarak kapsamaktadır (Şekil 4.8.B). Bu bulgular Semaphorin-4D'nin GO:2000405, GO:0099054 ve GO:0031641 ile ilişkilerinin biyolojik olarak anlamlı olabileceğine işaret etmektedir.

# 6. SONUÇ VE ÖNERİLER

Bu tez çalışması kapsamında heterojen biyomedikal çizge verisi ve çizge tabanlı derin öğrenme mimarilerinin kullanımıyla protein fonksiyon tahmini için yeni bir yaklaşım önerilmiştir.

Ablasyon analizinde her bir düğüm tipi ve bu düğüm tipini içeren kenar tiplerinin çıkarıldığı çizge verisi için ayrı modeller eğitilmiştir. Bu şekilde düğüm tiplerinin öğrenmeye olan katkısının gözlemlenmesi amaçlanmıştır. Bunun için iki tip ablasyon analizi gerçekleştirilmiştir. İlk analizde veriden belirli bir düğüm tipi ile bu düğüm tipinin yer aldığı kenarlar çıkartılarak model kalan veri ile eğitilmiştir. Sabit bir hiperparametre seti üzerinden eğitilen bu modellerde en yüksek tahmin performansı moleküler fonksiyon, biyolojik süreç ve hücresel bileşen kategorilerinde sırasıyla ilaç adayı bileşik, ilaç ve moleküler fonksiyon ablasyon modelleridir. Her üç kategoride de hastalık ve ilaç düğümlerinin performansı olumlu, EC numarası düğümlerinin çıkartılması ise olumsuz etkilemiştir. Her kategorideki en iyi iki ablasyon modelinin optimize edildikten sonraki performansları optimize tüm veri modelinin performansına ulaşamamıştır. Bu durum sabit hiperparametre seti ile eğitilmiş modellerle elde edilen performansların karşılaştırmasının yanıltıcı olabileceğini göstermektedir. Tüm düğüm tiplerinin öğrenime olan etkisi ve bu etkinin sebeplerinin optimize edilmiş ablasyon modelleri üzerinden araştırılmasının daha anlamlı olacağı düşünülmektedir. Bu çıkarıma uygun olarak, ikinci tip ablasyon analizinde tüm modellerin optimize edilmiş tahmin performansları karşılaştırılmıştır. Bu ablasyon analizinde modeller, verideki yalnızca bir düğüm tipi ve bu düğüm tipinin yer aldığı kenarlar kullanılarak elde edilen veri setleri üzerinden eğitilmiştir. Bu analizde de her üç kategoride hiçbir ablasyon modelinin tüm veri ile eğitilmiş model performansına ulaşamadığı görülmüştür. Bu durum fonksiyon tahmini görevinde çeşitli biyolojik bileşenlerin birbiri ile olan ilişkilerinden yararlanmanın sağladığı faydaya işaret etmektedir. CAFA3 karşılaştırması için eğitilen modellerde de ablasyon deneylerinin yapılması ve sonuçların benzer fonksiyon tahmini yöntemleri ile kıyaslanması ilginç bir araştırma olacaktır.

CAFA3 veri seti üzerinden yapılan Fmax performansı karşılaştırmalarında önerilen model tüm fonksiyonel terim kategorilerinde temel fonksiyon tahmini

yöntemleri Naive ve BLAST'a göre daha yüksek performans göstermiştir. Yarışmada en iyi performansa ulaşan 10 tahmin modeli ve çizge tabanlı bir tahmin modeli ile karşılaştırılabilir Fmax değerleri elde edilmiştir. Smin skoru üzerinde yapılan değerlendirmede ise önerilen model moleküler fonksiyon tahmininde yüksek tahmin performansıyla en iyi üç yöntem arasında yer almıştır. Biyolojik süreç tahmininde yine Naive ve BLAST'a göre daha iyi, diğer modellere yakın Smin değerlerine ulaşılmıştır. Hücresel bileşen tahmininde ise diğer tüm modellere kıyasla daha düşük Smin performansı gözlemlenmiştir. Bu kategorideki yüksek bilgi içerikli terimlerin daha başarılı tahmini için girdi verisine proteinlerin hücre içi lokalizasyonuna dair bilgileri yakalamaya yardımcı olacak ilişki veya öznitelik tipleri eklenebilir. Proteinler için deneysel hücre içi lokalizasyon anotasyonları sunan veri tabanlarından (113–115) yararlanılarak çizgede yer alan bilinen protein-hücresel bileşen ilişkileri zenginleştirilebilir. Protein öznitelik vektörleri sekansa ek olarak yapısal ve fizikokimyasal özellikleri temsil edebilen gömme metodlarından yararlanarak genişletilebilir. Bu eklemelerin yalnızca hücresel bileşen terimi tahmini için değil genel tahmin performansına da katkıda bulunacağı düşünülmektedir. Denenen yeni öznitelik vektörlerinin öğrenmeye olan etkisinin incelenmesi için bu öznitelik vektörlerini içeren veri ile eğitilen modeller ablasyon analizine dahil edilebilir.

Model performansını etkileyen bir diğer önemli faktör de son HGT katmanından çıkan güncellenmiş protein ve fonksiyonel terim gömme vektörlerini tahmin skorlarına dönüştüren tahmin modülüdür. Bu tez çalışmasında yer verilen sonuçlar bu modül için basit bir iç çarpım kullanan modeller ile elde edilmiştir. Bunun yerine tam bağlı ileri beslemeli sinir ağı gibi bir yapı kullanmak gömme vektörlerindeki daha karmaşık kalıpların öğrenilmesini sağlayabilir. Bu yapılarla birlikte gelen katman sayısı ve regülarizasyon teknikleri gibi ek hiperparametreler üzerinden yeni deneyler yapmak, tahmin performansını optimize etmek için fırsat sunacaktır.

Tahminlerin biyolojik anlamlılığı, literatürde çok sayıda çalışmada ele alınmış iki protein üzerinden değerlendirilmiştir. Bu proteinlere yüksek tahmin olasılığı ile ilişkilendirilen fonksiyonların ilgili proteinlere manuel olarak da anote edildiği görülmüştür. Çizge verisinde bu protein ve terimler taratıldığında fonksiyon

aktarımının protein-protein etkileşimleri ve fonksiyonel terimlerin birbiri ile olan hiyerarşik ilişkileri gibi farklı ilişkiler üzerinden gerçekleştiği görülmüştür. Bu durum fonksiyon tahmini için birden fazla ilişki tipinden yararlanmanın avantajlarına dikkat çekmektedir. Anotasyon veri tabanlarında yer almadığı için yanlış-pozitif olarak değerlendirilen fonksiyon tahminlerinin ilgili protein ile olası ilişkisi literatürde taratıldığında, bu fonksiyonel ilişkileri genel olarak ele alan ve üzerine daha detaylı araştırmalar yapılmasını öneren çalışmalara rastlanmıştır.

# 7. KAYNAKLAR


1. Branden C, Tooze J. Introduction to protein structure. 2012 [a.yer 01 Haziran 2023]; Erişim adresi: https://www.google.com/books?hl=tr&lr=&id=eUYWBAAAQBAJ&oi=fnd&pg=PP1&dq=Branden+C,+Tooze+J.+(1999).+Introduction+to+Protein+Structure+2nd+ed.&ots=PBaZLDf0k-&sig=tMZK8qe9pkNPEHJF7-T6iHx7lzE

2. Shehu A, Barbará D, Molloy K. A survey of computational methods for protein function prediction. Big Data Analytics in Genomics [Internet]. 01 Ocak 2016 [a.yer 01 Haziran 2023];225-98. Erişim adresi: https://link.springer.com/chapter/10.1007/978-3-319-41279-5_7

3. Bateman A, Martin MJ, Orchard S, Magrane M, Ahmad S, Alpi E, vd. UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Res [Internet]. 06 Ocak 2023 [a.yer 28 Nisan 2023];51(D1):D523-31. Erişim adresi: https://academic.oup.com/nar/article/51/D1/D523/6835362

4. Carbon S, Douglass E, Good BM, Unni DR, Harris NL, Mungall CJ, vd. The Gene Ontology resource: enriching a GOld mine. Nucleic Acids Res [Internet]. 08 Ocak 2021 [a.yer 28 Nisan 2023];49(D1):D325-34. Erişim adresi: https://pubmed.ncbi.nlm.nih.gov/33290552/

5. Friedberg I. Automated protein function prediction—the genomic challenge. Brief Bioinform [Internet]. 01 Eylül 2006 [a.yer 01 Haziran 2023];7(3):225-42. Erişim adresi: https://academic.oup.com/bib/article/7/3/225/326173

6. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, vd. The reactome pathway knowledgebase 2022. Nucleic Acids Res [Internet]. 07 Ocak 2022 [a.yer 28 Nisan 2023];50(D1):D687-92. Erişim adresi: https://academic.oup.com/nar/article/50/D1/D687/6426058

7. Szklarczyk D, Gable AL, Lyon D, Junge A, Wyder S, Huerta-Cepas J, vd. STRING v11: protein–protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. Nucleic Acids Res [Internet]. 08 Ocak 2019 [a.yer 01 Haziran 2023];47(D1):D607-13. Erişim adresi: https://academic.oup.com/nar/article/47/D1/D607/5198476

8. Oughtred R, Rust J, Chang C, Breitkreutz BJ, Stark C, Willems A, vd. The BioGRID database: A comprehensive biomedical resource of curated protein, genetic, and chemical interactions. Protein Sci [Internet]. 01 Ocak 2021 [a.yer 01 Haziran 2023];30(1):187. Erişim adresi: /pmc/articles/PMC7737760/

9. Himmelstein DS, Lizee A, Hessler C, Brueggeman L, Chen SL, Hadley D, vd. Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife. 22 Eylül 2017;6.

10. Chandak P, Huang K, Zitnik M. Building a knowledge graph to enable precision medicine. Scientific Data 2023 10:1 [Internet]. 02 Şubat 2023 [a.yer 28 Mayıs 2023];10(1):1-16. Erişim adresi: https://www.nature.com/articles/s41597-023-01960-3

11. Zheng S, Rao J, Song Y, Zhang J, Xiao X, Fang EF, vd. PharmKG: a dedicated knowledge graph benchmark for bomedical data mining. Brief Bioinform [Internet]. 20 Temmuz 2021 [a.yer 28 Mayıs 2023];22(4). Erişim adresi: https://academic.oup.com/bib/article/22/4/bbaa344/6042240

12. Santos A, Colaço AR, Nielsen AB, Niu L, Strauss M, Geyer PE, vd. A knowledge graph to interpret clinical proteomics data. Nat Biotechnol [Internet]. 01 Mayıs 2022 [a.yer 01 Haziran 2023];40(5):692. Erişim adresi: /pmc/articles/PMC9110295/

13. Doğan T, Atas H, Joshi V, Atakan A, Rifaioglu AS, Nalbat E, vd. CROssBAR: comprehensive resource of biomedical relations with knowledge graph representations. Nucleic Acids Res [Internet]. 20 Eylül 2021 [a.yer 28 Nisan 2023];49(16):e96-e96. Erişim adresi: https://academic.oup.com/nar/article/49/16/e96/6310792

14. Kanehisa M, Furumichi M, Sato Y, Kawashima M, Ishiguro-Watanabe M. KEGG for taxonomy-based analysis of pathways and genomes. Nucleic Acids Res [Internet]. 06 Ocak 2023 [a.yer 28 Nisan 2023];51(D1):D587-92. Erişim adresi: https://academic.oup.com/nar/article/51/D1/D587/6775388

15. Wishart DS, Feunang YD, Guo AC, Lo EJ, Marcu A, Grant JR, vd. DrugBank 5.0: a major update to the DrugBank database for 2018. Nucleic Acids Res [Internet]. 01 Ocak 2018 [a.yer 28 Nisan 2023];46(D1):D1074-82. Erişim adresi: https://pubmed.ncbi.nlm.nih.gov/29126136/

16. Mendez D, Gaulton A, Bento AP, Chambers J, De Veij M, Félix E, vd. ChEMBL: towards direct deposition of bioassay data. Nucleic Acids Res [Internet]. 08 Ocak 2019 [a.yer 28 Nisan 2023];47(D1):D930-40. Erişim adresi: https://academic.oup.com/nar/article/47/D1/D930/5162468

17. Amberger JS, Bocchini CA, Scott AF, Hamosh A. OMIM.org: Leveraging knowledge across phenotype-gene relationships. Nucleic Acids Res. 08 Ocak 2019;47(D1):D1038-43.

18. Orphanet: an online database of rare diseases and orphan drugs. Copyright, INSERM 1997. [Internet]. [a.yer 28 Nisan 2023]. Erişim adresi: http://www.orpha.net

19. Malone J, Holloway E, Adamusiak T, Kapushesky M, Zheng J, Kolesnikov N, vd. Modeling sample variables with an Experimental Factor Ontology. Bioinformatics. 03 Mart 2010;26(8):1112-8.

20. Köhler S, Gargano M, Matentzoglu N, Carmody LC, Lewis-Smith D, Vasilevsky NA, vd. The human phenotype ontology in 2021. Nucleic Acids Res. 08 Ocak 2021;49(D1):D1207-17.

21. Cao W, Yan Z, He Z, He Z. A Comprehensive Survey on Geometric Deep Learning. IEEE Access. 2020;8:35929-49.

22. Atz K, Grisoni F, Schneider G. Geometric deep learning on molecular representations. Nature Machine Intelligence 2021 3:12 [Internet]. 15 Aralık 2021 [a.yer 01 Haziran 2023];3(12):1023-32. Erişim adresi: https://www.nature.com/articles/s42256-021-00418-8

23. Yan TC, Yue ZX, Xu HQ, Liu YH, Hong YF, Chen GX, vd. A systematic review of state-of-the-art strategies for machine learning-based protein function prediction. Comput Biol Med [Internet]. 01 Mart 2023 [a.yer 30 Ocak 2023];154:106446. Erişim adresi: https://linkinghub.elsevier.com/retrieve/pii/S0010482522011544

24. Nickel M, Tresp V, Kriegel HP. A Three-Way Model for Collective Learning on Multi-Relational Data. 2011;

25. Bordes A, Usunier N, Garcia-Durán A, Weston J, Yakhnenko O. Translating Embeddings for Modeling Multi-relational Data.

26. Lin Y, Liu Z, Sun M, Liu Y, Zhu X. Learning Entity and Relation Embeddings for Knowledge Graph Completion [Internet]. Erişim adresi: www.aaai.org

27. Wang Z, Zhang J, Feng J, Chen Z. Knowledge Graph Embedding by Translating on Hyperplanes. Proceedings of the AAAI Conference on Artificial Intelligence [Internet]. 21 Haziran 2014 [a.yer 01 Haziran 2023];28(1):1112-9. Erişim adresi: https://ojs.aaai.org/index.php/AAAI/article/view/8870

28. Rezaul Karim M, Cochez M, Jares JB, Uddin M, Beyan O, Decker S. Drug-drug interaction prediction based on knowledge graph embeddings and convolutional-LSTM network. ACM-BCB 2019 - Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics [Internet]. 04 Eylül 2019 [a.yer 01 Haziran 2023];113-23. Erişim adresi: https://dl.acm.org/doi/10.1145/3307339.3342161

29. Ma T, Lin X, Song B, Yu PS, Zeng X. KG-MTL: Knowledge Graph Enhanced Multi-Task Learning for Molecular Interaction. IEEE Trans Knowl Data Eng. 2022;

30. Defferrard M, Bresson X, Vandergheynst P. Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering. Adv Neural Inf Process Syst [Internet]. 30 Haziran 2016 [a.yer 01 Haziran 2023];3844-52. Erişim adresi: https://arxiv.org/abs/1606.09375v3

31. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE. Neural Message Passing for Quantum Chemistry. 34th International Conference on Machine Learning, ICML 2017 [Internet]. 04 Nisan 2017 [a.yer 01 Haziran 2023];3:2053-70. Erişim adresi: https://arxiv.org/abs/1704.01212v2

32. Hamilton WL, Ying R, Leskovec J. Inductive Representation Learning on Large Graphs. Adv Neural Inf Process Syst [Internet]. 07 Haziran 2017 [a.yer 01 Haziran 2023];2017-December:1025-35. Erişim adresi: https://arxiv.org/abs/1706.02216v4

33. Zitnik M, Agrawal M, Leskovec J. Modeling polypharmacy side effects with graph convolutional networks. Bioinformatics [Internet]. 01 Şubat 2018 [a.yer 01 Haziran 2023];34(13):i457-66. Erişim adresi: http://arxiv.org/abs/1802.00543

34. Zhou J, Cui G, Hu S, Zhang Z, Yang C, Liu Z, vd. Graph neural networks: A review of methods and applications. AI Open. 01 Ocak 2020;1:57-81.

35. Li Y, Qiao G, Wang K, Wang G. Drug-target interaction predication via multi-channel graph neural networks. Brief Bioinform [Internet]. 01 Ocak 2022 [a.yer 01 Haziran 2023];23(1). Erişim adresi: https://pubmed.ncbi.nlm.nih.gov/34661237/

36. You R, Yao S, Mamitsuka H, Zhu S. DeepGraphGO: Graph neural network for large-scale, multispecies protein function prediction. Bioinformatics. 01 Temmuz 2021;37:I262-71.

37. Yuan Q, Chen J, Zhao H, Zhou Y, Yang Y. Structure-aware protein-protein interaction site prediction using deep graph convolutional network. Bioinformatics [Internet]. 01 Ocak 2021 [a.yer 01 Haziran 2023];38(1):125-32. Erişim adresi: https://pubmed.ncbi.nlm.nih.gov/34498061/

38. Leskovec J. CS224W: Machine Learning with Graphs. [a.yer 01 Haziran 2023]; Erişim adresi: http://cs224w.stanford.edu

39. Kipf TN, Welling M. Semi-Supervised Classification with Graph Convolutional Networks. 5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings [Internet]. 09 Eylül 2016 [a.yer 01 Haziran 2023]; Erişim adresi: https://arxiv.org/abs/1609.02907v4

40. Veličković P, Casanova A, Liò P, Cucurull G, Romero A, Bengio Y. Graph Attention Networks. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings [Internet]. 30 Ekim 2017 [a.yer 01 Haziran 2023]; Erişim adresi: https://arxiv.org/abs/1710.10903v3

41. Lai B, Xu J. Accurate protein function prediction via graph attention networks with predicted structure information. Brief Bioinform. 01 Ocak 2022;23(1).

42. Vaswani A, Brain G, Shazeer N, Parmar N, Uszkoreit J, Jones L, vd. Attention Is All You Need.

43. Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P. Large-scale chemical language representations capture molecular structure and properties. Nature Machine Intelligence 2022 4:12 [Internet]. 21 Aralık 2022 [a.yer 01 Haziran 2023];4(12):1256-64. Erişim adresi: https://www.nature.com/articles/s42256-022-00580-7

44. Irwin R, Dimitriadis S, He J, Bjerrum EJ. Chemformer: a pre-trained transformer for computational chemistry. Mach Learn Sci Technol [Internet]. 31 Ocak 2022 [a.yer 01 Haziran 2023];3(1):015022. Erişim adresi: https://iopscience.iop.org/article/10.1088/2632-2153/ac3ffb

45. Yüksel A, Ulusoy E, Ünlü A, Doğan T. SELFormer: Molecular Representation Learning via SELFIES Language Models. 10 Nisan 2023 [a.yer 01 Haziran 2023]; Erişim adresi: https://arxiv.org/abs/2304.04662v2

46. Jiang Y, Jin S, Jin X, Xiao X, Wu W, Liu X, vd. Pharmacophoric-constrained heterogeneous graph transformer model for molecular property prediction. Communications Chemistry 2023 6:1 [Internet]. 03 Nisan 2023 [a.yer 05 Nisan 2023];6(1):1-9. Erişim adresi: https://www.nature.com/articles/s42004-023-00857-x

47. Dwivedi VP, Bresson X. A Generalization of Transformer Networks to Graphs. 17 Aralık 2020 [a.yer 01 Haziran 2023]; Erişim adresi: https://arxiv.org/abs/2012.09699v2

48. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, vd. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res [Internet]. 01 Eylül 1997 [a.yer 01 Haziran 2023];25(17):3389-402. Erişim adresi: https://academic.oup.com/nar/article/25/17/3389/1061651

49. Sigrist CJA, Cerutti L, De Castro E, Langendijk-Genevaux PS, Bulliard V, Bairoch A, vd. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res [Internet]. 24 Ekim 2010 [a.yer 01 Haziran 2023];38(Database issue):D161. Erişim adresi: /pmc/articles/PMC2808866/

50. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer ELL. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res [Internet]. 01 Ocak 1999 [a.yer 01 Haziran 2023];27(1):260-2. Erişim adresi: https://pubmed.ncbi.nlm.nih.gov/9847196/

51. Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics [Internet]. 15 Ocak 2020 [a.yer 01 Haziran 2023];36(2):422-9. Erişim adresi: https://academic.oup.com/bioinformatics/article/36/2/422/5539866

52. Rojano E, Jabato FM, Perkins JR, Córdoba-Caballero J, García-Criado F, Sillitoe I, vd. Assigning protein function from domain-function associations using DomFun. BMC Bioinformatics [Internet]. 01 Aralık 2022 [a.yer 28 Mayıs 2023];23(1):1-19. Erişim adresi: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-022-04565-6

53. Doğan T, MacDougall A, Saidi R, Poggioli D, Bateman A, O'Donovan C, vd. UniProt-DAAC: domain architecture alignment and classification, a new method for automatic functional annotation in UniProtKB. Bioinformatics [Internet]. 01 Ağustos 2016 [a.yer 01 Haziran 2023];32(15):2264-71. Erişim adresi: https://academic.oup.com/bioinformatics/article/32/15/2264/1742842

54. You R, Zhang Z, Xiong Y, Sun F, Mamitsuka H, Zhu S. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. Bioinformatics [Internet]. 15 Temmuz 2018 [a.yer 30 Ocak 2023];34(14):2465-73. Erişim adresi: https://academic.oup.com/bioinformatics/article/34/14/2465/4924212

55. Koonin E V., Galperin MY. Sequence — Evolution — Function. Sequence — Evolution — Function. 2003;

56. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, vd. Highly accurate protein structure prediction with AlphaFold. Nature 2021 596:7873 [Internet]. 15 Temmuz 2021 [a.yer 01 Haziran 2023];596(7873):583-9. Erişim adresi: https://www.nature.com/articles/s41586-021-03819-2

57. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, vd. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Res [Internet]. 02 Temmuz 2018 [a.yer 01 Haziran 2023];46(W1):W296-303. Erişim adresi: https://academic.oup.com/nar/article/46/W1/W296/5000024

58. Ovchinnikov S, Park H, Varghese N, Huang PS, Pavlopoulos GA, Kim DE, vd. Protein structure determination using metagenome sequence data. Science [Internet]. 20 Ocak 2017 [a.yer 01 Haziran 2023];355(6322):294-8. Erişim adresi: https://pubmed.ncbi.nlm.nih.gov/28104891/

59. Greener JG, Kandathil SM, Jones DT. Deep learning extends de novo protein modelling coverage of genomes using iteratively predicted structural constraints. Nature Communications 2019 10:1 [Internet]. 04 Eylül 2019 [a.yer 01 Haziran 2023];10(1):1-13. Erişim adresi: https://www.nature.com/articles/s41467-019-11994-0

60. Roy A, Yang J, Zhang Y. COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. Nucleic Acids Res [Internet]. 01 Temmuz 2012 [a.yer 28 Mayıs 2023];40(W1):W471-7. Erişim adresi: https://academic.oup.com/nar/article/40/W1/W471/1071850

61. Zhang C, Freddolino PL, Zhang Y. COFACTOR: improved protein function prediction by combining structure, sequence and protein–protein interaction information. Nucleic Acids Res [Internet]. 03 Temmuz 2017 [a.yer 28 Mayıs 2023];45(W1):W291-9. Erişim adresi: https://academic.oup.com/nar/article/45/W1/W291/3787871

62. Cai Y, Wang J, Deng L. SDN2GO: An integrated deep learning model for protein function prediction. Front Bioeng Biotechnol. 29 Nisan 2020;8:391.

63. Cao R, Cheng J. Integrated protein function prediction by mining function associations, sequences, and protein–protein and gene–gene interaction networks. Methods. 15 Ocak 2016;93:84-91.

64. Zhang F, Song H, Zeng M, Li Y, Kurgan L, Li M. DeepFunc: A Deep Learning Framework for Accurate Prediction of Protein Functions from Protein Sequences and Interactions. Proteomics [Internet]. 01 Haziran 2019 [a.yer 01 Haziran 2023];19(12):1900019. Erişim adresi: https://onlinelibrary.wiley.com/doi/full/10.1002/pmic.201900019

65. Wass MN, Barton G, Sternberg MJE. CombFunc: predicting protein function using heterogeneous data sources. Nucleic Acids Res [Internet]. 01 Temmuz 2012 [a.yer 01 Haziran 2023];40(W1):W466-70. Erişim adresi: https://academic.oup.com/nar/article/40/W1/W466/1077759

66. Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, vd. Structure-based protein function prediction using graph convolutional networks. Nat Commun. 01 Aralık 2021;12(1).

67. Zhao C, Liu T, Wang Z. PANDA2: Protein function prediction using graph neural networks. NAR Genom Bioinform. 01 Mart 2022;4(1).

68. Zhou N, Jiang Y, Bergquist TR, Lee AJ, Kacsoh BZ, Crocker AW, vd. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biol [Internet]. 19 Kasım 2019 [a.yer 31 Ocak 2023];20(1). Erişim adresi: https://pubmed.ncbi.nlm.nih.gov/31744546/

69. Altenhoff AM, Train CM, Gilbert KJ, Mediratta I, de Farias TM, Moi D, vd. OMA orthology in 2021: Website overhaul, conserved isoforms, ancestral gene order and more. Nucleic Acids Res. 08 Ocak 2021;49(D1):D373-9.

70. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, vd. InterPro in 2022. Nucleic Acids Res [Internet]. 06 Ocak 2023 [a.yer 28 Nisan 2023];51(D1):D418-27. Erişim adresi: http://www.ncbi.nlm.nih.gov/pubmed/36350672

71. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, vd. The GOA database: gene Ontology annotation updates for 2015. Nucleic Acids Res [Internet]. 28 Ocak 2015 [a.yer 28 Nisan 2023];43(Database issue):D1057-63. Erişim adresi: https://pubmed.ncbi.nlm.nih.gov/25378336/

72. Bairoch A. The ENZYME database in 2000. Nucleic Acids Res. 01 Ocak 2000;28(1):304-5.

73. Suzek BE, Wang Y, Huang H, McGarvey PB, Wu CH. UniRef clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics [Internet]. 03 Mart 2015 [a.yer 28 Nisan 2023];31(6):926. Erişim adresi: /pmc/articles/PMC4375400/

74. Rao R, Bhattacharya N, Thomas N, Duan Y, Chen X, Canny J, vd. Evaluating Protein Transfer Learning with TAPE. 19 Haziran 2019;

75. Edera AA, Milone DH, Stegmayer G. Anc2vec: embedding gene ontology terms by preserving ancestors relationships. Brief Bioinform [Internet]. 10 Mart 2022 [a.yer 28 Nisan 2023];23(2). Erişim adresi: https://academic.oup.com/bib/article/23/2/bbac003/6523148

76. Mikolov T, Chen K, Corrado G, Dean J. Efficient Estimation of Word Representations in Vector Space. 1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings [Internet]. 16 Ocak 2013 [a.yer 28 Nisan 2023]; Erişim adresi: https://arxiv.org/abs/1301.3781v3

77. Melidis DP, Nejdl W. Capturing Protein Domain Structure and Function Using Self-Supervision on Domain Architectures. Algorithms 2021, Vol 14, Page 28 [Internet]. 19 Ocak 2021 [a.yer 28 Nisan 2023];14(1):28. Erişim adresi: https://www.mdpi.com/1999-4893/14/1/28/htm

78. Weininger D. SMILES, a Chemical Language and Information System: 1: Introduction to Methodology and Encoding Rules. J Chem Inf Comput Sci [Internet]. 01 Şubat 1988 [a.yer 28 Nisan 2023];28(1):31-6. Erişim adresi: https://pubs.acs.org/doi/abs/10.1021/ci00057a005

79. Schwaller P, Probst D, Vaucher AC, Nair VH, Kreutter D, Laino T, vd. Mapping the space of chemical reactions using attention-based neural networks. Nature Machine Intelligence 2021 3:2 [Internet]. 28 Ocak 2021 [a.yer 28 Nisan

2023];3(2):144-52. Erişim adresi: https://www.nature.com/articles/s42256-020-00284-w

80. Řeh uřek R, Sojka P. Software Framework for Topic Modelling with Large Corpora. Içinde: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. Valletta, Malta: ELRA; 2010. s. 45-50.

81. Ali M, Hoyt CT, Domingo-Fernández D, Lehmann J, Jabeen H. BioKEEN: a library for learning and evaluating biological knowledge graph embeddings. Bioinformatics [Internet]. 15 Eylül 2019 [a.yer 28 Nisan 2023];35(18):3538-40. Erişim adresi: https://academic.oup.com/bioinformatics/article/35/18/3538/5320556

82. Grover A, Leskovec J. node2vec: Scalable Feature Learning for Networks. Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. 03 Temmuz 2016 [a.yer 28 Nisan 2023];13-17-August-2016:855-64. Erişim adresi: https://arxiv.org/abs/1607.00653v1

83. Peng C, Dieck S, Schmid A, Ahmad A, Knaus A, Wenzel M, vd. CADA: phenotype-driven gene prioritization based on a case-enriched knowledge graph. NAR Genom Bioinform [Internet]. 01 Eylül 2021 [a.yer 28 Nisan 2023];3(3). Erişim adresi: /pmc/articles/PMC8415429/

84. Krenn M, Häse F, Nigam A, Friederich P, Aspuru-Guzik A. Self-Referencing Embedded Strings (SELFIES): A 100% robust molecular string representation. Mach Learn Sci Technol [Internet]. 31 Mayıs 2019 [a.yer 28 Nisan 2023];1(4). Erişim adresi: http://arxiv.org/abs/1905.13741

85. Hu Z, Dong Y, Wang K, Sun Y. Heterogeneous Graph Transformer. The Web Conference 2020 - Proceedings of the World Wide Web Conference, WWW 2020 [Internet]. 20 Nisan 2020 [a.yer 28 Nisan 2023];2704-10. Erişim adresi: https://dl.acm.org/doi/10.1145/3366423.3380027

86. Fey M, Lenssen JE. Fast Graph Representation Learning with PyTorch Geometric. 06 Mart 2019 [a.yer 28 Nisan 2023]; Erişim adresi: https://arxiv.org/abs/1903.02428v3

87. Zhou N, Jiang Y, Bergquist TR, Lee AJ, Kacsoh BZ, Crocker AW, vd. The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome Biol [Internet]. 19 Kasım 2019 [a.yer 28 Nisan 2023];20(1). Erişim adresi: https://pubmed.ncbi.nlm.nih.gov/31744546/

88. Radivojac P, Clark WT, Oron TR, Schnoes AM, Wittkop T, Sokolov A, vd. A large-scale evaluation of computational protein function prediction. Nature Methods 2013 10:3 [Internet]. 27 Ocak 2013 [a.yer 08 Mayıs 2023];10(3):221-7. Erişim adresi: https://www.nature.com/articles/nmeth.2340

89. The pandas development team. pandas-dev/pandas: Pandas. 24 Nisan 2020 [a.yer 28 Nisan 2023]; Erişim adresi: https://zenodo.org/record/7857418

90. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, vd. Array programming with NumPy. Nature 2020 585:7825 [Internet]. 16

Eylül 2020 [a.yer 28 Nisan 2023];585(7825):357-62. Erişim adresi: https://www.nature.com/articles/s41586-020-2649-2

91. Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, vd. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics [Internet]. 01 Haziran 2009 [a.yer 28 Nisan 2023];25(11):1422-3. Erişim adresi: https://academic.oup.com/bioinformatics/article/25/11/1422/330687

92. Steven Bird, Ewan Klein, Edward Loper. Natural Language Processing with Python. O'Reilly Media Inc.; 2009.

93. Pedregosa FABIANPEDREGOSA F, Michel V, Grisel OLIVIERGRISEL O, Blondel M, Prettenhofer P, Weiss R, vd. Scikit-learn: Machine Learning in Python. Journal of Machine Learning Research [Internet]. 2011 [a.yer 28 Nisan 2023];12(85):2825-30. Erişim adresi: http://jmlr.org/papers/v12/pedregosa11a.html

94. Hunter JD. Matplotlib: A 2D graphics environment. Comput Sci Eng. 2007;9(3):90-5.

95. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, vd. PyTorch: An Imperative Style, High-Performance Deep Learning Library. Adv Neural Inf Process Syst [Internet]. 03 Aralık 2019 [a.yer 28 Nisan 2023];32. Erişim adresi: https://arxiv.org/abs/1912.01703v1

96. Yüksel A, Ulusoy E, Ünlü A, Deniz G, Doğan T. SELFormer: Molecular Representation Learning via SELFIES Language Models. 10 Nisan 2023 [a.yer 28 Nisan 2023]; Erişim adresi: https://arxiv.org/abs/2304.04662v1

97. Sänger M, Leser U. Large-scale entity representation learning for biomedical relationship extraction. Bioinformatics [Internet]. 19 Nisan 2021 [a.yer 28 Nisan 2023];37(2):236-42. Erişim adresi: https://academic.oup.com/bioinformatics/article/37/2/236/5877941

98. Jiang Y, Oron TR, Clark WT, Bankapur AR, D'Andrea D, Lepore R, vd. An expanded evaluation of protein function prediction methods shows an improvement in accuracy. Genome Biol [Internet]. 07 Eylül 2016 [a.yer 01 Haziran 2023];17(1):1-19. Erişim adresi: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1037-6

99. Peng Y, Jiang Y, Radivojac P. Enumerating consistent sub-graphs of directed acyclic graphs: an insight into biomedical ontologies. Bioinformatics [Internet]. 01 Temmuz 2018 [a.yer 01 Haziran 2023];34(13):i313-22. Erişim adresi: https://academic.oup.com/bioinformatics/article/34/13/i313/5045754

100. Koh A, Lee MN, Yang YR, Jeong H, Ghim J, Noh J, vd. C1-Ten is a protein tyrosine phosphatase of insulin receptor substrate 1 (IRS-1), regulating IRS-1 stability and muscle atrophy. Mol Cell Biol [Internet]. 01 Nisan 2013 [a.yer 01 Haziran 2023];33(8):1608-20. Erişim adresi: https://pubmed.ncbi.nlm.nih.gov/23401856/

101. QuickGO::Term GO:0046626 [Internet]. [a.yer 01 Haziran 2023]. Erişim adresi: https://www.ebi.ac.uk/QuickGO/term/GO:0046626

102. Kim E, Kim DH, Singaram I, Jeong H, Koh A, Lee J, vd. Cellular phosphatase activity of C1-Ten/Tensin2 is controlled by Phosphatidylinositol-3,4,5-triphosphate binding through the C1-Ten/Tensin2 SH2 domain. Cell Signal [Internet]. 01 Kasım 2018 [a.yer 01 Haziran 2023];51:130-8. Erişim adresi: https://pubmed.ncbi.nlm.nih.gov/30092354/

103. ENZYME - 3.1.3.48 protein-tyrosine-phosphatase [Internet]. [a.yer 01 Haziran 2023]. Erişim adresi: https://enzyme.expasy.org/EC/3.1.3.48

104. Sema4d - Semaphorin-4D - Mus musculus (Mouse) | UniProtKB | UniProt [Internet]. [a.yer 01 Haziran 2023]. Erişim adresi: https://www.uniprot.org/uniprotkb/O09126/entry

105. Vodrazka P, Korostylev A, Hirschberg A, Swiercz JM, Worzfeld T, Deng S, vd. The semaphorin 4D-plexin-B signalling complex regulates dendritic and axonal complexity in developing neurons via diverse pathways. Eur J Neurosci [Internet]. Ekim 2009 [a.yer 01 Haziran 2023];30(7):1193-208. Erişim adresi: https://pubmed.ncbi.nlm.nih.gov/19788569/

106. Basile JR, Gavard J, Gutkind JS. Plexin-B1 utilizes RhoA and Rho kinase to promote the integrin-dependent activation of Akt and ERK and endothelial cell motility. Journal of Biological Chemistry [Internet]. 30 Kasım 2007 [a.yer 01 Haziran 2023];282(48):34888-95. Erişim adresi: http://www.jbc.org/article/S0021925820546243/fulltext

107. Ahmed A, Eickholt BJ. Intracellular kinases in semaphorin signaling. Adv Exp Med Biol [Internet]. 2007 [a.yer 01 Haziran 2023];600:24-37. Erişim adresi: https://pubmed.ncbi.nlm.nih.gov/17607944/

108. Elhabazi A, Lang V, Hérold C, Freeman GJ, Bensussan A, Boumsell L, vd. The human semaphorin-like leukocyte cell surface molecule CD100 associates with a serine kinase activity. J Biol Chem [Internet]. 19 Eylül 1997 [a.yer 01 Haziran 2023];272(38):23515-20. Erişim adresi: https://pubmed.ncbi.nlm.nih.gov/9295286/

109. Kikutani H, Kumanogoh A. Semaphorins in interactions between T cells and antigen-presenting cells. Nature Reviews Immunology 2003 3:2 [Internet]. Şubat 2003 [a.yer 01 Haziran 2023];3(2):159-67. Erişim adresi: https://www.nature.com/articles/nri1003

110. Takamatsu H, Okuno T, Kumanogoh A. Regulation of immune cell responses by semaphorins and their receptors. Cell Mol Immunol [Internet]. Mart 2010 [a.yer 01 Haziran 2023];7(2):83. Erişim adresi: /pmc/articles/PMC4076735/

111. Koropouli E, Kolodkin AL. Semaphorins and the Dynamic Regulation of Synapse Assembly, Refinement, and Function. Curr Opin Neurobiol [Internet]. Ağustos 2014 [a.yer 01 Haziran 2023];0:1. Erişim adresi: /pmc/articles/PMC4122587/

112. Bernard F, Moreau-Fauvarque C, Heitz-Marchaland C, Zagar Y, Dumas L, Fouquet S, vd. Role of transmembrane semaphorin Sema6A in oligodendrocyte

differentiation and myelination. Glia [Internet]. Ekim 2012 [a.yer 01 Haziran 2023];60(10):1590-604. Erişim adresi: https://pubmed.ncbi.nlm.nih.gov/22777942/

113. Sprenger J, Lynn Fink J, Karunaratne S, Hanson K, Hamilton NA, Teasdale RD. LOCATE: a mammalian protein subcellular localization database. Nucleic Acids Res [Internet]. Ocak 2008 [a.yer 01 Haziran 2023];36(Database issue). Erişim adresi: https://pubmed.ncbi.nlm.nih.gov/17986452/

114. Thumuluri V, Almagro Armenteros JJ, Johansen AR, Nielsen H, Winther O. DeepLoc 2.0: multi-label subcellular localization prediction using protein language models. Nucleic Acids Res [Internet]. 05 Temmuz 2022 [a.yer 01 Haziran 2023];50(W1):W228-34. Erişim adresi: https://academic.oup.com/nar/article/50/W1/W228/6576357

115. Yu CS, Cheng CW, Su WC, Chang KC, Huang SW, Hwang JK, vd. CELLO2GO: A Web Server for Protein subCELlular LOcalization Prediction with Functional Gene Ontology Annotation. PLoS One [Internet]. 09 Haziran 2014 [a.yer 01 Haziran 2023];9(6):99368. Erişim adresi: /pmc/articles/PMC4049835/

# 8. EKLER

## EK-1: Tez Çalışması ile İlgili Etik Kurul İzinleri

T.C.
HACETTEPE ÜNİVERSİTESİ
Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu

Sayı : 16969557-1764
Konu :

21.09.2021

**Doç. Dr. Tunca DOĞAN**
**Mühendislik Fakültesi**
**Bilgisayar Mühendisliği Bölümü**
**Yapay Zeka Mühendisliği Anabilim Dalı**
**Öğretim Üyesi**

**Sayın Doç. Dr. DOĞAN,**

Kurulumuza değerlendirilmek üzere sunduğunuz GO 21/993 kayıt numaralı ve ***"Heterojen Biyomedikal Verinin Bilgi Çizgeleri ve Derin Öğrenme Tabanlı Analizi ile Protein Fonksiyonlarının Otomatik Tahmini"*** başlıklı proje Kurulumuzun 21.09.2021 tarihli toplantısında değerlendirilmiş olup, çalışmanın erişime açık veri tabanlarından veri toplanması yolu ile yapılacağı görülmüştür. Gönüllü insanlar üzerinde gerçekleştirilecek nitelikte olmayan bu tip çalışmalar Etik Kurulların kapsamı dışında kalmaktadır.

Bu yazı ilgili protokolün bilimsel ve etik açıdan incelendiğini belirtilmek için Etik Kurul kararı yerine geçmek üzere hazırlanmıştır.

Prof. Dr. G. Burça DOGAN
Başkan

EK :
Toplantı Katılım Tutanağı.

Hacettepe Üniversitesi Girişimsel Olmayan Klinik Araştırmalar Etik Kurulu
06100 Sıhhiye-Ankara
Telefon: 0 (312) 305 1082 • Faks: 0 (312) 310 0580 • E-posta: goetik@hacettepe.edu.tr

Ayrıntılı Bilgi için:

## EK-2: Tez Çalışması Orijinallik Raporu

turnitin

# Dijital Makbuz

Bu makbuz ödevinizin Turnitin'e ulaştığını bildirmektedir. Gönderiminize dair bilgiler şöyledir:

Gönderinizin ilk sayfası aşağıda gönderilmektedir.

| | |
|---|---|
| Gönderen: | Erva Ulusoy |
| Ödev başlığı: | Erva Ulusoy - Biyoinformatik YL Tezi (son sürüm) |
| Gönderi Başlığı: | Erva Ulusoy - Biyoinformatik YL Tezi (son sürüm) |
| Dosya adı: | Erva_Ulusoy_YL_Tez_final_2.pdf |
| Dosya boyutu: | 5.03M |
| Sayfa sayısı: | 101 |
| Kelime sayısı: | 22,308 |
| Karakter sayısı: | 143,890 |
| Gönderim Tarihi: | 18-Tem-2023 04:38ÖS (UTC+0300) |
| Gönderim Numarası: | 2133096190 |

TC.
HACETTEPE ÜNİVERSİTESİ
SAĞLIK BİLİMLERİ ENSTİTÜSÜ

HETEROJEN BİYOMEDİKAL VERİNİN BİLGİ ÇİZGELERİ VE DERİN ÖĞRENME TABANLI ANALİZİ İLE PROTEİN FONKSİYONLARININ OTOMATİK TAHMİNİ

Erva ULUSOY

Biyoinformatik Programı
YÜKSEK LİSANS TEZİ

ANKARA
2023

Erva Ulusoy - Biyoinformatik YL Tezi (son sürüm)

ORİJİNALLİK RAPORU

| %6 | %5 | %1 | %4 |
|---|---|---|---|
| BENZERLİK ENDEKSİ | İNTERNET KAYNAKLARI | YAYINLAR | ÖĞRENCİ ÖDEVLERİ |

BİRİNCİL KAYNAKLAR

| | | |
|---|---|---|
| 1 | Submitted to Hacettepe University<br>Öğrenci Ödevi | %3 |
| 2 | openaccess.hacettepe.edu.tr:8080<br>İnternet Kaynağı | %1 |
| 3 | acikbilim.yok.gov.tr<br>İnternet Kaynağı | <%1 |
| 4 | www.openaccess.hacettepe.edu.tr:8080<br>İnternet Kaynağı | <%1 |
| 5 | stars.library.ucf.edu<br>İnternet Kaynağı | <%1 |
| 6 | dspace.yildiz.edu.tr<br>İnternet Kaynağı | <%1 |
| 7 | docplayer.biz.tr<br>İnternet Kaynağı | <%1 |
| 8 | Submitted to University of Birmingham<br>Öğrenci Ödevi | <%1 |
| 9 | Zinnet Duygu Aksehir, Erdal Kiliic. "The Effect of Statistical Attributes on the Determination of Stock Trading Actions", 2022 7th | <%1 |

# 9. ÖZGEÇMİŞ