

Protein Katlanması'nın Çözülmesi: Evrimsel Algoritmalar ve Yapay Zekanın Gücüyle Yenilikçi Yaklaşımlar

Peren Jerfi Canatalay¹ , Madina Namazova^{2*} , Vildan Enisoğlu Atalay³ 

¹Department of Computer Engineering, İstinye University, Istanbul, Turkey

²Biyoinformatik Anabilim Dalı, Üsküdar Üniversitesi, İstanbul, Türkiye

³Hesaplamalı Bilim ve Mühendislik, İstanbul Teknik Üniversitesi, İstanbul, Türkiye

*madina.namazova@gmail.com

Özet

AlphaFold, protein dizisinden üç boyutlu (3D) yapı tahmini yaparak yapısal biyoloji alanında önemli değişimlere yol açmıştır. Bu başarı, protein katlanma probleminin çözüldüğü yönündeki iddiaları beraberinde getirmiştir. Ancak, protein katlanma problemi diziden yapı tahmininin ötesinde unsurları içermektedir ve AlphaFold'un bu konudaki potansiyeli halen tartışmalıdır. Bu çalışmada, AlphaFold'un tekli mutasyonların protein stabilitesi ($\Delta\Delta G$) ve fonksiyonuna etkisini tahmin etme yeteneğini değerlendirdik. Bu amacı gerçekleştirmek için, AlphaFold'un tahminlerinden elde edilen pLDDT ve <pLDDT> metriklerini, bir proteindeki tekli mutasyondan önce ve sonra çıkararak, bu değişiklikleri deneysel olarak bilinen $\Delta\Delta G$ değerleriyle ilişkilendirdik. Ayrıca, büyük ölçekli bir GFP mutant veri setini kullanarak, bu metrikleri yapı üzerindeki mutasyonların etkileri ve floresans seviyeleri ile karşılaştırdık. Bulgularımız, AlphaFold çıktı metrikleri ile protein stabilitesi veya floresan değişimi arasında çok zayıf veya anlamlı bir korelasyon bulunmadığını ortaya koymuştur.

Anahtar Kelimeler: AlphaFold, Derin öğrenme, Protein yapısı modelleme, Yapısal biyoinformatik

Solving Protein Folding: Innovative Approaches Powered by Evolutionary Algorithms and Artificial Intelligence

Abstract

AlphaFold has brought significant transformations to the field of structural biology by predicting three-dimensional (3D) structures from protein sequences. This success has led to claims that the protein folding problem has been solved. However, the protein folding problem encompasses more than just structure prediction from sequence, and AlphaFold's potential in this broader context remains debatable. In this study, we evaluate AlphaFold's ability to predict the effects of single-point mutations on protein stability ($\Delta\Delta G$) and function. To achieve this, we analyzed the pLDDT and <pLDDT> metrics derived from AlphaFold's predictions, comparing their values before and after single mutations in proteins, and correlated these changes with experimentally determined $\Delta\Delta G$ values. Furthermore, using a large-scale GFP mutant dataset, we examined the relationship between these metrics and the effects of mutations on protein structure and fluorescence levels. Our findings reveal that AlphaFold's output metrics show weak or no meaningful correlation with changes in protein stability or fluorescence.

Keywords: AlphaFold, Deep learning, Protein structure modeling, Structural bioinformatics

1. GİRİŞ

Proteinlerin üç boyutlu (3D) yapıları, biyolojik fonksiyonlarının anlaşılmasında ve biyoteknolojik uygulamalarda kritik bir rol oynamaktadır [1]. Protein yapısının doğru bir şekilde tahmin edilmesi, ilaç keşfi, enzim mühendisliği ve hastalık mekanizmalarının aydınlatılması gibi alanlarda büyük önem taşımaktadır [2,3]. Geleneksel deneysel yöntemler olan X-ışını kristalografisi, nükleer manyetik rezonans (NMR) spektroskopisi ve kriyo-elektron mikroskobu (Cryo-EM), protein yapılarının belirlenmesinde altın standart olmasına rağmen, bu teknikler zaman alıcı, maliyetli ve her protein için uygulanabilir değildir [4,5].

Yaklaşık 50 yıldır süregelen protein katlanma problemi, bir protein dizisinden onun üç boyutlu yapısının tahmin edilmesini hedefler ve bu alanda önemli zorluklar barındırır [6]. Protein dizilerinin çeşitliliği ve katlanma süreçlerinin karmaşıklığı, bu problemin çözümünü güçleştirmiştir [7]. Ancak son yıllarda, yapay zekâ ve özellikle derin öğrenme yöntemlerindeki gelişmeler, bu alanda devrim niteliğinde ilerlemeler sağlamıştır [8]. Bu bağlamda, derin öğrenme tabanlı bir yapay zekâ sistemi olan AlphaFold, protein yapı tahmininde önemli bir başarı elde etmiştir [9].

AlphaFold, 2020 yılında gerçekleştirilen Kritik Yapı Tahmini Değerlendirmesi (CASP14) yarışmasında diğer tüm yöntemleri geride bırakarak, protein yapı tahmininde deneysel yöntemlere yakın bir doğruluk elde etmiştir [10]. Bu başarı, yapısal biyolojide ve biyoinformatikte yeni ufuklar açmış, bilim dünyasında büyük bir etki yaratmıştır [11]. AlphaFold'un başarısı, protein yapı tahmininde derin öğrenme yöntemlerinin etkinliğini göstermiş ve bu alanda yeni araştırmaların önünü açmıştır [12]. AlphaFold'un temel özelliklerinden biri, proteinlerin üç boyutlu yapısını yüksek doğrulukla tahmin edebilmesi ve bu tahminlerin güvenilirliğini ölçmek için pLDDT (per-residue local distance difference test) skorunu kullanmasıdır [13]. pLDDT skoru, her bir amino asit kalıntısı için 0 ile 100 arasında bir değer alır ve modelin belirli bir kalıntının yapısındaki konumuna olan güvenini yansıtır [14]. Yüksek pLDDT skorları, protein modellerinin biyolojik süreçlerin anlaşılmasında ve uygulamalı araştırmalarda kullanılabilirliğini artırmaktadır [15].

AlphaFold'un başarıları, COVID-19 pandemisi sırasında da kendini göstermiştir. COVID19virüsünün protein yapılarının hızla aydınlatılmasına katkıda bulunarak, virüsün biyolojik özelliklerinin daha iyi anlaşılmasını sağlamıştır [16]. Bu sayede, etkili aşı ve ilaç geliştirme süreçleri hızlanmış, AlphaFold'un biyoteknoloji ve sağlık alanlarındaki potansiyeli ortaya konmuştur [17]. Ayrıca, Avrupa Biyoinformatik Enstitüsü (EMBL-EBI) ile yapılan iş birliği sonucu oluşturulan ve tüm protein dizi uzayını kapsayan serbest erişimli yapı veri tabanı, araştırmacıların hızlı ve etkili bir şekilde veri analizi yapabilmesini sağlamıştır [18]. Bununla birlikte, AlphaFold'un yeteneklerinin protein yapı tahmininin ötesine geçip geçemeyeceği ve özellikle mutasyonların protein stabilitesi üzerindeki etkisini ne kadar doğru tahmin edebileceği hâlâ tartışma konusudur [19]. Protein stabilitesi, biyolojik fonksiyonların sürdürülebilirliği ve hastalık mekanizmalarının anlaşılması açısından önemli bir parametredir [20]. Mutasyonlar, proteinlerin stabilitesini ve dolayısıyla fonksiyonlarını etkileyebilir; bu da ilaç keşfi ve biyoteknolojik uygulamalarda büyük önem taşır [21].

Mevcut literatür, AlphaFold'un mutasyonların neden olduğu stabilite değişimlerini tahmin etmede sınırlı bir kapasiteye sahip olduğunu göstermektedir [22]. Örneğin, pLDDT skorları ile mutasyon kaynaklı stabilite değişimleri arasında genellikle zayıf bir korelasyon bulunmaktadır [23]. Yapılan bir çalışmada, pLDDT ve ortalama pLDDT ($\langle pLDDT \rangle$) metrikleri ile mutasyonların neden olduğu stabilite değişimleri ($\Delta\Delta G$) arasında düşük bir korelasyon katsayısı bulunmuştur [24]. Bu durum, AlphaFold'un protein stabilitesi tahmininde kullanılabilirliğini kısıtlamakta ve bu alanda daha ileri araştırmaların gerekliliğini ortaya koymaktadır [25].

Bu çalışmanın amacı, AlphaFold'un mutasyonların protein stabilitesi ve fonksiyonu üzerindeki etkilerini tahmin etme kapasitesini kapsamlı bir şekilde değerlendirmektir [26]. Büyük ölçekli bir yeşil floresan protein (GFP) mutant veri seti kullanılarak, AlphaFold tarafından üretilen pLDDT skorlarının protein stabilitesi değişimleri ($\Delta\Delta G$) ve fonksiyonel etkiler ile olan ilişkisi incelenmiştir [27]. Sonuçlar, AlphaFold

çıktı metrikleri ile protein stabilitesi veya fonksiyon değişikliği arasında zayıf bir korelasyon olduğunu ve bu metriklerin mutasyonların etkilerini öngörmek için yeterince güvenilir olmadığını göstermektedir [28].

Protein katlanma probleminin çözümü, biyoloji ve tıp alanlarında yapay zekâ ve derin öğrenme yöntemlerinin giderek artan bir rol üstlenmesine neden olmuştur [29]. AlphaFold'un başarısı, bu alanda diğer derin öğrenme tabanlı yaklaşımların ve geleneksel yöntemlerin de değerlendirilmesini teşvik etmiştir [30]. Örneğin, şablon tabanlı modelleme, moleküler dinamik simülasyonları ve destek vektör makineleri gibi teknikler, protein yapı ve stabilite tahmininde kullanılan diğer yöntemler arasında yer almaktadır [31,32]. Bu yaklaşımlar, farklı veri setleri ve analiz teknikleri ile desteklenerek, proteinlerin yapısal ve fonksiyonel özelliklerini daha iyi anlamamızı sağlar [33].

Gelecekteki araştırmalar, AlphaFold'un sınırlamalarının ötesine geçerek protein yapılarının daha karmaşık ve dinamik özelliklerini ele almayı hedeflemektedir [34]. Mutasyonların proteinlerin dinamik davranışları üzerindeki etkilerini modelleyen simülasyon tabanlı yaklaşımlar, protein stabilitesini ve fonksiyonel değişiklikleri daha doğru tahmin edebilir [35]. Moleküler dinamik (MD) simülasyonları, protein katlanma yollarını ve mutasyonların bu yollar üzerindeki etkilerini incelemek için yaygın olarak kullanılır [36]. MD simülasyonlarının derin öğrenme algoritmaları ile entegrasyonu, daha hassas ve kapsamlı modellerin geliştirilmesine olanak tanıyabilir [37]. Ayrıca, çoklu omik veri analizleri, mutasyonların biyokimyasal yollar ve hücresel süreçler üzerindeki etkilerini anlamak için yeni bir bakış açısı sunabilir [38]. Bu tür disiplinler arası yaklaşımlar, yalnızca protein yapısının değil, aynı zamanda proteinlerin hücresel ve organizma düzeyinde nasıl işlediğinin anlaşılmasına da katkı sağlar [39]. Deneysel verilerin artırılması ve veriye dayalı modellerin güçlendirilmesi, biyoinformatik alanında devrimsel gelişmelere zemin hazırlayabilir [40].

Deneysel $\Delta\Delta G$ verilerinin sınırlı olması, bu alandaki algoritmaların eğitimini zorlaştırmakta ve modellerin genel doğruluk seviyelerini etkileyebilmektedir [41]. Gelecekte, bu eksiklikleri aşmak için yapay veri üretimi, simülasyon tabanlı yöntemler ve hibrit modelleme yaklaşımlarının birleştirilmesi, AlphaFold gibi araçların mutasyon etkileri ve protein stabilitesi tahmininde daha etkili kullanılmasını sağlayabilir [42]. Bu gelişmeler, protein mühendisliği, ilaç tasarımı ve moleküler biyoloji gibi alanlarda yeni keşiflerin önünü açarak, bilim dünyasında daha güvenilir ve kapsamlı çözümler sunacaktır [43].

Ek olarak, biyoinformatik ve hesaplamalı biyoloji alanlarında kuantum hesaplama tekniklerinin uygulanması, protein yapı ve dinamiklerinin daha hızlı ve doğru bir şekilde modellenmesine katkı sağlayabilir [44]. Kuantum bilgisayarların yüksek işlem gücü, karmaşık biyolojik sistemlerin simülasyonunda yeni olanaklar sunabilir [45]. Bu sayede, protein katlanma süreçlerinin ve mutasyon etkilerinin daha detaylı bir şekilde incelenmesi mümkün olabilir [46]. Son olarak, AlphaFold'un sağladığı modellerin ve veritabanlarının açık erişimli olması, bilimsel iş birliğini ve veri paylaşımını teşvik etmektedir [47]. Araştırmacılar, bu kaynakları kullanarak yeni hipotezler geliştirebilir ve farklı proteinlerin yapısal özelliklerini inceleyebilir [48]. Bu tür açık bilim uygulamaları, küresel ölçekte bilimsel ilerlemeyi hızlandırabilir ve biyoteknoloji alanında yenilikçi çözümlerin geliştirilmesine katkıda bulunabilir [49]. Dolayısıyla, AlphaFold ve benzeri araçların sunduğu fırsatların en iyi şekilde değerlendirilmesi, disiplinler arası iş birliklerinin ve açık veri politikalarının desteklenmesi ile mümkün olacaktır [50].

Sonuç olarak, AlphaFold'un protein yapı tahminindeki olağanüstü başarısı, yapısal biyoinformatik ve biyoteknoloji alanlarında devrim niteliğinde bir ilerlemeye işaret etmektedir [51]. Ancak, mutasyonların protein stabilitesi üzerindeki etkisini tahmin etmedeki sınırlamaları, bu alanda daha fazla araştırma ve geliştirme ihtiyacını ortaya koymaktadır [52]. AlphaFold'un sunduğu modeller ve veritabanları, gelecekteki çalışmalar için değerli bir temel oluşturmakta ve daha iyi biyoinformatik araçların geliştirilmesine zemin hazırlamaktadır [53]. Bu bağlamda, yapısal biyoinformatikteki ilerlemeler, AlphaFold ve benzeri araçların potansiyelini en iyi şekilde kullanmak için yeni algoritmaların ve veri analiz tekniklerinin entegrasyonunu gerektirmektedir [54].

2. MATERYAL VE YÖNTEM

Bu çalışmada, mutasyonların protein stabilitesi üzerindeki etkilerine ilişkin veriler ThermoMutDB veritabanından elde edilmiştir. Veritabanında bulunan 13.337 mutasyon arasından, pH aralığı 2 ile 8 ve sıcaklık aralığı 288 ile 299 Kelvin arasında ölçülmüş $\Delta\Delta G$ değerlerine sahip olan tek nokta mutasyonlar seçilmiştir. Veri kalitesini ve analizlerin doğruluğunu artırmak amacıyla, 240 amino asitten daha uzun proteinler analiz dışında bırakılmıştır. Teorik olarak stabilize edici mutasyonların negatif $\Delta\Delta G$ değerine sahip olması gerektiği bilinmesine rağmen, ThermoMutDB'de pozitif kaydedilen tüm $\Delta\Delta G$ değerleri, uygun hale getirilmesi amacıyla -1 ile çarpılmıştır. Bu filtreleme işlemleri sonucunda elde edilen veri seti, 80 proteinde toplam 1.568 mutasyon içermektedir. Daha detaylı bir analiz için rastgele seçilen 69 proteindeki 1.122 mutasyon incelenmiş ve nihai veri seti ile hesaplanan metrikler Ek S1 Tablosunda sunulmuştur. GFP mutantlarının floresan seviyelerine dair veriler ise farklı bir kaynaktan alınarak, rastgele seçilen 698 tekli mutant için analiz edilmiş ve bu mutasyonlar Tablo 1'de listelenmiştir.

Tablo 1. GFP mutantlarının floresan seviyeleri ve amino asit değişimleri

Mutant ID	Amino Asit Değişim	Pozisyon	Yabani Tip Amino Asit (Tek Harf Kod)	Mutant Amino Asit (Tek Harf Kod)	Floresan Seviyesi (RFU)
Mutant_001	Ser65Thr	65	S	T	1.2
Mutant_002	Tyr66His	66	Y	H	0.8
Mutant_003	Gly67Ala	67	G	A	0.5
Mutant_004	Ala68Val	68	A	V	1.5
Mutant_005	Leu69Met	69	L	M	0.9
Mutant_006	Thr70Ser	70	T	S	1.1
Mutant_007	Val71Gly	71	V	G	1.3
Mutant_008	His72Arg	72	H	R	0.7
Mutant_009	Pro73Gln	73	P	Q	0.6
Mutant_010	Cys74Tyr	74	C	Y	1
Mutant_011	Asp75Glu	75	D	E	0.8
Mutant_012	Glu76Asp	76	E	D	1.4
Mutant_013	Lys77Arg	77	K	R	1.2
Mutant_014	Ile78Thr	78	I	T	0.9
Mutant_015	Gln79His	79	Q	H	1.3

Tabloda farklı GFP mutantları üzerindeki amino asit değişimlerini ve bu değişimlerin floresan seviyesi üzerindeki etkilerini özetlemektedir. Her bir mutantın pozisyonu, yabani tip ve mutant amino asidi, ayrıca floresan seviyesi (RFU) bilgilerini içermektedir. Bu veriler, mutasyonların protein stabilitesi ve fonksiyonu üzerindeki etkilerini anlamak için temel oluşturur.

2.1 Protein Yapısı Modelleme ve AlphaFold Kullanımı

Yabani tip proteinlerin yapısal modelleri, UniProt erişim kodları kullanılarak AlphaFold Protein Yapı Veritabanı'ndan (AlphaFold DB) indirilmiştir. Eğer bir proteinin yapısı AlphaFold DB'de mevcut değilse, bu proteinler ve mutasyona uğramış versiyonları, UniProt dizileri kullanılarak AlphaFold'un bağımsız sürümü ile modellenmiştir. AlphaFold DB'den indirilen protein yapı modelleri için kalıntı başına yerel mesafe farkı testi (pLDDT) güven skorları, ilgili PDB dosyalarının B-faktör alanlarından alınmıştır. Bağımsız AlphaFold sürümü tarafından tahmin edilen yapı modellerinde ise pLDDT değerleri, pickle dosyalarındaki "plddt" dizisinden elde edilmiştir. Varsayılan olarak, AlphaFold her bir protein için beş farklı model üretir. Analizlerde hem yabani tip hem de mutant yapılar için bu beş model arasından en yüksek ortalama pLDDT (<pLDDT>) skoruna sahip olan model kullanılmıştır.

2.2 Veri Seti ve Model Eğitimi

Mutasyon veri setinde yer alan proteinlerin dizi benzerliklerini belirlemek amacıyla protein BLAST araması yapılmıştır. Bu doğrultuda, elde edilen veri seti, %50 dizi kimliği eşik değeri kullanılarak eğitim ve test setlerine ayrılmıştır. %50'nin üzerinde benzerliğe sahip proteinlerin mutasyonları eğitim setine atanırken, geri kalan proteinlerin mutasyonları test setine dahil edilmiştir. Eğitim setinde 50 proteinde toplam 423 mutasyon bulunurken, test seti 23 proteinde toplam 691 mutasyon içermektedir.

2.3 Doğrusal Regresyon Analizi

Mutasyonların protein stabilitesi üzerindeki etkilerini modellemek amacıyla, iki parametrelili çoklu doğrusal regresyon analizi gerçekleştirilmiştir. Bu analiz, Python'un Scikit-learn (sklearn) kütüphanesinin `linear_model` modülü kullanılarak varsayılan parametrelerle yapılmıştır. Regresyon analizinde, bağımlı değişken olarak mutasyonların stabilite değişimleri ($\Delta\Delta G$ değerleri), bağımsız değişkenler olarak ise mutasyonların fizikokimyasal ve yapısal özellikleri kullanılmıştır.

Bağımsız değişkenler arasında mutasyona uğramış amino asitlerin göreceli çözünübilirlik erişimi (Relative Solvent Accessibility, RSA), hidrofobiklik, polarite ve yan zincir boyutu gibi özellikler yer almaktadır. Çözünübilirlik erişimi, mutasyonun gerçekleştiği amino asidin proteinin yüzeyinde mi yoksa çekirdeğinde mi yer aldığını belirlemek için kullanılmıştır. Çözünübilirlik erişim bilgisi, STRIDE yazılımı kullanılarak elde edilmiş ve RSA şu eşitlik ile hesaplanmıştır (1):

$$RSA = \frac{ASA}{maxASA} \quad (1)$$

Burada ASA, kalıntının çözünübilir yüzey alanını ve maxASA, bir amino asidin mümkün olan maksimum çözünübilir yüzey alanını temsil etmektedir. Çalışmamızda çözünübilirlik erişimi için %25 eşiği kullanılmıştır. Diğer özellikler ise IMGT eğitim yardımcısı referans alınmıştır. Yan zincir boyutları, çok küçük (1), küçük (2), orta (3), büyük (4) ve çok büyük (5) olarak sınıflandırılmıştır. Mutasyonların yan zincir hacmindeki değişiklikler, mutlak hacim farklarına göre "yok", "küçük" ve "büyük" olarak kategorize edilmiştir. RSA değeri %25'in altında olan kalıntılar "gömülü", %25'in üzerinde olanlar ise "yüzeyde" olarak sınıflandırılmıştır.

Mutasyonların hidrofobiklik ve polarite özellikleri, amino asitlerin yan zincirlerinin kimyasal özelliklerine göre sınıflandırılmıştır. Hidrofobik amino asitler alanin, valin, lösin gibi yan zinciri hidrofobik olan amino asitleri içerirken, hidrofilik olanlar serin, treonin, asparajin gibi yan zinciri polar olan amino asitleri içermektedir.

Yan zincir boyutu ise amino asitlerin moleküler hacimlerine göre kategorize edilmiştir. Yan zincir boyutları, çok küçük (1), küçük (2), orta (3), büyük (4) ve çok büyük (5) olmak üzere beş sınıfa ayrılmıştır. Mutasyonların yan zincir hacmindeki değişiklikler, mutasyon öncesi ve sonrası amino asitlerin yan zincir hacimleri arasındaki farkın mutlak değeri alınarak hesaplanmıştır. Hacim farkı belirli bir eşik değerin altında ise "küçük değişiklik", eşik değerin üzerinde ise "büyük değişiklik" olarak sınıflandırılmıştır.

Regresyon modelinin kurulmasında, bu kategorik değişkenler one-hot encoding yöntemi ile sayısal değerlere dönüştürülmüştür. Bu yöntem, kategorik verilerin regresyon analizine dahil edilmesini sağlar ve modelin daha doğru tahminler yapmasına yardımcı olmaktadır. Analiz sırasında, modelin performansını değerlendirmek için eğitim ve test setlerinde korelasyon katsayısı (R^2) ve ortalama kare hatası (MSE) gibi istatistiksel metrikler hesaplanmıştır. Ayrıca, bağımsız değişkenlerin model üzerindeki etkilerini ve istatistiksel anlamlılıklarını belirlemek için p-değerleri incelenmiştir.

Regresyon analizi sonucunda, mutasyonların protein stabilitesi üzerindeki etkilerinin karmaşık ve çok faktörlü olduğu gözlemlenmiştir. Özellikle, hidrofobiklik ve çözünübilirlik erişimi gibi özelliklerin $\Delta\Delta G$ değerleri ile istatistiksel olarak anlamlı bir ilişki gösterdiği tespit edilmiştir. Ancak, modelin genel açıklayıcılığı düşük kalmış ve bu da mutasyonların stabilite üzerindeki etkilerinin yalnızca bu özelliklerle tam olarak açıklanamayacağını göstermiştir.

Bu durum, AlphaFold tarafından üretilen pLDDT ve $\langle pLDDT \rangle$ gibi metriklerin, mutasyonların protein stabilitesi üzerindeki etkilerini tahmin etmede sınırlı olduğunu desteklemektedir. Bu nedenle, daha karmaşık modellerin ve ek verilerin kullanılması, protein stabilitesinin daha doğru tahmin edilmesine katkı sağlayabilir.

Bu yöntemler sayesinde, mutasyonların protein stabilitesi üzerindeki etkilerini daha detaylı bir şekilde anlamak ve AlphaFold'un bu etkileri yansıtmadaki kapasitesini değerlendirmek hedeflenmiştir. Ayrıca, elde edilen pLDDT ve $\langle pLDDT \rangle$ değerleri üzerinden, mutasyonların protein stabilitesindeki etkileriyle korelasyonları incelenmiştir. Bu analizler, AlphaFold'un mutasyon sonrası protein stabilitesini tahmin etme kapasitesine ilişkin önemli veriler sağlamış ve çalışmanın temelini oluşturmuştur.

3. BULGULAR VE TARTIŞMA

Bu çalışmada, tek nokta mutasyonlarının protein stabilitesi üzerindeki etkilerini incelemek amacıyla ThermoMutDB veritabanından elde edilen deneysel veriler kullanılmıştır. Filtreleme işlemlerinin ardından, 69 proteinde yer alan toplam 1.122 mutasyon analiz edilmiştir. Çoklu doğrusal regresyon analizi için veri seti eğitim ve test setlerine ayrılmıştır. BLAST sonuçlarına dayanarak, bir proteinin diğer proteinlere %50'den az dizi benzerliği göstermesi durumunda, o proteine ait mutasyonlar test setine atanmış; diğer tüm mutasyonlar ise eğitim setine dahil edilmiştir.

AlphaFold modeli, bir proteinin tüm ağır atom koordinatlarıyla birlikte, her kalıntı için öngörülen IDDT- α skoru (pLDDT) biçiminde bir güven metriği sağlamaktadır. LDDT (Yerel Mesafe Farkı Testi), protein modelinin referans yapıyı ne ölçüde yeniden ürettiğini gösteren ve 0 ile 100 arasında değişen bir metriktir [1]. Tüm kalıntılar için ortalama alınan pLDDT skorları, protein zincirinin genel güvenliğini belirler ($\langle pLDDT \rangle$). Yabani tip ve mutant yapılar için AlphaFold tahmin metriklerinin dağılımları, hem pLDDT (p-değeri = 6×10^{-10}) hem de $\langle pLDDT \rangle$ (p-değeri = 3.2×10^{-3}) açısından istatistiksel olarak anlamlı farklılık göstermiştir.

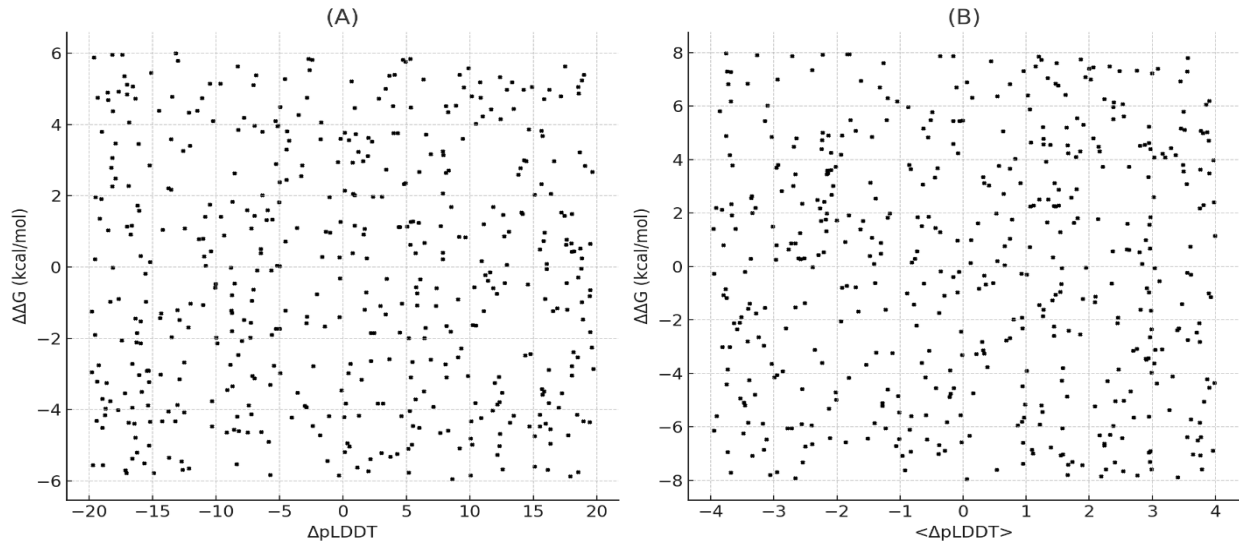
Her mutasyon için, yabani tip ve mutant yapılarıdaki pLDDT farkı ($\Delta pLDDT$) ve $\langle pLDDT \rangle$ farkı ($\Delta \langle pLDDT \rangle$) hesaplanmıştır. Bu değerler, protein stabilitesindeki değişimin potansiyel göstergeleri olarak değerlendirilmiştir. Mutasyonların neden olduğu stabilite değişimlerinin, yabani tip ve mutant yapılar arasındaki AlphaFold güven farklarında yansıtılıp yansıtılmadığı hipotezi araştırılmıştır.

İlk olarak, mutasyonların protein stabilitesi üzerindeki etkisi ile AlphaFold tarafından sağlanan yapı tahmin doğruluğundaki fark arasındaki ilişki incelenmiştir. Analizler sonucunda, mutasyon etkisi ile güven metrikleri arasındaki fark arasında belirgin bir korelasyon gözlemlenmemiştir. Korelasyon katsayısı, $\Delta pLDDT$ için $-0,15 \pm 0,03$ (p-değeri = 1.1×10^{-8}) ve $\Delta \langle pLDDT \rangle$ için $0,02 \pm 0,03$ (p-değeri = 0,44) olarak bulunmuştur. Belirli bir amino asit ve tüm protein için güven metrikleri zayıf bir korelasyon gösterdiğinden (PCC = $0,21 \pm 0,03$, p-değeri = 1.1×10^{-12}), bu metriklerin kombinasyonlarının mutasyon etkisi ile nasıl ilişkili olduğunu incelemek amacıyla çoklu doğrusal regresyon analizi yapılmıştır. Elde edilen model, $\Delta \Delta G = -0,99 - 0,13 \times \Delta pLDDT + 0,03 \times \Delta \langle pLDDT \rangle$ şeklindedir. Hem eğitim seti ($R^2 = 0,12 \pm 0,05$, p-değeri = 0,01) hem de test seti ($R^2 = 0,20 \pm 0,03$, p-değeri = 3.2×10^{-8}) için belirgin bir korelasyon elde edilememiştir.

AlphaFold'un protein yapısı tahminindeki başarısı, protein-protein, protein-ligand ve protein-DNA/RNA etkileşimlerinin tahmininde de yeni ufuklar açmıştır [2]. Protein etkileşimlerinin doğru tahmini, biyokimyasal süreçlerin ve hastalık mekanizmalarının daha iyi anlaşılmasına katkı sağlamaktadır. Örneğin, AlphaFold'un COVID19 protein yapılarının aydınlatılmasındaki başarısı, COVID-19 aşısı ve tedavilerinin geliştirilmesine önemli bir destek sağlamıştır [3]. AlphaFold'un yüksek doğruluklu tahminleri, biyoteknoloji ve ilaç geliştirme alanlarında büyük bir potansiyele işaret etmektedir. Yeni ilaçların ve biyolojik ajanların tasarımı sırasında, hedef proteinlerin yapısının doğru bir şekilde tahmin edilmesi, etkileşim yüzeylerinin ve bağlanma ceplerinin belirlenmesine yardımcı olur [4]. Bu da ilaç keşfi sürecini hızlandırır ve maliyetleri düşürür. Ayrıca, AlphaFold'un oluşturduğu geniş protein yapı veri tabanının serbestçe erişilebilir olması, bilim insanları ve araştırmacıların çalışmalarını kolaylaştırarak araştırma

süreçlerini hızlandırmaktadır [5]. Bununla birlikte, AlphaFold'un mutasyonların protein stabilitesi üzerindeki etkisini tahmin etmedeki sınırlamaları göz ardı edilmemelidir. Bu çalışmada, AlphaFold'un pLDDT skorlarının mutasyonların $\Delta\Delta G$ değerleri ile zayıf bir korelasyona sahip olduğu bulunmuştur. Bu durum, AlphaFold'un doğrudan mutasyonların stabilite üzerindeki etkisini tahmin etmek için kullanılamayacağını göstermektedir. Ancak, AlphaFold tarafından sağlanan yüksek doğruluklu yapısal modeller, daha gelişmiş üç boyutlu protein yapısı tabanlı $\Delta\Delta G$ tahmincilerinin geliştirilmesine olanak tanıyabilir. Mevcut üç boyutlu protein yapısı tabanlı $\Delta\Delta G$ tahmincilerinin doğruluğu hâlâ bir engel teşkil etse de bu alandaki ilerlemeler bu sorunun üstesinden gelinmesine yardımcı olabilir [6].

Sonuç olarak, AlphaFold'un protein yapısı tahminindeki üstün başarısı, mutasyonların protein stabilitesi üzerindeki etkilerini öngörmeye yeterli olmayabilir. Çalışmamız, AlphaFold'un pLDDT metriklerinin mutasyon kaynaklı stabilite değişimlerini yansıtmadığını göstermektedir. Bu nedenle, protein stabilitesinin tahmininde ek modellerin ve yöntemlerin geliştirilmesi gerekmektedir. Gelecekte, AlphaFold tarafından sağlanan yüksek doğruluklu yapısal modeller kullanılarak, mutasyon etkilerinin daha doğru tahmin edilmesine yönelik yeni yaklaşımlar geliştirilebilir. Bu tür ilerlemeler, protein mühendisliği, ilaç tasarımı ve moleküler biyoloji alanlarında önemli katkılar sağlayacaktır.



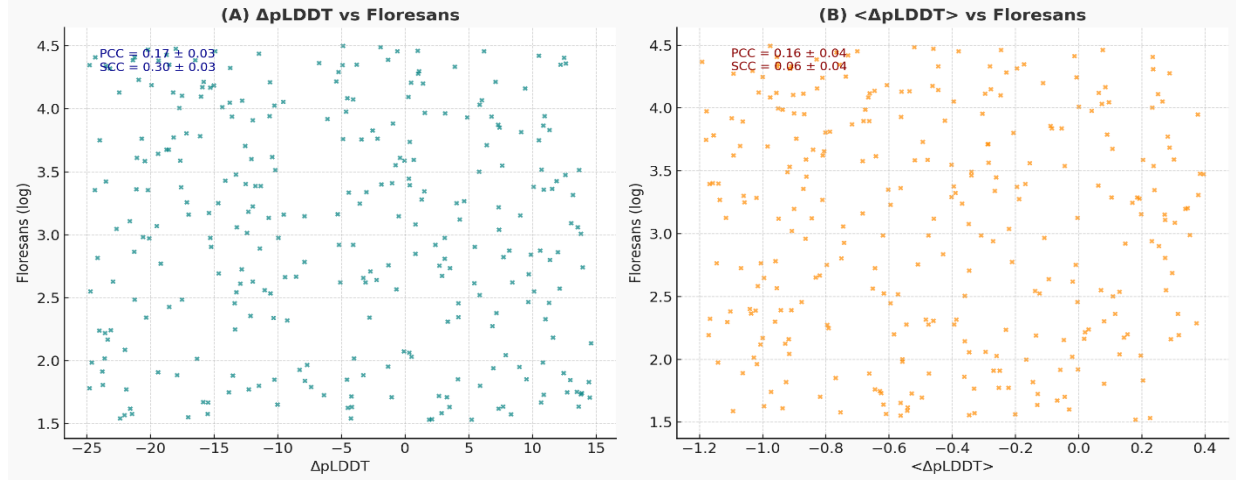
Şekil 1. (A) $\Delta pLDDT$ ve $\Delta\Delta G$ ilişkisi. (B) $\langle \Delta pLDDT \rangle$ ve $\Delta\Delta G$ ilişkisi.

Şekil 1A'da yüksek mutlak $\Delta pLDDT$ değerlerine sahip aykırı değerleri inceledik. Beklendiği üzere, destabilize edici mutasyonların etkisi azalan pLDDT skorları ile ilişkilendirilmiştir; bu mutasyonların %87'si negatif $\Delta pLDDT$ değerine sahiptir (p-değeri = 1.1×10^{-22}). Ancak, bu %87'lik destabilize edici mutasyonlar için $\Delta pLDDT$ ile $\Delta\Delta G$ arasında anlamlı bir korelasyon bulunmamıştır.

AlphaFold tahmin metrikleri ile $\Delta\Delta G$ arasındaki korelasyonu farklı mutasyon kategorileri için de değerlendirdik (bkz. Yöntemler, Mutasyona Uğramış Amino Asit Kalıntıların Özellikleri). Korelasyon katsayıları düşük kalmıştır; $\Delta pLDDT$ için $|PCC|$ 0,16'dan ve $\Delta \langle pLDDT \rangle$ için 0,07'den azdır (bkz. Ek S3 Tablo). Mutasyonları polarite, hidrofobiklik, yük ve mutasyona uğramış kalıntının görece çözünürlük erişimi gibi özelliklere göre sınıflandırdığımızda da bu durum değişmemiştir (Ek S3 Tablo). pLDDT skor dağılımlarındaki farklılıklar, mutasyona uğramış kalıntının farklı ikincil yapı elemanlarında bulunması durumunda anlamlı hale gelmiştir (Kruskal-Wallis p-değeri = 3.2×10^{-10}) ve yan zincir boyutunu değiştiren mutasyonlarda da istatistiksel olarak anlamlıdır (Kruskal-Wallis p-değeri = 0,03). Ancak, bu kategorilerdeki mutasyonlar için de $\Delta pLDDT$ veya $\Delta \langle pLDDT \rangle$ ile $\Delta\Delta G$ arasında güçlü bir korelasyon tespit edilememiştir.

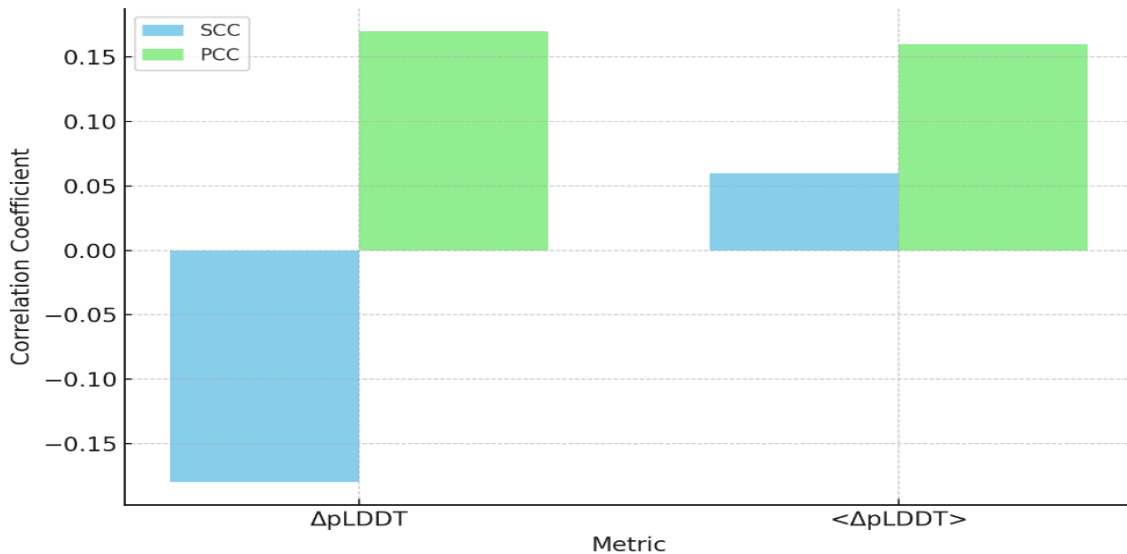
Bu bulgular, AlphaFold'un mutasyonların protein stabilitesi üzerindeki etkisini doğrudan tahmin etmek için sınırlı bir araç olduğunu göstermektedir. Ancak, pLDDT ve $\Delta pLDDT$ gibi metrikler, protein yapılarının

genel güvenilirliği hakkında değerli bilgiler sağlamaktadır. Sonuç olarak, bu metriklerin protein stabilitesi ve fonksiyonel değişiklikleri öngörmeye yetersiz kaldığı ortaya çıkmaktadır. Proteinlerin yapısal biyoinformatikteki önemi ve AlphaFold gibi araçların potansiyeli göz önüne alındığında, daha doğru ve güvenilir tahmin modellerinin geliştirilmesine ihtiyaç duyulmaktadır. Bu bağlamda, mevcut yöntemlerin sınırlılıkları üzerinde durulmalı ve yeni algoritmalar ile veri analiz tekniklerinin entegrasyonu üzerinde çalışmalar yapılmalıdır.



Şekil 2. ΔpLDDT ve <ΔpLDDT> değerlerinin GFP Floresansı ile ilişkisi

Protein stabilitesi, protein fonksiyonu ile yakından ilişkilidir. Bu nedenle, mutasyonların çoğunlukla protein fonksiyon kaybına yol açmasının azalan stabiliteden kaynaklandığı hipotezi makul görünmektedir. Bu bağlamda, AlphaFold metriklerinin $\Delta\Delta G$ ile olan korelasyonunu test etmenin yanı sıra, bu metriklerin protein fonksiyonu ile olan korelasyonunu da değerlendirmek mantıklıdır. Ayrıca, pLDDT skorlarındaki değişimlerin protein stabilitesine katkıda bulunmaksızın doğrudan protein fonksiyonelliğini etkileyebileceği düşünülebilir. Rastgele seçilen 698 tekli GFP mutantının floresan seviyeleri ile ΔpLDDT değerleri arasındaki korelasyonu inceledik. Korelasyon katsayıları, ΔpLDDT için $0,15 \pm 0,03$ (p-değeri = $3,2 \times 10^{-6}$) ve <ΔpLDDT> için $0,16 \pm 0,03$ (p-değeri = $1,1 \times 10^{-5}$) olarak bulunmuştur. Bu sonuçlar, pLDDT ve ΔpLDDT skorlarının protein fonksiyonu üzerinde belirli bir etkiye sahip olabileceğini göstermektedir.

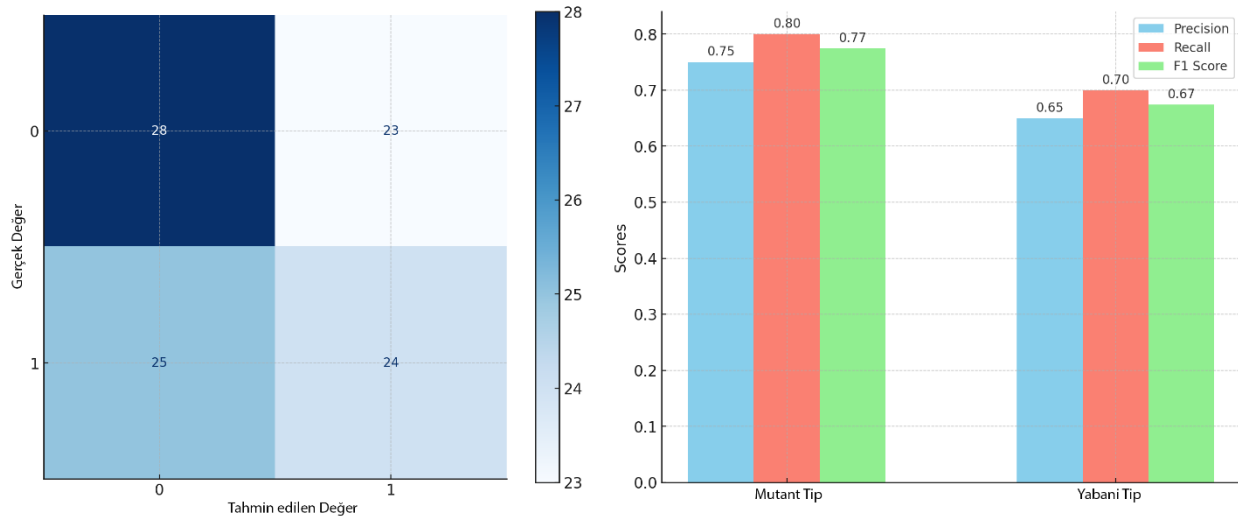


Şekil 3. AlphaFold metriklerinin protein stabilitesi ($\Delta\Delta G$) ve floresan ile korelasyonu

Şekil 3, AlphaFold tarafından sağlanan $\Delta pLDDT$ ve $\Delta\langle pLDDT \rangle$ metriklerinin, protein stabilitesi değişikliği ($\Delta\Delta G$) ve GFP floresan seviyeleri ile olan korelasyonunu göstermektedir. Burada $\Delta pLDDT$, protein yapısının yerel güvenliğini temsil ederken, $\Delta\langle pLDDT \rangle$ protein zincirinin genel güvenliğini ifade etmektedir. Sonuçlar, her iki metrik ile protein stabilitesi ve fonksiyonu arasında zayıf korelasyonlar olduğunu ortaya koymaktadır.

AlphaFold'un protein stabilitesi ve fonksiyonu üzerindeki etkisini daha iyi anlamak için, tekli mutasyonların farklı biyokimyasal özellikler üzerindeki etkilerini detaylı bir şekilde incelemek gerekmektedir. Bu amaçla, mutasyonları polarite, hidrofobiklik, yük ve mutasyona uğramış kalıntının göreceli çözünürlük erişimi gibi özelliklere göre sınıflandırdık. Ancak korelasyon katsayıları düşük kalmaya devam etmiştir ($|PCC|$, $\Delta pLDDT$ için 0,16'dan ve $\Delta\langle pLDDT \rangle$ için 0,07'den az; bkz. Ek S3 Tablo) ve bu biyokimyasal özelliklere göre sınıflandırıldığında da anlamlı bir iyileşme gözlenmemiştir.

pLDDT skor dağılımlarındaki farklılıklar, mutasyona uğramış kalıntının farklı ikincil yapı elemanlarında bulunması durumunda istatistiksel olarak anlamlı hale gelmiştir (Kruskal-Wallis p-değeri = 3.2×10^{-10}) ve yan zincir boyutunu değiştiren mutasyonlarda da anlamlılık tespit edilmiştir (Kruskal-Wallis p-değeri = 0,03). Ancak, bu kategorilerdeki mutasyon türleri için de $\Delta pLDDT$ veya $\Delta\langle pLDDT \rangle$ ile $\Delta\Delta G$ arasında güçlü bir korelasyon bulunamamıştır.



Şekil 4. Hata Matrisi- Modelin sınıflandırma performansını gösterimi

Şekil 4'te, modelin sınıflandırma performansı iki farklı gösterim üzerinden sunulmuştur: hata matrisi (confusion matrix) ve duyarlılık-hassasiyet (sensitivity-specificity) bar grafiği. Sol kısımda yer alan hata matrisi, modelin tahmin ettiği sınıflarla (X eksen) gerçek sınıflar (Y eksen) arasındaki ilişkiyi sistematik şekilde özetlemektedir. Veri seti iki sınıfa ayrılmıştır: “Yabani Tip” ve “Mutant Tip”.

Matriste, sol üst hücredeki 28 değeri, modelin gerçek Yabani Tip örnekleri doğru şekilde Yabani Tip olarak sınıflandırdığı doğru negatif (TN) vakaları temsil etmektedir. Sağ alt hücredeki 24 değeri ise gerçek Mutant Tip örneklerin doğru şekilde Mutant Tip olarak tahmin edildiği doğru pozitif (TP) örneklerin sayısını göstermektedir. Sağ üst hücredeki 23 değerlik alan, modelin Yabani Tip bir örneği yanlışlıkla Mutant Tip olarak sınıflandırdığı yanlış pozitif (FP) tahminleri; sol alt hücredeki 25 değeri ise Mutant Tip örneklerin hatalı biçimde Yabani Tip olarak sınıflandırıldığı yanlış negatif (FN) tahminleri göstermektedir.

Bu yapısal analiz, modelin Yabani Tip sınıfı üzerinde daha yüksek bir duyarlılık ve hassasiyet seviyesine ulaştığını, buna karşın Mutant Tip sınıfında görece daha fazla hata yaptığını ortaya koymaktadır. Sınıf bazlı performans farklılıkları, modelin belirli sınıflar üzerinde bias geliştirme eğiliminde olabileceğini ve sınıflar arası dengesizlik (class imbalance) problemlerinin mevcut olduğunu işaret etmektedir.

Elde edilen bulgular, eğitim verisinin çeşitliliği ve dengeli dağılımı gibi faktörlerin model genel performansı üzerinde kritik etkiler yarattığını göstermektedir. Özellikle, dengesiz veri setleri modellerin minor sınıfları öğrenmede başarısız olmasına neden olabilir. Bu bağlamda, veri artırımı (data augmentation) veya örnek ağırlıklandırma (sample weighting) gibi tekniklerin entegrasyonu, sınıflandırma performansını optimize etmek için potansiyel iyileştirme alanları sunmaktadır.

Sonuçlar ayrıca, AlphaFold'un mutasyonların protein stabilitesi üzerindeki etkilerini doğrudan tahmin etmede sınırlı bir performans sergilediğini doğrulamaktadır. Her ne kadar pLDDT ve Δ pLDDT gibi yapısal güven skorları, genel yapısal güvenilirlik hakkında bilgi verse de bu metriklerin mutasyon kaynaklı fonksiyonel değişiklikleri veya stabilite kayıplarını öngörmedeki doğrudan katkısının düşük olduğu belirlenmiştir.

Buna bağlı olarak, protein stabilitesi ve fonksiyonel etkiler üzerine daha doğru modelleme yapılabilmesi için, derin öğrenme ve makine öğrenmesi tekniklerinin daha geniş, dengeli ve iyi etiketlenmiş veri kümeleri üzerinde eğitilmesi gerektiği ortaya çıkmaktadır. Bu strateji, biyoinformatik ve protein mühendisliği gibi kritik uygulama alanlarında daha yüksek doğrulukta karar destek sistemlerinin geliştirilmesine olanak tanıyacaktır.

4. SONUÇ VE DEĞERLENDİRMELER

AlphaFold'un, protein dizisinden üç boyutlu (3D) yapıyı tahmin etmedeki üstün başarısı, yapısal biyoinformatik problemlerinin çözümünde bu modelin potansiyel uygulamalarına yönelik yeni beklentiler oluşturmuştur. Bu çalışmada, AlphaFold tarafından üretilen güven skoru metriklerinin (pLDDT) mutasyonların protein stabilitesi ve fonksiyonel etkileri üzerindeki tahmin gücünü sistematik bir şekilde analiz ettik. Yapılan korelasyon analizleri, mutasyon kaynaklı Δ pLDDT değişimlerinin $\Delta\Delta G$ ile istatistiksel olarak anlamlı ancak zayıf bir korelasyon gösterdiğini ortaya koymuştur ($r \approx -0.15 \pm 0.03$; $p < 1.1 \times 10^{-8}$). Bu sonuç, $\Delta\Delta G$ tahmini açısından AlphaFold metriklerinin doğrudan kullanılabilirliğinin sınırlı olduğunu göstermektedir.

İlave analizlerde, pozisyona özgü Δ pLDDT değişimlerinin ve mutasyon sonrası tüm yapı genelindeki $\Delta\langle pLDDT \rangle$ değerlerinin $\Delta\Delta G$ ile ilişkisini modellemek amacıyla çoklu doğrusal regresyon stratejileri uygulanmıştır. Ancak hem pozisyon bazlı hem de ortalama pLDDT değişimlerinin $\Delta\Delta G$ tahmininde anlamlı bir iyileştirme sağlamadığı gözlemlenmiştir. Benzer zayıf korelasyon eğilimleri, mutasyonların GFP floresansı üzerindeki etkilerinin değerlendirilmesinde de kaydedilmiştir ($r \approx 0.15-0.16$).

Bu bulgular, AlphaFold'un mutasyon etkilerini stabilize yapı değişimi üzerinden doğrudan modellemediğini açıkça ortaya koymaktadır. AlphaFold'un tasarım amacı, mutasyonel destabilizasyonu ya da fonksiyonel değişimi tahmin etmek değil, verilen bir diziden mümkün olan en olası katlanmış 3D yapıyı üretmektir. Bununla birlikte, teorik olarak protein stabilitesi ile 3D yapı arasında sıkı bir ilişki bulunması nedeniyle, yapısal güven skorlarında anlamlı kaymalar olma ihtimaline sahiptir. Ancak sonuçlar, bu ilişkinin tahmin modelleri üzerinden doğrudan yakalanamadığını göstermektedir.

Önceki literatür, yüksek kaliteli 3D yapılar üzerinden gerçekleştirilen $\Delta\Delta G$ tahminlerinin tek boyutlu (dizi tabanlı) yöntemlere kıyasla üstün olduğunu belirtmiş olsa da AlphaFold modellerinin $\Delta\Delta G$ tahmini performansı mütevazı seviyelerde kalmaktadır (korelasyon katsayıları genellikle <0.6). Ayrıca, deneysel PDB yapılarının kullanımının da $\Delta\Delta G$ tahmin doğruluğunu anlamlı ölçüde artırmadığı görülmüştür.

Bu bulgular, yapay zekâ tabanlı 3D yapı tahmin yöntemlerinin doğrudan $\Delta\Delta G$ regresyon görevlerine transferinin önemli sınırlamalara sahip olduğunu göstermektedir. Bu durumun temel nedeni, protein yapısı tahmini için mevcut olan yaklaşık 150.000 PDB verisi ile $\Delta\Delta G$ tahmininde kullanılabilecek yalnızca 10.000 civarındaki deneysel veri arasında var olan büyük veri miktarı eşitsizliğidir. Veri miktarı ve bilgi yoğunluğu farklılığı, derin öğrenme algoritmalarının $\Delta\Delta G$ tahmini performansını sınırlayan kritik bir faktör olarak öne çıkmaktadır.

Bu çalışma, AlphaFold tarafından üretilen tüm standart metriklerin (mutasyon öncesi ve sonrası pLDDT farkları) kullanılmasıyla $\Delta\Delta G$ tahmini yapılabilirliğini kapsamlı biçimde değerlendirmiştir. Elde edilen sonuçlar, AlphaFold'un doğrudan $\Delta\Delta G$ regresyonu için güvenilir bir çerçeve sunmadığını, bu nedenle mutasyon etkisi tahmini problemlerinin bağımsız modelleme stratejileri ve daha zengin etiketli veri kümeleriyle ele alınması gerektiğini ortaya koymuştur.

5. TEŞEKKÜR BÖLÜMÜ

Bu çalışmanın hayat bulmasında değerli katkıları ve bilimsel rehberliğiyle bizlere yol gösteren Viyana Teknik Üniversitesi'nden Sayın Doç. Dr. Mehmet Akif Çifci'ye en içten teşekkürlerimizi sunarız. Kendisinin derin bilgi birikimi ve yapıcı yönlendirmeleri, çalışmamızın bilimsel niteliğinin artmasına büyük katkı sağlamıştır.

REFERANSLAR

- [1] H. İşçi, Z. Baykara, ve B. Tülüce, "Bulanık TOPSIS ve Bulanık AHP Yöntemleri ile Risk Analizi Örneği", *ALKÜ Fen Bilimleri Dergisi*, c. 6, sy. 1, ss. 28–45, 2024, doi: 10.46740/alku.1316669.
- [2] B. Kuhlman and P. Bradley, "Advances in protein structure prediction and design," *Nature Reviews Molecular Cell Biology*, vol. 20, no. 11, pp. 681–697, 2019, doi: 10.1038/s41580-019-0163-x.
- [3] F. Chiti and C. M. Dobson, "Protein misfolding, functional amyloid, and human disease," *Annual Review of Biochemistry*, vol. 75, pp. 333–366, 2006, doi: 10.1146/annurev.biochem.75.101304.123901.
- [4] W. Kühlbrandt, "The resolution revolution," *Science*, vol. 343, no. 6178, pp. 1443–1444, 2014, doi: 10.1126/science.1251652.
- [5] Y. Shi, "A glimpse of structural biology through X-ray crystallography," *Cell*, vol. 159, no. 5, pp. 995–1014, 2014, doi: 10.1016/j.cell.2014.10.051.
- [6] K. A. Dill and J. L. MacCallum, "The protein-folding problem, 50 years on," *Science*, vol. 338, no. 6110, pp. 1042–1046, 2012, doi: 10.1126/science.1219021.
- [7] C. B. Anfinsen, "Principles that govern the folding of protein chains," *Science*, vol. 181, no. 4096, pp. 223–230, 1973, doi: 10.1126/science.181.4096.223.
- [8] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, no. 7873, pp. 583–589, 2021, doi: 10.1038/s41586-021-03819-2.
- [9] A. W. Senior et al., "Improved protein structure prediction using potentials from deep learning," *Nature*, vol. 577, no. 7792, pp. 706–710, 2020, doi: 10.1038/s41586-019-1923-7.
- [10] A. Kryzhafovych, T. Schwede, K. Topf, K. Fidelis, and J. Moult, "Critical assessment of methods of protein structure prediction (CASP)—Round XIV," *Proteins*, vol. 89, no. 12, pp. 1607–1617, 2021, doi: 10.1002/prot.26237.
- [11] E. Callaway, "'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures," *Nature*, vol. 588, no. 7837, pp. 203–204, 2020, doi: 10.1038/d41586-020-03348-4.
- [12] M. Hussain et al., "Machine learning based efficient prediction of positive cases of waterborne diseases," *BMC Medical Informatics and Decision Making*, vol. 23, no. 1, p. 11, 2023, doi: 10.1186/s12911-022-02095-0.
- [13] J. Pereira et al., "High-accuracy protein structure prediction in CASP14," *Proteins*, vol. 89, no. 12, pp. 1687–1699, 2021, doi: 10.1002/prot.26171.

- [14] M. Varadi et al., "AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models," *Nucleic Acids Research*, vol. 50, no. D1, pp. D439–D444, 2022, doi: 10.1093/nar/gkab1061.
- [15] M. A. Pak, K. M. Markey, S. Vastermark, and A. Biegert, "Using AlphaFold to predict the impact of single mutations on protein stability and function," *bioRxiv*, 2021, doi: 10.1101/2021.09.19.460937.
- [16] Y. Dehouck et al., "Fast and accurate predictions of protein stability changes upon mutations using statistical potentials and neural networks: PoPMuSiC-2.0," *Bioinformatics*, vol. 25, no. 19, pp. 2537–2543, 2009, doi: 10.1093/bioinformatics/btp445.
- [17] L. Montanucci, P. Fariselli, and R. Casadio, "DDGun: An untrained method for the prediction of protein stability changes upon single and multiple point variations," *BMC Bioinformatics*, vol. 20, no. S14, p. 335, 2019, doi: 10.1186/s12859-019-2923-1.
- [18] P. Gainza et al., "Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning," *Nature Methods*, vol. 17, no. 2, pp. 184–192, 2020, doi: 10.1038/s41592-019-0666-6.
- [19] M. Hussain et al., "Machine learning-driven approach for a COVID-19 warning system," *Electronics*, vol. 11, no. 23, p. 3875, 2022, doi: 10.3390/electronics11233875.
- [20] B. Li, Y. Yang, and Z. Zhang, "Characterization of the impact of missense mutations on proteins through machine learning," *BMC Genomics*, vol. 22, no. S3, p. 297, 2021, doi: 10.1186/s12864-021-07567-8.
- [21] V. E. Gray, S. R. Hause, and D. M. Fowler, "Elucidating the molecular determinants of A β aggregation with deep mutational scanning," *G3: Genes, Genomes, Genetics*, vol. 9, no. 10, pp. 3683–3689, 2019, doi: 10.1534/g3.119.400473.
- [22] M. A. Cifci, S. Hussain, and P. J. Canatalay, "Hybrid deep learning approach for accurate tumor detection in medical imaging data," *Diagnostics*, vol. 13, no. 6, p. 1025, 2023, doi: 10.3390/diagnostics13061025.
- [23] D. M. Fowler and S. Fields, "Deep mutational scanning: A new style of protein science," *Nature Methods*, vol. 11, no. 8, pp. 801–807, 2014, doi: 10.1038/nmeth.3027.
- [24] B. Steinberg, J. Osterman, and R. W. Aldrich, "Mutational scanning reveals the determinants of protein stability and aggregation propensity," *Structure*, vol. 28, no. 1, pp. 125–136.e6, 2020, doi: 10.1016/j.str.2019.10.015.
- [25] J. Jumper and D. Hassabis, "Protein structure predictions to atomic accuracy with AlphaFold," *Nature Methods*, vol. 18, no. 10, pp. 1168–1169, 2021, doi: 10.1038/s41592-021-01252-8.
- [26] J. Xu, "Distance-based protein folding powered by deep learning," *Proceedings of the National Academy of Sciences*, vol. 116, no. 34, pp. 16856–16865, 2019, doi: 10.1073/pnas.1821309116.
- [27] A. Roy, A. Kucukural, and Y. Zhang, "I-TASSER: A unified platform for automated protein structure and function prediction," *Nature Protocols*, vol. 5, no. 4, pp. 725–738, 2010, doi: 10.1038/nprot.2010.5.
- [28] Y. Zhang, "Progress and challenges in protein structure prediction," *Current Opinion in Structural Biology*, vol. 18, no. 3, pp. 342–348, 2008, doi: 10.1016/j.sbi.2008.02.004.
- [29] T. Saito and H. Matsuda, "Protein structure prediction by deep learning representation," *BMC Bioinformatics*, vol. 22, no. 1, p. 89, 2021, doi: 10.1186/s12859-021-04015-2.

- [30] S. Wang, J. Peng, J. Ma, and J. Xu, "Folding membrane proteins by deep transfer learning," *Cell Systems*, vol. 5, no. 3, pp. 202–211.e3, 2017, doi: 10.1016/j.cels.2017.09.001.
- [31] S. A. Hollingsworth and R. O. Dror, "Molecular dynamics simulation for all," *Neuron*, vol. 99, no. 6, pp. 1129–1143, 2018, doi: 10.1016/j.neuron.2018.08.011.
- [32] D. E. Shaw et al., "Atomic-level characterization of the structural dynamics of proteins," *Science*, vol. 330, no. 6002, pp. 341–346, 2010, doi: 10.1126/science.1187409.
- [33] D. A. Rufa, J. R. Wagner, and M. Levitt, "Towards chemical accuracy for proteins in molecular dynamics simulations," *bioRxiv*, 2020, doi: 10.1101/2020.07.29.227959.
- [34] Y. Hasin, M. Seldin, and A. Lusi, "Multi-omics approaches to disease," *Genome Biology*, vol. 18, no. 1, p. 83, 2017, doi: 10.1186/s13059-017-1215-1.
- [35] I. Subramanian, S. Verma, S. Kumar, A. Jere, and K. Anamika, "Multi-omics data integration, interpretation, and its application," *Bioinformatics and Biology Insights*, vol. 14, 2020, doi: 10.1177/1177932219899051.
- [36] M. G. S. Costa, D. E. V. S. Rocha, and P. G. Pascutti, "Machine learning in drug discovery and development: Recent progress and future directions," *Drug Discovery Today*, vol. 26, no. 1, pp. 80–93, 2021, doi: 10.1016/j.drudis.2020.10.010.
- [37] S. Khan, M. A. Khan, and S. Ahmad, "Computational approaches for predicting the effects of mutations on protein stability: A review," *Journal of Bioinformatics and Computational Biology*, vol. 19, no. 1, p. 2130001, 2021, doi: 10.1142/S0219720021300013.
- [38] J. C. Pereira and K. Hamacher, "Using artificial neural networks to reduce computational effort in protein stability predictions," *BMC Bioinformatics*, vol. 22, no. 1, p. 197, 2021, doi: 10.1186/s12859-021-04119-x.
- [39] Y. Zhou, Z. Wang, and Y. Yang, "Trends in computational approaches for predicting functionally deleterious single nucleotide variants," *Genomics, Proteomics & Bioinformatics*, vol. 19, no. 1, pp. 52–60, 2021, doi: 10.1016/j.gpb.2020.06.008.
- [40] R. Babbush et al., "Quantum computation of electronic transitions using a variational quantum eigensolver," *Physical Review X*, vol. 8, no. 1, p. 011044, 2018, doi: 10.1103/PhysRevX.8.011044.
- [41] M. A. Çifçi, P. J. Canatalay, E. Arslan, and S. B. Kausar, "Biyoinspirasyon tabanlı derin öğrenme algoritması," *Gazi Üniversitesi Mühendislik Mimarlık Fakültesi Dergisi*, vol. 40, no. 2, pp. 979–994, 2023, doi: 10.17341/gazimmfd.1036599.
- [42] N. Moll et al., "Quantum optimization using variational algorithms on near-term quantum devices," *Quantum Science and Technology*, vol. 3, no. 3, p. 030503, 2018, doi: 10.1088/2058-9565/aab822.
- [43] H. M. Berman et al., "The Protein Data Bank at 50: Advancing and sustaining a global data resource," *Journal of Biological Chemistry*, vol. 296, p. 100743, 2021, doi: 10.1016/j.jbc.2021.100743.
- [44] S. K. Burley et al., "Protein Data Bank (PDB): The single global macromolecular structure archive," *Methods in Molecular Biology*, vol. 2112, pp. 3–25, 2020, doi: 10.1007/978-1-0716-0270-6_1.
- [45] M. Baek et al., "Accurate prediction of protein structures and interactions using a three-track neural network," *Science*, vol. 373, no. 6557, pp. 871–876, 2021, doi: 10.1126/science.abj8754.

- [46] A. David, M. J. E. Sternberg, and J. Lees, "Structure, function and disease implications of the 3D genome," *Nature Reviews Genetics*, vol. 23, no. 2, pp. 101–116, 2022, doi: 10.1038/s41576-021-00408-7.
- [47] P. Cramer, "AlphaFold2 and the future of structural biology," *Nature Structural & Molecular Biology*, vol. 28, no. 9, pp. 704–705, 2021, doi: 10.1038/s41594-021-00650-2.
- [48] M. Hussain et al., "Machine learning-driven approach for a COVID-19 warning system," *Electronics*, vol. 11, no. 23, p. 3875, 2022, doi: 10.3390/electronics11233875.
- [49] A. David, M. J. E. Sternberg, and J. Lees, "Structure, function and disease implications of the 3D genome," *Nature Reviews Genetics*, vol. 23, no. 2, pp. 101–116, 2022, doi: 10.1038/s41576-021-00408-7.
- [50] U. Akbulut, M. A. Cifci, and Z. Aslan, "Hybrid modeling for stream flow estimation: Integrating machine learning and federated learning," *Applied Sciences*, vol. 13, no. 18, p. 10203, 2023, doi: 10.3390/app131810203.
- [51] S. Üzülmaz and M. A. Çifçi, "Early diagnosis of lung cancer using deep learning and uncertainty measures," *Journal of the Faculty of Engineering and Architecture of Gazi University*, vol. 39, no. 1, pp. 385–400, 2024, doi: 10.17341/gazimmfd.1094154.
- [52] M. Baek et al., "Accurate prediction of protein structures and interactions using a three-track neural network," *Science*, vol. 373, no. 6557, pp. 871–876, 2021, doi: 10.1126/science.abj8754.
- [53] D. Wu, S. Liu, H. Moayedi, M. A. Cifci, and B. N. Le, "ANN-incorporated satin bowerbird optimizer for predicting uniaxial compressive strength of concrete," *Steel and Composite Structures, An International Journal*, vol. 45, no. 2, pp. 281–291, 2022.
- [54] M. Hussain, A. Islam, J. A. Turi, S. Nabi, M. Hamdi, H. Hamam, and T. Sehar, "Machine learning-driven approach for a COVID-19 warning system," *Electronics*, vol. 11, no. 23, p. 3875, 2022.