Organik Yarı İletkenlerin HOMO ve LUMO Enerji Seviyelerinin Yapay Sinir Ağları ile Tahmini

Simla Burcu HARMA
Computer Eng. Department
TOBB ETU
Ankara, Turkey
simlaharma@gmail.com

Eşref Oğuzhan Yıldırım
Computer Eng. Department
TOBB ETU
Ankara, Turkey
e.oguzhan.yildirim@gmail.com

Abstract

Bu projede moleküllerin kararlılığını, yarı-iletkenlik özelliklerini belirleyen HOMO ve LUMO orbitallerinin enerji değerleri, organik yarı iletkenlerde tahmin edilmeye çalışılmıştır. Büyük moleküler kütüphanelerde zaman alıcı ve maliyetli olan bu işlem, yapay sinir ağları ile hızlandırılmıştır. Bu tahminler, Multi Layer Perceptron, Generalized Regression Neural Networks ve Support Vector Regression modelleri kullanılmış; Feature Selection, Lasso Regularization ve early stopping teknikleriyle sonuçlar iyileştirilmeye çalışılmıştır. Sonuç olarak bu alanda en az hatayı veren modelin Feature Selection uygulanmış dataseti input alan Support Vector Regression olduğu görülmüştür.

1 Introduction

OYİ'lerin elektronik yapılarının kuantum kimyasal hesaplarla elde edilmesi mümkündür. Ancak geniş kapsamlı büyük moleküler kütüphanelerde bu hesapların yapılması zordur, ancak IBM WCG (www.worldcommunitygrid.org) gibi çok büyük hesaplama imkanlarıyla mümkündür. Bu sebeple son yıllarda bu hesapları hızlandıracak makine öğrenmesi yaklaşımlarına ilgi artmıştır. Bu motivasyonla, bu projede küçük bir molekül seti ile (147 molekül) yapay sinir ağlarının (YSA) elektronik yapı tahmin gücünü araştırdık. Molekül seti, moleküler imzalarla (Faulon) reprezente edilir. Bunlar YSA inputları olarak kullanılarak HOMO ve LUMO enerji seviyeleri tahmin edilir. Moleküllerin bir kısmı YSA'nın (training set) eğitimi için, diğer kısmı testi (test set) için kullanılır. Molekül setleri seçilirken moleküllerin benzer olmaması sağlanır (Tanimoto uzaklığı kullanılabilir). çıktılar, tahmin edilen enerji seviyelerinin kuantum kimyasal olarak hesaplanan enerji seviyelerine regresyonu ile kontrol edilebilir.

Moleküllerin kararlılığını, yarı-iletkenlik özelliklerini belirleyen HOMO ve LUMO orbitallerinin enerji değerlerini, organik yarı iletkenlerde tahmin etme üzerine kurulu projemize benzer bir çalışma, organik fotovoltaikler için 2015 yılında Pyzer- Knapp vd. tarafından yapılmıştır. [3]"Learning from the Harvard Clean Energy Project: The Use of Neural Networks to Accelerate Materials Discovery " isimli makalelerinde belirttikleri gibi "overfitting"den kaçınmak için "early stopping" tekniğini kullanmışlardır. 200.000 molekül ile eğitim yapıp 50.000 molekül üzerinde test etmişler ve maliyet fonksiyonu olarak Mean Absolute Error(MAE) kullanmışlardır.

Oluşturduğumuz YSA modellerini test ederken 147 molekül ve 115 özellikten oluşan dataset kullanılmış ve buna dair ayrıntılar ileriki bölümlerde verilmiştir. Problemin tanımını vererek başlayıp sonrasında implementasyon tekniklerini ve nasıl geliştirdiğimizi anlattık. Elde ettiğimiz sonuçlar ve sonuçlarla ilgili hesaplama ve yorumlara yer verdik.

2 Description of the Problem and Problem Domain

Organik yarı iletkenler (OYİ) konjuge moleküler veya polimer yapılarından oluşurlar. Bu konjuge sistemleri sayesinde elektrolüminesen, fotovoltaik etkiler ya da yarı iletkenlik gibi (opto)elektronik özellikler gösterirler. Özellikle plastiklerle uyumlu olmaları, çözeltiden hazırlanabilmeleri ve çok yüksek işleme sıcaklıkları gerektirmemeleri sebebiyle malzeme araştırmalarında temiz enerjiden biyoelektronik materyallere uzanan geniş bir yelpazede önemli bir yere sahiptirler. Ayrıca, zengin karbon kimyası yeni OYİ)'lerin dizaynı için imkan sağlar.

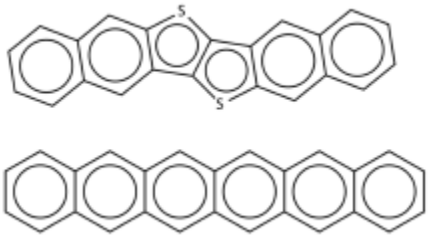


Figure 1: Bazı organik yarı iletkenler.

OYİ'lerin elektronik yapılarının kuantum kimyasal hesaplarla elde edilmesi mümkündür. Ancak geniş kamsamlı büyük moleküler kütüphanelerde bu hesapların yapılması zordur, ancak IBM WCG (www.worldcommunitygrid.org) gibi çok büyük hesaplama imkanlarıyla mümkündür. Bu sebeple son yıllarda bu hesapları hızlandıracak makine öğrenmesi yaklaşımlarına ilgi artmıştır. Bu motivasyonla, bu projede küçük bir molekül seti ile (150 molekül) yapay sinir ağlarının (YSA) elektronik yapı tahmin gücünü araştırdık. Molekül seti, moleküler imzalarla (Faulon) reprezente edilir. Bunlar YSA inputları olarak kullanılarak HOMO ve LUMO enerji seviyeleri tahmin edilir. Moleküllerin bir kısmı YSA'nın (training set) eğitimi için, diğer kısmı testi (test set) için kullanılır. Molekül setleri seçilirken moleküllerin benzer olmaması sağlanır (Tanimoto uzaklığı kullanılabilir). çıktılar, tahmin edilen enerji seviyelerinin kuantum kimyasal olarak hesaplanan enerji seviyelerine regresyonu ile kontrol edilebilir.

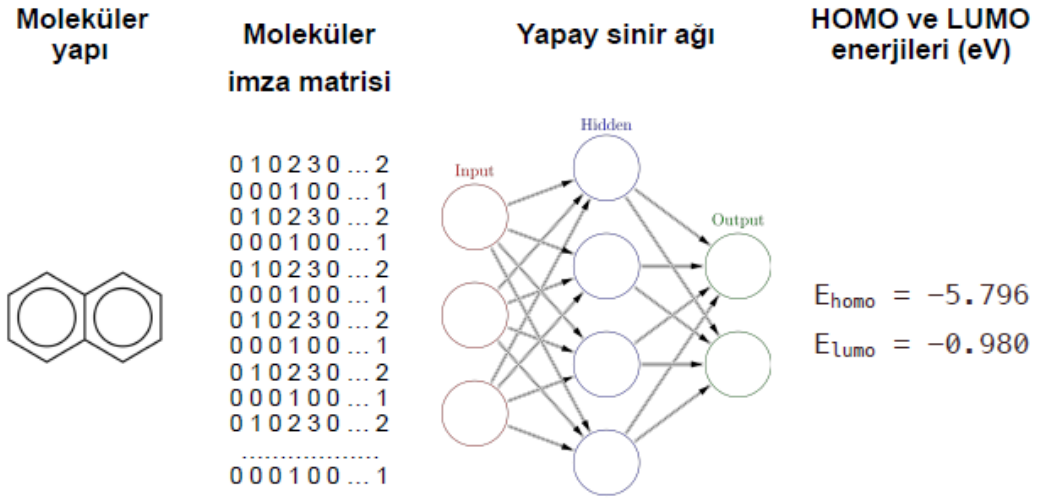


Figure 2: Projede kullanılan verilerin modellemesi.

Molekül setindeki farklı atomlar ve her atomun moleküller içindeki farklı bağlantılarını reprezente eden bir moleküler imza seti oluşturulur. Her molekül bu basis'de her imzadan kaç tane olduğunu gösteren doğal sayı vektörleri halinde ifade edilir. Bu imzalar moleküllerin yapısal tanımlayıcıları olarak kullanılır.

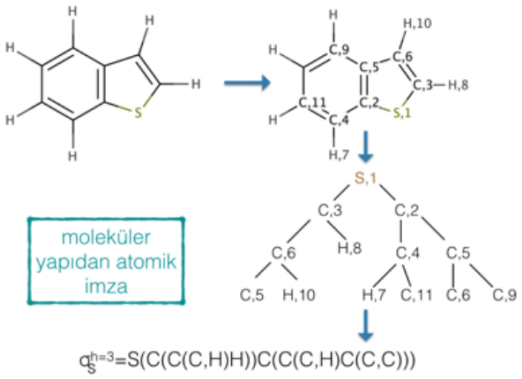


Figure 3: Benzo(b)tiyofen molekülünde Sülfür atomu için h=3 derinliğindeki atomik imzanın eldesi.

Hesaplamak istediğimiz parametreler moleküllerin HOMO (dolu olan, en yüksek enerjili moleküler orbital) ve LUMO (boş olan, en düşük enerjili moleküler orbital) orbitallerinin enerji değeridir. Bu enerji seviyeleri moleküllerin kararlılığını, yarıiletkenlik özelliklerini belirlediği için önemlidir. Aynı zamanda da yapabileceğimiz en basit tahminlerden birisidir.

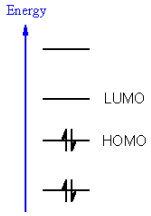


Figure 4: Enerji seviyeleri modellemesi.

3 Description of Implemented Technique/ Improvements

147 molekülden oluşan veri setimizi TOBB ETÜ Tıp Fakültesi öğretim üyelerinden Yrd. Doç. Dr. Şule Atahan Evrenk hazırlamıştır. Molekül setini hazırlarken kullanılan organik yarı iletken moleküllerinin elektronik yapılarını etkileyecek özellikler belirlenmiştir(Feature Extraction). İlk başta 500 özelliğe(feature) sahip olan molekül seti moleküllerin farklı kimyasal yönleri ele alınarak hocamız tarafından 16 özelliğe düşürülmüştür. İlk aşamada 16 boyutlu 147 girdi(moleküler imza matrisi) Python ile yazdığımız çok Katmanlı Yapay Sinir Ağına çeşitli parametrelerle(Gizli katmandaki nöron sayısı, learning rate, iterasyon sayısı) verilmiştir. Buradan elde ettiğimiz sonuçlar 4 'te ayrıntılı olarak açıklanmıştır. Projenin ara rapor dönemine kadar yapılan çalışmalar bu şekildedir. Daha iyi sonuçlar alabilmek adına, bir sonraki aşamaya 115 özellikle devam edilmiş ve öncekilere göre daha başarılı olduğu görülmüştür. Veri toplama ve düzenleme metodumuz yukarıda anlatıldığı gibidir. Projemizin uygulamasında kullandığımız yapay sinir ağı modelleri Multi Layer Perceptron(1 ve 2 hidden layer ile), Generalized Regression, Support Vector Machines olup ayrıca aldığımız sonuçları iyileştirebilmek adına Feature Selection metodları denenmiştir. Kodlamada bazı modellerin implementasyonunda(Multi Layer Perceptron, Support Vector Machines) ve veri setini program için uygun hale getirmede Python programlama dili kullanılmış; bazı kısımlar ise MATLAB nntool ile yapılmıştır. Bunlar yapılırken en büyük problemlerden biri olan "overfitting"den kaçınmak[1] için çeşitli yöntemler(Feature Selection, regularization, early stopping[2] gibi) denenmiştir. Sonunda bütün YSA modelleri karşılaştırılmış ve Figure 5'teki grafik elde edilmiştir.

4 Results and Evaluation/Discussion of the Results

HOMO ve LUMO enerji seviyeleri için yukarıda belirtilen modellerde çeşitli parametreler uygulanmıştır.

Multi Layer Perceptron(MLP) modelinde ilk başta tek hidden layer(25 nöronlu), sonra iki hidden layer(25-15 nöronlu) kullanılmıştır. HOMO değerleri için 2 hidden layer'lı MLP'nin, LUMO değerleri için ise tek hidden layerlı MLP'nin daha iyi sonuç verdiği görülmüştür.

Kullanılan bir diğer model ise genellikle fonksiyon yaklaşımlarında kullanılan Generalized Regression Neural Network(GRNN) olmuştur. İki katmandan oluşan GRNN'in ilk katmanında(Radial Basis Layer) 140, ikinci katmanında(Special Linear Layer) 1 nöron kullanılmıştır; ancak buradan elde edilen sonuçlar MLP'ye göre daha az tatmin edicidir.

Son model olarak, Support Vector Regression(SVR) kullanılmıştır. SVR modelinin kerneli tüm parametrelerle denenip en son linear olarak seçilmiş ve bu modelin hem HOMO hem de LUMO değerleri için MLP'den daha iyi sonuç verdiği görülmüştür.

Son aşamada MLP ve SVR için Feature Selection ve Lasso Regularization kullanılmış ve alınan sonuçlar analiz edilmiştir. Feature Selection'da, 115 olan özellik sayısını HOMO için 42'ye, LUMO içinse 30'a düşürmek iyi sonuç vermiştir. Ancak Lasso Regularization'ın bu dataset için uygun olmadığı tespit edilmiştir, bunun nedeni datasetin zaten sparse bir yapıya sahip olması şeklinde yorumlanmıştır.

Sonuçlar Mean Squared Error(MSE) kullanılarak karşılaştırılmış ve aşağıdaki grafik oluşturulmuştur:

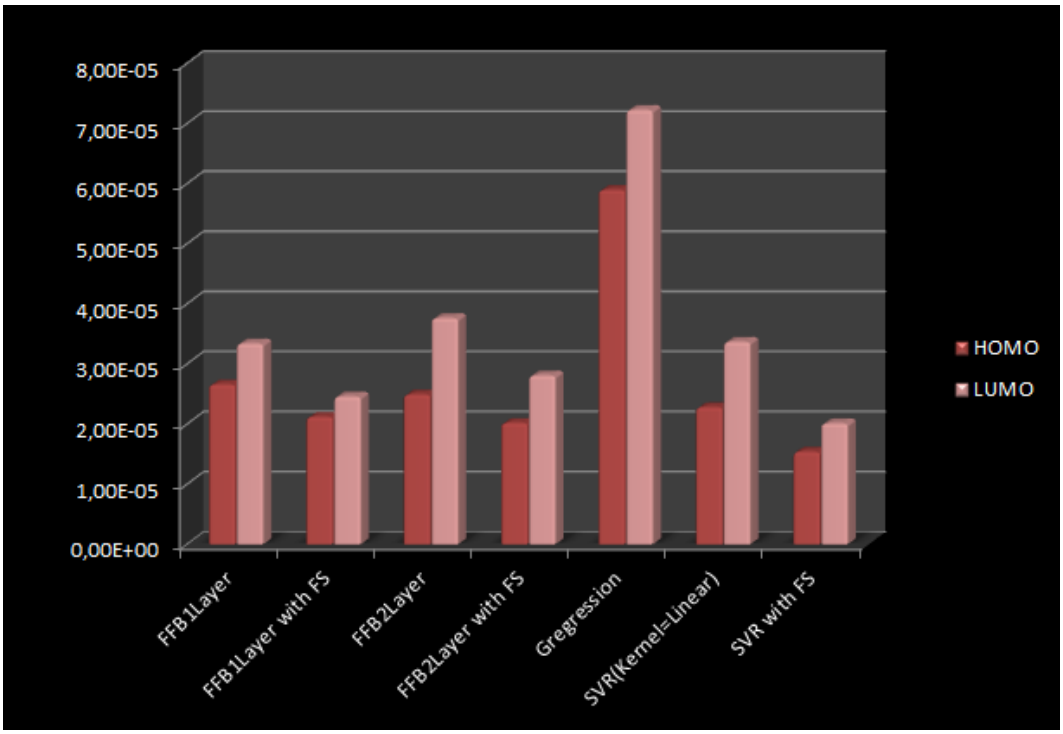
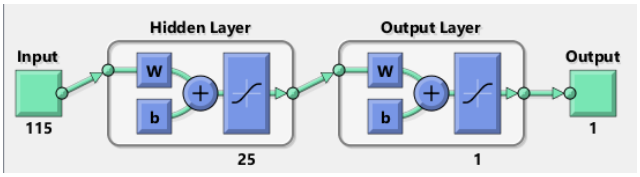
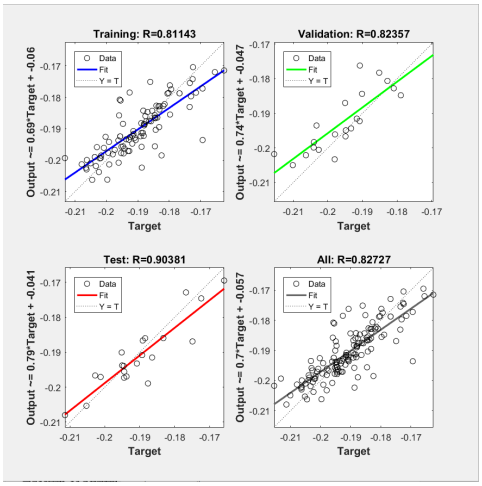


Figure 5: Modellerin MSE değerleri.

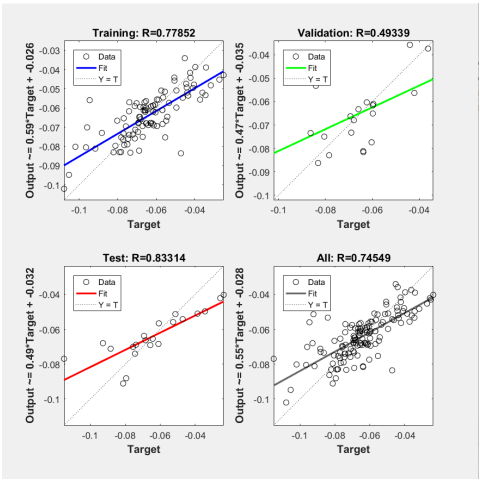
Ayrıca, MATLAB'da MLP(1 ve 2 hidden layer'lı) için aşağıdaki grafikler elde edilmiştir.



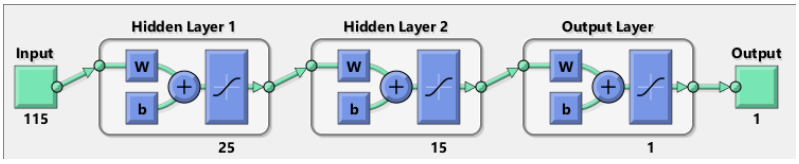
(a) Tek hidden layer'lı MLP modeli



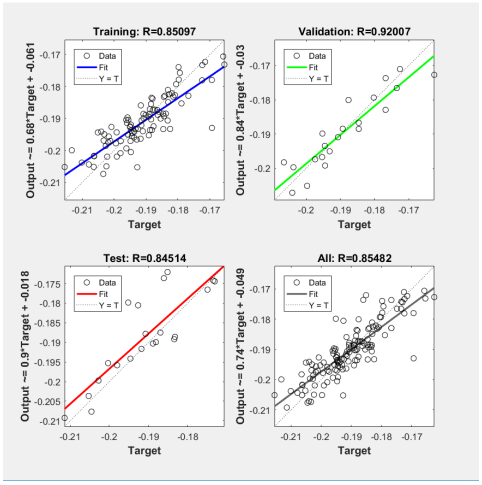
(b) HOMO değerleri tahmini



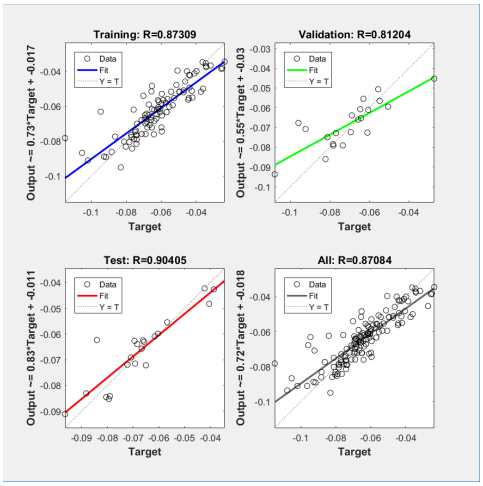
(c) LUMO değerleri tahmini



(a) İki hidden layer'lı MLP modeli



(b) HOMO değerleri tahmini



(c) LUMO değerleri tahmini

5 Future Directions and Ideas

Modeller test edilip incelendiğinde, genel anlamda iyi sonuç vermelerine rağmen datasetinin küçüklüğü daha iyi sonuçlar almada bir engel teşkil etmiştir. Ayrıca özelliklerin moleküller için daha spesifik olan yönleri ele alınabilir, çünkü datasette tekrar eden çok fazla özellik bulunmaktadır. İleriki çalışmalarda bu sorunlar detaylandırılıp dataset genişletilebilir. Bununla birlikte moleküllerin daha farklı özellikleri de target değer olarak kullanılabilir.

6 Conclusion

Projemizden elde ettiğimiz sonuçlar incelendiğinde, bir YSA modeli oluşturulurken ve eğitilirken küçük datasete sahip olmanın olumsuz yanları gözlemlenmiştir. Yeterli eğitim ve test verisi olmadığı zaman güzel sonuçlar almak zorlaşmaktadır. Buna rağmen modellerimizden elde ettiğimiz sonuçlar tatmin edicidir. Model karşılaştırması için MSE fonksiyonu kullanılmış ve modeller birden çok kez denenerek en iyi sonuç elde edilmeye çalışılmıştır. Raporumuzda belirtilen sonuçlar pek çok denemeden aldığımız en iyi sonuçlardır.

Support Vector Regression'ın kullandığımız modeller içinde en iyi sonucu verdiği gözlemlenmiştir. Ayrıca Feature Selection yapmanın faydaları da sonuçlarımıza yansımıştır(Figure 5).

Organik Yarı İletkenlerin HOMO ve LUMO enerji seviyelerinin hesaplanması zaman alıcı ve maliyetli olduğundan projemiz bu konuda yapılan çalışmalarda yardımcı kaynak olarak kullanılabilir.

References

[1] S. Lawrence , C. L. Giles , Proc. IEEE-INNS- ENNS Int. Joint Conf. Neural Networks IJCNN 2000 , IEEE , Piscataway, NJ 2000 , Vol. 1, pp. 114 - 119.

[2] L. Prechelt , Neural Networks 1998 , 11 , 761 .

[3] Pyzer-Knapp, Edward O, Li, Kewei, and Aspuru-Guzik, Alan. Learning from the Harvard clean energy project: The use of neural networks to accelerate materials discovery. Advanced Functional Materials, 25(41):6495–6502, 2015.