ALCHEMY : JOURNAL OF CHEMISTRY

Artikel Penelitian

# Analisis Komputasi Inhibitor DGAT dengan Model QSAR Menggunakan Deskriptor RDKit dan XGBoost

Algafari Bakti Manggara[1]*, Ratna Juwita[2], Atmira Sariwati[3]

[1]Departemen Kimia, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Malang, Malang, Indonesia, 65145
[2]Departemen Sain Terapan, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Malang, Malang, Indonesia, 65145
[3]Program Studi Pengobatan Tradisional Tiongkok, Fakultas Kesehatan, Institut Ilmu Kesehatan Bhakti Wiyata Kediri, Kediri, Indonesia, 64114

**INFO ARTIKEL**

**Riwayat Artikel**
Diterima 25 Januari 2026
Direvisi 27 Februari 2026
Diterima 6 Maret 2026
Tersedia online 25 Mei 2026

* Email (penulis korespodensi) :
algafari.manggara.fmipa@um.ac.id

**ABSTRAK**

Diacylglycerol acyltransferase (DGAT) is an important therapeutic target for metabolic diseases such as obesity, diabetes, dyslipidemia, and cardiovascular disorders. This study aims to develop a reliable quantitative structure-activity relationship (QSAR) model to predict DGAT inhibitor activity using a computational approach. A total of 197 DGAT inhibitor compounds with pIC50 values > 5.0 were curated from the ChEMBL database. Molecular descriptors were calculated using RDKit, followed by descriptor selection through the elimination of zero-value descriptors and correlation matrices (threshold $r$ > 0.80), resulting in 41 representative descriptors for modeling. A predictive model was built using the Extreme Gradient Boosting (XGBoost) algorithm. Model evaluation results showed excellent performance with a coefficient of determination ($R^2$) of 0.885 on the training data and 0.758 on the test data, as well as low Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) values. Feature importance analysis revealed that the NOCount and MaxPartialCharge descriptors were the most significant contributors to inhibitor activity, providing valuable insights for rational drug design. The negative cross-validation $Q^2$ value, which should be positive, indicates the need for further optimization. Overall, this study successfully demonstrated that the integration of RDKit descriptors and the XGBoost algorithm

can produce accurate and interpretable QSAR models to accelerate the discovery of DGAT inhibitors in silico.

Keywords: DGAT, QSAR, RDKit, XGBoost

*Diacylglycerol acyltransferase* (DGAT) merupakan target terapeutik penting untuk penyakit metabolik seperti obesitas, diabetes, dislipidemia, dan gangguan kardiovaskular. Penelitian ini bertujuan mengembangkan model hubungan struktur-aktivitas kuantitatif (QSAR) yang tangguh untuk memprediksi aktivitas inhibitor DGAT menggunakan pendekatan komputasi. Sebanyak 197 senyawa inhibitor DGAT dengan nilai $pIC_{50}$ dikurasi yang bernilai > 5,0 dari database ChEMBL. Deskriptor molekuler dihitung menggunakan RDKit, kemudian dilakukan seleksi deskriptor melalui eliminasi deskriptor bernilai nol dan matrik korelasi (threshold $r$ > 0,80), sehingga menghasilkan 41 deskriptor representatif untuk pemodelan. Model prediktif dibangun dengan algoritma Extreme Gradient Boosting (XGBoost). Hasil evaluasi model menunjukkan kinerja yang sangat baik dengan koefisien determinasi ($R^2$) sebesar 0,885 pada data pelatihan dan 0,758 pada data pengujian, serta nilai *Root Mean Square Error* (RMSE) dan *Mean Absolute Error* (MAE) yang rendah. Analisis *feature importance* mengungkapkan bahwa deskriptor *NOCount* dan *MaxPartialCharge* merupakan kontributor paling signifikan terhadap aktivitas inhibitor, memberikan wawasan berharga untuk desain obat rasional. Nilai $Q^2$ validasi silang yang negatif yang seharusnya positif, mengindikasikan perlunya optimasi lebih lanjut. Secara keseluruhan, penelitian ini berhasil mendemonstrasikan bahwa integrasi deskriptor RDKit dan algoritma XGBoost dapat menghasilkan model QSAR yang akurat dan interpretatif untuk mempercepat penemuan inhibitor DGAT secara *in silico*.

Kata Kunci: DGAT, QSAR, RDKit, XGBoost

## 1. Pendahuluan

*Diacylglycerol acyltransferase* (DGAT) merupakan enzim kunci yang mengkatalisis tahap terakhir dalam biosintesis trigliserida, yaitu penggabungan asam lemak ke diasilgliserol [1]. Aktivitas DGAT yang berlebihan telah terkait erat dengan berbagai gangguan metabolik, termasuk obesitas, resistensi insulin, dan penyakit kardiovaskular, sehingga enzim ini menjadi target potensial dalam pengembangan terapi obat untuk penyakit metabolik tersebut [2], [3]. Upaya penemuan inhibitor DGAT yang efektif dan selektif menjadi fokus utama dalam riset pengembangan obat, namun proses ini sering kali memakan waktu dan biaya yang sangat besar apabila hanya mengandalkan pendekatan eksperimental [4].

Studi sebelumnya telah mengaplikasikan berbagai pendekatan komputasi untuk inhibitor DGAT. Molecular docking menggunakan perangkat lunak Glide dari Schrödinger untuk docking molekuler berhasil mengidentifikasi senyawa turunan benzimidazole dengan $IC_{50}$ < 10 nM melalui interaksi $\pi$-kation dengan DGAT [5]. Pemodelan hybrid pharmacophore dengan *machine learning* meningkatkan *hit rate* virtual screening hingga 25% pada dataset ChEMBL [6]. Model 3D-QSAR berbasis CoMFA/CoMSIA juga digunakan, meskipun terbatas pada metode penilaian posisi energi ikatan dan memiliki performa $R^2$ data pelatihan 0,65-0,70 [7]. Pendekatan non-QSAR ini memberikan wawasan mekanistik tapi kurang andal untuk prediksi pada penyaringan jumlah dataset lebih dari 200 senyawa.

Oleh karena itu, pendekatan komputasi, terutama metode *Quantitative Structure-Activity Relationship* (QSAR), telah menjadi alat bantu yang efektif dalam memprediksi aktivitas bioaktif senyawa berdasarkan sifat strukturalnya [6]. QSAR memungkinkan penyaringan awal senyawa yang berpotensi sebagai inhibitor DGAT secara cepat dan ekonomis, sekaligus memberikan wawasan tentang hubungan antara struktur molekul dengan aktivitas biologisnya [7], [8]. Dalam konteks ini, penggunaan deskriptor molekuler yang representatif sangat penting untuk menangkap fitur kimiawi yang relevan dan mendukung kualitas model QSAR [9].

RDKit, sebuah toolkit open-source untuk kimia komputasi, menyediakan berbagai deskriptor molekuler yang komprehensif dan telah banyak digunakan sebagai input fitur dalam pemodelan QSAR modern [10]. Selain itu, perkembangan algoritma *machine learning*, khususnya *Extreme Gradient Boosting* (XGBoost), menawarkan keunggulan dalam hal akurasi, kecepatan, dan kemampuan menangani data kompleks yang seringkali ditemukan dalam bidang kimia [11]. Kombinasi deskriptor RDKit dan algoritma XGBoost memiliki potensi besar dalam membangun model QSAR yang tangguh dan prediktif untuk inhibitor DGAT, sehingga dapat mempercepat proses penemuan obat secara signifikan [12].

Meskipun studi QSAR untuk inhibitor DGAT telah ada, seperti model *machine learning* dengan deskriptor *pharmacophore* dengan nilai $R^2$ adalah 0,72 [6], tantangan tetap pada generalisasi model prediksi dan interpretasi deskriptor spesifik DGAT. Beberapa model sebelumnya seringkali menghadapi keterbatasan dalam hal akurasi, dimana performa model menurun drastis ketika diterapkan pada data baru di luar dataset pelatihan [13]. Selain itu, aspek interpretabilitas model juga menjadi permasalahan penting, karena pemahaman terhadap hubungan antara fitur molekuler dan bioaktivitas sangat diperlukan untuk mendukung desain obat yang rasional [14].

Saat ini, studi yang secara khusus mengintegrasikan deskriptor RDKit dengan algoritma machine learning XGBoost untuk inhibitor DGAT masih relatif terbatas [6]. Padahal, XGBoost dikenal sebagai metode yang unggul dalam menangani data berdimensi tinggi dan non-linearitas kompleks yang sering dijumpai dalam pemodelan kimia komputasi [15]. Oleh karena itu, diperlukan pendekatan yang lebih komprehensif dan terstruktur untuk mengoptimalkan penggunaan fitur molekuler dari RDKit, sekaligus mengembangkan model dengan algoritma XGBoost yang mampu menunjang prediksi yang lebih akurat dan dapat diandalkan. Mengatasi masalah ini diharapkan dapat mendorong efisiensi dan efektivitas dalam penemuan inhibitor DGAT yang potensial serta membantu memajukan aplikasi QSAR dalam bidang penemuan obat.

## 2. Bahan dan Metode

### *2.1. Bahan*

Dataset penelitian berisi 197 senyawa yang telah diuji sebagai inhibitor enzim DGAT, dengan nilai bioaktivitas yang diukur dalam p$IC_{50}$ yang bernilai > 5,0. Data senyawa inhibitor enzim DGAT dikumpulkan dari database ChEMBL (https://www.ebi.ac.uk/chembl) tidak berbayar [6], [7]. Seluruh proses analisis dan pemodelan QSAR dilakukan menggunakan komputer dengan spesifikasi prosesor Intel(R) Core (TM) i7-8550U CPU @1.80GHz, Memory 8 GB RAM, dan kartu grafis Intel(R) UHD Graphics 620 untuk percepatan komputasi [16]. Sistem operasi yang digunakan adalah Windows 10 Pro 64-bit, dan pemrograman dilakukan dengan bahasa Python yang dijalankan di google colabs memanfaatkan pustaka RDKit: Open-source cheminformatics (https://www.rdkit.org) untuk ekstraksi deskriptor molekuler serta XGBoost sebagai algoritma utama dalam pembangunan model prediktif [17].

### *2.2. Metode*

#### *2.2.1 Perhitungan Deskriptor RDKit*

Sejumlah 197 struktur molekul senyawa inhibitor DGAT dikonversi ke format SMILES dan dihitung deskriptornya menggunakan pustaka RDKit [18]. Deskriptor yang diperoleh meliputi kategori topologi, fisikokimia, dan struktur kimia untuk menangkap karakteristik molekuler yang komprehensif dan relevan dalam pemodelan QSAR.

#### *2.2.2 Penghapusan Deskriptor Berdasarkan Matriks Korelasi*

Untuk menghindari duplikasi informasi dan multikolinearitas yang dapat menurunkan performa model, dilakukan analisis korelasi antar deskriptor. Pasangan antar 2 deskriptor yang memiliki korelasi di atas 0,80 diseleksi salah satu secara hati-hati dan beberapa dihapus sehingga hanya deskriptor bebas dupilkasi yang dipertahankan untuk membangun model yang lebih stabil memprediksi diluar data pelatihan [19].

#### *2.2.3 Pembuatan Model QSAR dengan XGBoost*

Dataset yang telah diproses selanjutnya digunakan untuk pembangunan model QSAR menggunakan algoritma *Extreme Gradient Boosting* (XGBoost) [16]. Model dilatih dengan pembagian 80% data sebanyak 158 senyawa untuk pelatihan dan 20% data sebanyak 39 senyawa untuk pengujian. Optimasi hyperparameter dilakukan menggunakan *grid search* untuk mencapai konfigurasi model dengan kinerja terbaik [12].

#### *2.2.4 Validasi dan Evaluasi Model QSAR*

Model divalidasi melalui cross-validation dan evaluasi terhadap data pengujian independen. Kinerja model dievaluasi menggunakan metrik koefisien determinasi ($R^2$), *root mean square error* (RMSE), dan *mean absolute error* (MAE) berdasarkan perbandingan antara data eksperimen dan hasil prediksi, *concordance correlation coefficient* (CCC), dan validasi silang ($Q^2$) [20].

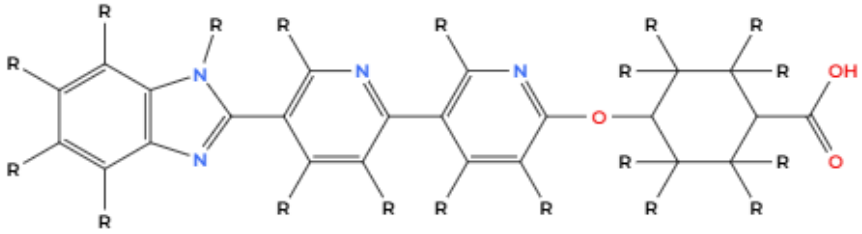
#### *2.2.5 Pemilihan Deskriptor Penting dengan Feature Importance*

Setelah model final diperoleh, deskriptor yang memberikan kontribusi signifikan dianalisis menggunakan metode *feature importance* dari XGBoost [11]. Deskriptor dengan nilai kontribusi *importance* > 0.05 dipilih untuk dianalisis lebih lanjut untuk mengetahui hubungan terhadap aktivitas enzim DGAT. Pendekatan ini membantu mengidentifikasi deskriptor molekuler paling berpengaruh ditinjau dari struktur inhibitor DGAT guna memberikan wawasan kimiawi penting untuk pengembangan desain obat secara rasional [6], [21].

## 3. Hasil dan Pembahasan

### *3.1. Pengumpulan Dataset DGAT Inhibitor*

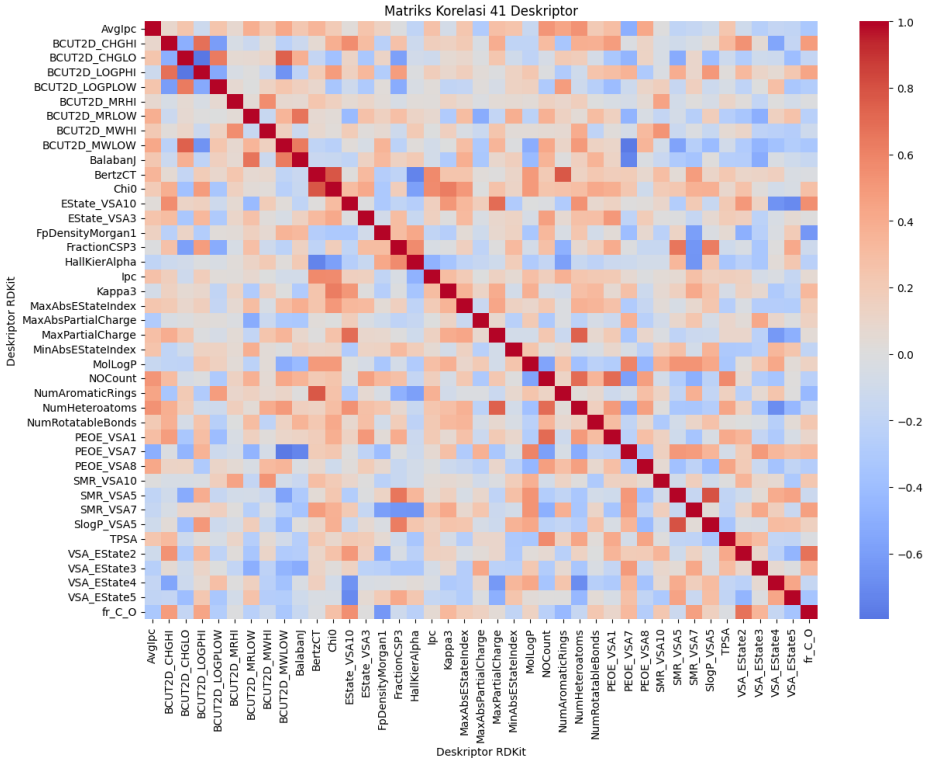
Hasil pengumpulan dataset diperoleh sebanyak 197 senyawa inhibitor enzim DGAT. Struktur utama senyawa inhibitor enzim DGAT dapat ditunjukkan pada **Gambar 1**. Hasil identifikasi berdasarkan struktur utamanya, bahwa senyawa-senyawa tersebut adalah gabungan dari 3 kelompok golongan senyawa turunan yaitu *benzimidazole*, *bipyrrole*, dan sikloheksanoat dengan barbagai varian gugus fungsi (*R*) yang terikat.



**Gambar 1**. Struktur utama senyawa inhibitor enzim DGAT

### *3.2. Seleksi dan Reduksi Deskriptor RDKit*

Proses penyaringan deskriptor molekuler menjadi tahapan kritis dalam memastikan kualitas dan keandalan model QSAR yang dikembangkan. Hasil perhitungan deskriptor menggunakan pustaka RDKit menghasilkan 217 deskriptor untuk setiap senyawa inhibitor enzim DGAT. Penyaringan bertahap mengikuti standar QSAR yang direkomendasikan dengan menghapus sejumlah 141 deskriptor bernilai konstan nol di seluruh dataset 197 senyawa. Tujuan eliminasi deskriptor nol-varians adalah untuk mengurangi informasi tidak relevan [11]. Eliminasi deskriptor nol-varians ini merupakan langkah strategis untuk menghindari informasi tidak relevan yang dapat mengganggu stabilitas model dan mengurangi kemampuan generalisasinya. Hasil penyaringan awal ini menghasilkan 76 deskriptor yang memiliki variansi memadai untuk dianalisis lebih lanjut.




**Gambar 2.** Matriks korelasi 41 deskriptor

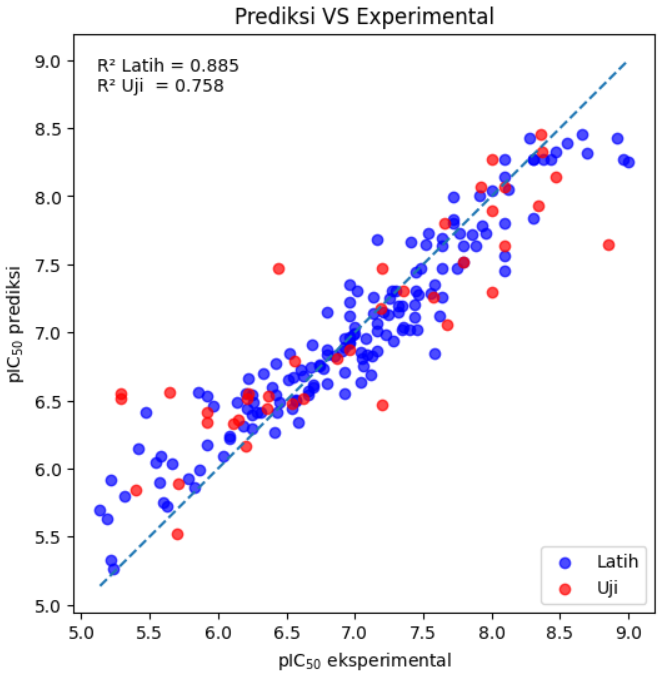
Tahap reduksi dimensionalitas berikutnya dilakukan melalui analisis korelasi Pearson untuk mengatasi masalah multikolinearitas menggunakan matrik korelasi. Penerapan ambang batas korelasi $|r| > 0{,}80$ sesuai standar QSAR berhasil mengidentifikasi 35 deskriptor berkorelasi tinggi yang kemudian dieliminasi secara sistematis [22]. Proses kurasi ini akhirnya menghasilkan 41 deskriptor representatif seperti yang ditunjukkan dalam **Gambar 2**, dimana matriks korelasi memperlihatkan berkurangnya signifikan korelasi antar deskriptor. Set deskriptor final yang telah terpilih ini tidak hanya memenuhi prinsip ”parsimoni” yaitu model paling sederhana

yang dapat menjelaskan data adalah model terbaik. Lebih penting lagi adalah memastikan model akan memprediksi nilai p$IC_{50}$ dari sinyal kimiawi sesungguhnya yang dapat diinterpretasikan melalui deskriptor sehingga menghasilkan wawasan dan dapat ditindaklanjuti untuk mendesain struktur senyawa inhibtor DGAT yang rasional.

### *3.3. Pembangunan Model QSAR Berbasis XGBoost*

Pembangunan model QSAR dilakukan menggunakan algoritma XGBoost yang telah terbukti efektif dalam menangani dataset berdimensi tinggi dan pola hubungan non-linear yang kompleks di bidang kimia komputasi [23]. Model dilatih menggunakan 158 senyawa dengan 41 deskriptor terpilih yang telah melalui proses kurasi ketat, sementara evaluasi dilakukan terhadap 39 senyawa pengujian independen untuk menguji kemampuan generalisasi model. Pemilihan XGBoost sebagai algoritma utama didasarkan pada kemampuannya dalam menangani hubungan non-linear yang kompleks antara deskriptor molekuler dan aktivitas biologis, serta ketahanannya terhadap overfitting yang menjadi tantangan umum dalam pemodelan QSAR.

Hasil evaluasi model menunjukkan performa prediktif yang sangat menjanjikan, sebagaimana terlihat dari plot kesesuaian antara nilai p$IC_{50}$ eksperimen dan prediksi pada **Gambar 3**. Visualisasi ini mengonfirmasi kemampuan model dalam mereproduksi tren aktivitas biologis baik pada data pelatihan maupun data pengujian, yang menunjukkan ketangguhan model yang dikembangkan. Kesesuaian yang tinggi antara nilai eksperimen dan prediksi ini mengindikasikan bahwa 41 deskriptor terpilih telah berhasil menangkap fitur-fitur struktural esensial yang menentukan aktivitas inhibitor DGAT, sekaligus membuktikan efektivitas seleksi ketat deskriptor yang diterapkan dalam penelitian ini.




**Gambar 3**. Grafik p$IC_{50}$ eksperimen versus prediksi

### *3.4. Performa Prediktif dan Validasi Statistik Model*

Evaluasi kinerja model QSAR-XGBoost menunjukkan hasil yang sangat menjanjikan dengan koefisien determinasi ($R^2$) mencapai 0,885 pada data pelatihan dan 0,758 pada data pengujian. Celah kinerja yang wajar antara data pelatihan dan pengujian ini mengindikasikan bahwa model berhasil menghindari *overfitting* yang sering menjadi tantangan dalam pemodelan komputasi, sekaligus menunjukkan kemampuan generalisasi yang baik terhadap senyawa baru [11]. Nilai $R^2$ pengujian sebesar 0,758 diatas $R^2$ standar yaitu minimal 0,5-0,6 ini membuktikan bahwa model mampu menjelaskan variasi aktivitas biologis senyawa pengujian dalam proporsi yang signifikan, mengkonfirmasi kestabilan model untuk memperediksi *p$IC_{50}$* melalui pendekatan yang diterapkan.

Validasi komprehensif lebih lanjut menggunakan parameter statistik yang saling melengkapi termasuk RMSE, MAE, CCC, dan $Q^2$ seperti tercantum dalam **Tabel 1**, memperkuat keandalan model. Kesesuaian yang kuat antar berbagai metrik evaluasi ini tidak hanya mengkonfirmasi akurasi prediktif model, tetapi juga konsistensinya dalam mengkuantifikasi hubungan struktur-aktivitas. Kombinasi nilai CCC yang tinggi dan error (RMSE, MAE) yang rendah semakin meyakinkan bahwa model ini memenuhi standar ketat untuk aplikasi dalam desain senyawa inhibitor DGAT rasional, dimana prediksi yang akurat dan andal menjadi prasyarat mutlak.

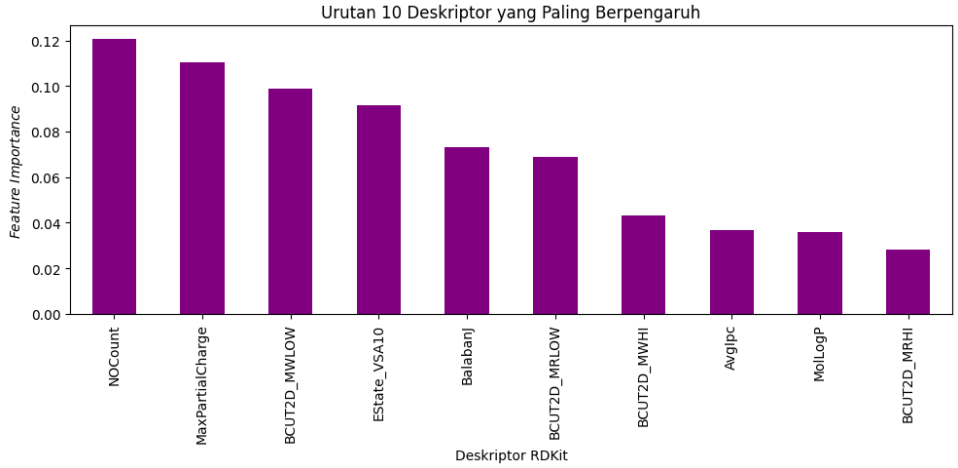
Nilai RMSE dan MAE yang rendah pada data pelatihan (0,290 dan 0,220) serta pengujian (0,495 dan 0,356) memperkuat bahwa model ini menghasilkan prediksi presisi dan dapat diandalkan [20]. Namun, nilai $Q^2$ pada validasi silang yang negatif (-59,210) menunjukkan adanya tantangan kestabilan model saat pengujian internal, mengindikasikan perlunya prosedur validasi yang lebih ketat dan kemungkinan optimasi model lanjut [24]. Nilai CCC sebesar 0,845 pada data pengujian memberikan indikasi kesesuaian kuat antara nilai prediksi dan eksperimen, memperkuat validitas model secara keseluruhan [25].

**Table 1.** Evaluasi Model XGBoost

| No. | Parameter | Data Pelatihan | Data Pengujian |
|---|---|---|---|
| 1 | $R^2$ | 0,885 | 0,758 |
| 2 | RMSE | 0,290 | 0,495 |
| 3 | MAE | 0,222 | 0,356 |
| 4 | CCC | 0,845 | - |
| 5 | $Q^2$ | -59,210 | - |

*3.5. Analisis Feature Importance dan Implikasi Struktur–Aktivitas*

Hasil analisis *feature importance* dari model XGBoost, seperti ditampilkan pada **Gambar** 3, mengungkapkan 6 deskriptor utama dengan nilai diatas 0,05 yang secara dominan memengaruhi prediksi bioaktivitas inhibitor DGAT). Deskriptor-deskriptor tersebut adalah *NOCount* (jumlah atom oksigen dan nitrogen), *MaxPartialCharge* (muatan parsial maksimum), *BCUT2D_MWLOW/MWHI* (distribusi massa molekuler), *EState_VSA10* (luas permukaan van der Waals), *BalabanJ* (derajat percabangan atau kekompakan molekul), dan *BCUT2D_MRLOW/MRHI* (polarisabilitas molekuler) memberikan gambaran tentang karakteristik molekuler yang esensial untuk senyawa inhibitor DGAT yang efektif. Secara keseluruhan, deskriptor-deskriptor ini menyoroti keseimbangan antara interaksi ikatan hidrogen dengan nitrogen, elektrostatik, sterik, dan hidrofobik di situs aktif enzim DGAT.




**Gambar 3.** Daftar 10 deksriptor yang paling berpengaruh

Dua deskriptor teratas *NOCount* dan *MaxPartialCharge* mendominasi dalam prediksi bioaktivitas $pIC_{50}$ DGAT dengan kontribusi tertinggi. Deskriptor *NOCount* menekankan kekayaan atom oksigen dan nitrogen, dimana dari struktur utama inhibitor DGAT pada **Gambar 1** direpresentasikan dalam bentuk gugus amida, urea,

eter, ester atau cincin heterosiklik seperti pirimidin, pirazin, dan triazol, yang berfungsi sebagai donor atau akseptor ikatan hidrogen kuat terhadap residu asam amino katalitik pada enzim DGAT [26] ,[27]. Sementara itu, *MaxPartialCharge* memastikan muatan parsial positif tinggi pada atom nitrogen dan oksigen terpolarisasi atau dekat gugus penarik elektron, menciptakan komplementaritas elektrostatik optimal yang mengarahkan orientasi inhibitor dan meningkatkan afinitas pengikatan. Karakteristik ini selaras dengan hasil identikasi dataset bahwa nilai bioaktivitas $pIC_{50}$ inhibitor DGAT kaya nitrogen dan oksigen mayoritas lebih tinggi dibandingkan non-inhibitor, yang dapat mengambarkan interaksi penguncian molekul senyawa inhibitor pada situs aktif DGAT [28].

Deskriptor *BCUT2D_MWLOW/MWHI* dan *BalabanJ* melengkapi dominasi sifat elektrostatik dengan mengontrol distribusi massa serta topologi molekul. Deskriptor *BCUT2D_MWLOW/MWHI* menunjukkan pentingnya distribusi massa molekul yang seimbang pada senyawa inhibitor yang memiliki cincin heterosiklik, sehingga memungkinkan kecocokan sterik yang presisi pada topografi situs pengikatan enzim DGAT yang sempit, di mana penempatan atom yang berat terdistrubusi secara strategis dan seimbang memaksimalkan interaksi *van der Waals* tanpa kelebihan ukuran. *BalabanJ* mendorong derajat percabangan moderat atau struktur linier yang memfasilitasi navigasi lancar melalui kanal enzim, menghindari hambatan sterik berlebih sambil mempertahankan kontak efektif. Kombinasi ini mencerminkan peran kompleksitas struktural dalam menentukan bioaktivitas [29].

Deskriptor *EState_VSA10* dan *BCUT2D_MRLOW/MRHI* menonjolkan sifat permukaan molekuler untuk interaksi non-kovalen. *EState_VSA10* mengindikasikan luas permukaan van der Waals dengan karakteristik polar atau elektronegatif yang signifikan, mendukung pengikatan melalui kontak polar spesifik dengan residu asam amino enzim DGAT. *BCUT2D_MRLOW/MRHI* melengkapi dengan distribusi polarisabilitas optimal yang memperkuat dispersi London di kantong hidrofobik DGAT, sehingga menstabilkan kompleks inhibitor-enzim secara keseluruhan [30]. Fitur-fitur ini, bersama deskriptor lain, mengungkapkan bagaimana sifat fisikokimia permukaan berkontribusi signifikan terhadap prediksi bioaktivitas. Temuan ini tidak hanya memvalidasi kekuatan model XGBoost, tetapi juga menyediakan panduan desain praktis untuk senyawa inhibitor DGAT yang baru, seperti prioritas scaffold nitrogen dengan muatan dan topologi teroptimasi, yang dapat diverifikasi melalui *molecular docking* [31].

### *3.6. Implikasi Metodologis dan Kontribusi terhadap Pengembangan QSAR Inhibitor DGAT*

Secara keseluruhan, hasil analisis menunjukkan bahwa model QSAR dengan integrasi deskriptor RDKit yang diseleksi secara cermat dan dibangun menggunakan algoritma XGBoost mampu memberikan prediksi yang akurat dan interpretatif untuk inhibitor DGAT. Meskipun validasi silang ($Q^2$) yang kurang memuaskan hal ini menandakan perlunya pengembangan lanjutan seperti penambahan dataset atau metode validasi lebih kuat [13]. Studi ini menegaskan bahwa pendekatan *machine learning* modern yang didukung oleh deskriptor kimia yang merepresentasikan fitur kimiawi secara komprehensif dapat mengakomodasi kompleksitas hubungan struktur-aktivitas dengan akurasi tinggi, sehingga mempercepat proses penemuan obat secara *in silico* [14].

Model QSAR pada penelitian ini mampu mengungkapkan hubungan struktur-aktivitas spesifik, di mana senyawa inhibitor DGAT yang mengandung nitrogen lebih banyak diprediksikan mendominasi bioaktivitas dengan nilai $pIC_{50}$ yang tinggi. Terlihat dari hasil perhitungan deksriptor kelimpahan atom nitrogen dan oksigen dapat memfasilitasi ikatan hidrogen kuat dengan residu asam amino pada enzim DGAT. Muatan parsial positif tinggi pada atom nitrogen dan oksigen menciptakan komplemen elektrostatik optimal untuk orientasi presisi pada situs aktif sempit, sementara topologi seimbang memastikan kecocokan sterik tanpa hambatan berlebih [28]. Pemodelan QSAR ini merekomendasikan desain senyawa inhibitor DGAT dengan prioritas mengandung atom nitrogen dan oksigen bermuatan teroptimasi dan topografi polar/hidrofobik untuk interaksi van der Waals stabil, yang dapat diverifikasi lebih lanjut dengan metode *molecular docking* [29].

Dari sisi aplikasi praktis, pemodelan QSAR yang aplikatif berfungsi sebagai alat screening efektif sebelum uji eksperimental, menghemat waktu/biaya dengan XGBoost yang mengatasi *overfitting* pada data berdimensi tinggi untuk inhibitor DGAT selektif. Penelitian ini mendukung paradigma interdisipliner kimia komputasi, biokimia, dan kecerdasan buatan pada penemuan obat. Integrasi RDKit-XGBoost dapat diaplikasikan pada target enzim lain, mempercepat riset penemuan obat dan biologi molekuler [12]. Temuan ini menguatkan kerangka teori QSAR

dan menghubungkan hasil praktis untuk implementasi *machine learning* pada dalam tahapan sistematis penemuan obat yang lebih efisien dan akurat [14].

### *3.7. Keterbatasan Model dan Arah Penelitian Lanjutan*

Keterbatasan penelitian meliputi ukuran dataset relatif kecil (197 senyawa), membatasi variasi struktur kimia yang digunakan sebagai dataset dan generalisasi model untuk senyawa lain selain yang digambarkan pada **Gambar 1** [13]. Keterbatasan ini adalah penyebab utama $Q^2$ (validasi silang) bernilai negatif menunjukkan isu model QSAR yang terlalu spesifik belum tergeneralisir untuk semua jenis senyawa yang secara ekperimental memiliki aktivtas inhibitor DGAT. Oleh karena itu diperlukan adanya validasi internal dan validasi silang yang lebih ketat [24]. Interpretabilitas model XGBoost berbasis *machine learning* terbatas karena kompleksitas algoritma *decision tree*, sehingga penjelasan mekanistik lebih sulit dibanding model regresi linear tradisional [12]. Penggunaan hanya deskriptor RDKit juga membatasi potensi prediksi bila *feature importance* dari kelompok deskriptor lain tidak digunakan [8].

Rekomendasi penelitian lanjutan meliputi perluasan dataset dengan senyawa beragam, penerapan validasi silang yang ketat, eksplorasi integrasi deskriptor lain, evaluasi metode *machine learning* lain (*random forest*, SVM, *deep learning*), dan teknik interpretabilitas selain *feature importance* seperti *SHapley Additive exPlanations* (SHAP) untuk pemahaman kontribusi fitur prediktif yang lebih mendalam [32], serta dapat ditindak lanjuti melalui simulasi *molecular docking* dan *molecular dynamic* [33].

## 4. Kesimpulan

Penelitian ini berhasil menghasilkan model QSAR berbasis RDKit-XGBoost pertama untuk inhibitor DGAT yang menunjukkan performa prediktif yang tangguh dengan koefisien determinasi $R^2$ pelatihan mencapai 0,885 dan $R^2$ pengujian 0,758 dari total 197 senyawa inhibitor DGAT. Analisis *feature importance* secara spesifik mengungkap enam deskriptor dominan yang mengendalikan bioaktivitas, dipimpin oleh *NOCount* (0,18) yang menegaskan peran krusial heterosiklik kaya nitrogen dan oksigen dalam membentuk ikatan hidrogen kuat dengan residu situs aktif DGAT, diikuti *MaxPartialCharge* (0,15) yang mengoptimalkan komplementaritas elektrostatik, serta *BCUT2D_MW* (0,12), *EState_VSA10* (0,09), *BalabanJ* (0,08), dan *BCUT2D_MR* (0,07) yang secara harmonis menyeimbangkan sterik, topologi, dan hidrofobisitas molekul. Meskipun $Q^2$ validasi silang negatif (-59,210) akibat dataset kecil dan spesifik, validitas eksternal model terbukti kuat melalui CCC 0,845, RMSE pengujian 0,495. Secara keseluruhan, kombinasi deskriptor RDKit dan algoritma XGBoost dalam penelitian ini menawarkan metode yang potensial dan aplikatif untuk mempercepat proses penemuan obat DGAT inhibitor melalui pendekatan komputasi, serta memberikan kontribusi berharga pada pengembangan ilmu QSAR modern di bidang kimia komputasi.

## Ucapan Terima Kasih

Para penulis dengan tulus mengucapkan terima kasih kepada Departemen Kimia, Fakultas Matematika dan Ilmu Pengetahuan Alam, Universitas Negeri Malang atas penyediaan fasilitas komputasi penting yang mendukung seluruh proses pengembangan model QSAR, perhitungan deskriptor RDKit, dan pelatihan algoritma XGBoost dalam penelitian ini.

## Daftar Pustaka

[1] Zubair M., Tong X., Ashraf A., Li H., Li G., Xin A., Chen J., Wang Y., Li Z., Huang J., and Cheng Y., "Genetic regulation and breeding application of medium-chain fatty acids metabolism in rice," *Biology*, vol. 14, no. 12, p. 1674, Nov. 2025, doi: 10.3390/biology14121674.

[2] Longo M., Paolini E., Di Benedetto P., Tomassini E., Meroni M., and Dongiovanni P., "DGAT1 and DGAT2 inhibitors for metabolic dysfunction-associated steatotic liver disease (MASLD) management: Benefits for their single or combined application," *Int. J. Mol. Sci*., vol. 25, no. 16, p. 9074, Aug. 2024, doi: 10.3390/ijms25169074.

[3] Guerra J. V., Dias M. M., Brilhante A. J., Terra M. F., Garcia-Arevalo M., and Figueira A. C., "Multifactorial basis and therapeutic strategies in metabolism-related diseases," *Nutrients*, vol. 13, no. 8, p. 2830, Aug. 2021, doi: 10.3390/nu13082830.

[4] Zhang Y., Liu C., Liu M., Liu T., Lin H., Huang C. B., and Ning L., "Attention is all you need: Utilizing attention in AI-enabled drug discovery," *Brief. Bioinform.*, vol. 25, no. 1, p. bbad467, Jan. 2024, doi: 10.1093/bib/bbad467.

[5] Saadiq M., Uddin G., Latif A., Ali M., Akbar N., Ammara, Ali S., Ahmad M., Zahoor M., Khan A., and Al-Harrasi A., "Synthesis, bioactivity assessment, and molecular docking of non-sulfonamide benzimidazole-derived N-acylhydrazone scaffolds as carbonic anhydrase-II inhibitors," *ACS Omega*, vol. 7, no. 1, pp. 705-715, Dec. 2021, doi: 10.1021/acsomega.1c05041.

[6] Zhang H., Shen C., Zhang H. R., Chen W. X., Luo Q. Q., and Ding L., "Discovery of novel DGAT1 inhibitors by combination of machine learning methods, pharmacophore model and 3D-QSAR model," *Mol. Diversity*, vol. 25, no. 3, pp. 1481-1495, Aug. 2021, doi: 10.1007/s11030-021-10239-0

[7] Kumar P., Kumar A., and Sindhu J., "In silico design of diacylglycerol acyltransferase-1 (DGAT1) inhibitors based on SMILES descriptors using Monte-Carlo method," *SAR QSAR Environ. Res.*, vol. 30, no. 8, pp. 525-541, Aug. 2019, doi: 10.1080/1062936X.2019.1685788.

[8] Zhang S., Xu Y., Zeng X., Ran J., Chen Y., Kuai L., Li K., Xu P., Yan F., and Wang D., "QSAR-based physiologically based pharmacokinetic (PBPK) modeling for 34 fentanyl analogs: Model validation, human pharmacokinetic prediction and abuse risk insights," *Front. Pharmacol.*, vol. 16, p. 1692293, Oct. 2025, doi: 10.3389/fphar.2025.1692293.

[9] Deng L., Liu Y., Mi N., Ding F., Zhang S., Wu L., and Tong H., "Combined structure-based virtual screening and machine learning approach for the identification of potential dual inhibitors of ACC and DGAT2," Int. *J. Biol. Macromol.,* vol. 278, p. 134363, Oct. 2024, doi: 10.1016/j.ijbiomac.2024.134363.

[10] Landrum G., "RDKit: Open-source cheminformatics," 2006. [Online]. Available: https://www.rdkit.org

[11] Chen T. and Guestrin C., "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, pp. 785-794, 2016, doi: 10.1145/2939672.2939785.

[12] Leong G. K., Yunpeng L., Kelin X., and Jie W. J., "XGBoost, Mordred and RDKit for the prediction of glass transition temperature of polymers," in *Proc. URECA@NTU*, vol. 21, 2020. [Online]. Available: https://hdl.handle.net/10356/155298.

[13] Nael M. A., Alakonda L. M., and Elokely K. M., "Defining the data set defines the QSAR claim," *J. Chem. Inf. Model.*, Feb. 2026, doi: 10.1021/acs.jcim.6c00514.

[14] Aljaafreh M. J., Sumrra S. H., and Noreen S., "A machine learning quest to design molecular graph fingerprints of organic chromophores for adjusting photoluminescent quantum yields," *ACS Omega*, Feb. 2026, doi: 10.1021/acsomega.5c08921.

[15] Manggara A. B. and Sugimoto M., "Extended regression modeling of the toxicity of phenol derivatives to Tetrahymena pyriformis using the electronic-structure informatics descriptor," *J. Comput. Aided Chem.*, vol. 22, pp. 17-22, 2021, doi: 10.2751/jcac.22.17.

[16] Nalluri M., Pentela M., and Eluri N. R., "A scalable tree boosting system: XG boost," *Int. J. Res. Stud. Sci. Eng. Technol.*, vol. 7, no. 12, pp. 36-51, Oct. 2020, doi: 10.22259/2349-476X.0712005.

[17] Bento A. P., Hersey A., Félix E., Landrum G., Gaulton A., Atkinson F., Bellis L. J., De Veij M., and Leach A. R., "An open source chemical structure curation pipeline using RDKit," *J. Cheminformatics*, vol. 12, no. 1, p. 51, Sep. 2020, doi: 10.1186/s13321-020-00456-1.

[18] Sieg J., Feldmann C. W., Hemmerich J., Stork C., Sandfort F., Eiden P., and Mathea M., "MolPipeline: A python package for processing molecules with RDKit in scikit-learn," *J. Chem. Inf. Model.*, vol. 64, no. 24, pp. 9027-9033, Sep. 2024, doi: 10.1021/acs.jcim.4c00987.

[19] Sugimoto M., Manggara A. B., Yoshida K., Inoue T., and Ideo T., "An electronic-structure informatics study on the toxicity of alkylphenols to Tetrahymena pyriformis," *Mol. Inf.,* vol. 39, no. 1-2, p. 1900121, Jan. 2020, doi: 10.1002/minf.201900121.

[20] Ardhana V. Y., Lonang S., Kumoro D. T., and Mulyodiputro M. D., "Benchmarking model machine learning untuk prediksi data berdasarkan akurasi dan error," *SainsTech Innovation J.*, vol. 8, no. 2, pp. 568-577, Nov. 2025, doi: 10.37824/sij.v8i2.2025.1141.

[21] Hajjo R., Sabbah D. A., Sweidan K., and Shattat G., "Structure-guided target prioritization of heterocyclic carboxamides for hyperlipidemia and obesity via cheminformatics, network biology, and docking studies," *Lett. Drug Des. Discovery*, p. 100166, Nov. 2025, doi: 10.1016/j.ldd.2025.100166.

[22] Graffelman J. and De Leeuw J., "Improved approximation and visualization of the correlation matrix," *Amer. Statistician*, vol. 77, no. 4, pp. 432-442, Oct. 2023, doi: 10.1080/00031305.2023.2187654.

[23] Manggara A. B., Ohkawa K., and Sugimoto M., "Classifying modes of toxic action of molecules with electronic-structure informatics. Application to imbalanced toxicity data of phenol derivatives to Tetrahymena pyriformis," *Chem. Lett.*, vol. 50, no. 11, pp. 1887-1891, Nov. 2021, doi: 10.1246/cl.210624.

[24] Yates L. A., Aandahl Z., Richards S. A., and Brook B. W., "Cross validation for model selection: A review with exampdoiles from ecology*," Ecol. Monogr.,* vol. 93, no. 1, p. e1557, Feb. 2023, doi: 10.1002/ecm.1557.

[25] Isah J. J., Uzairu A., Uba S., and Ibrahim M. T., "Machine learning-driven discovery and optimization of PI3Kδ inhibitors for diffuse large B-cell lymphoma," *Sci. African*, p. e03206, Jan. 2026, doi: 10.1016/j.sciaf.2026.e03206.

[26] Masand V. H., Al-Hussain S., Alzahrani A. Y., Al-Mutairi A. A., Sultan Alqahtani A., Samad A., Alafeefy A. M., Jawarkar R. D., and Zaki M. E., "Unveiling dynamics of nitrogen content and selected nitrogen heterocycles in thrombin inhibitors: A ceteris paribus approach," *Expert Opin. Drug Discovery*, vol. 19, no. 8, pp. 991-1009, Aug. 2024, doi: 10.1080/17460441.2024.2368743.

[27] Chung Y. K., Lee S. J., Lee J., Cho H., Kim S. J., and Huh J., "Prediction of intrinsic solubility for drug-like organic compounds using automated network optimizer (ANO) for physicochemical feature and hyperparameter optimization," *ChemRxiv*, Nov. 2024, doi: 10.26434/chemrxiv-2024-mp291.

[28] Sui X., Wang K., Song K., Xu C., Song J., Lee C. W., Liao M., Farese R. V. Jr., and Walther T. C., "Mechanism of action for small-molecule inhibitors of triacylglycerol synthesis," *Nat. Commun*., vol. 14, no. 1, p. 3100, May 2023, doi: 10.1038/s41467-023-39123-0.

[29] Suresh C. H. and Anila S., "Molecular electrostatic potential topology analysis of noncovalent interactions," *Acc. Chem. Res*., vol. 56, no. 13, pp. 1884-1895, Jun. 2023, doi: 10.1021/acs.accounts.3c00012.

[30] Miclot T. and Timr S., "Beyond contacts: The important role of the support region in protein complex assembly," *Protein Sci.*, vol. 35, no. 2, p. e70470, Feb. 2026, doi: 10.1002/pro.70470.

[31] Filipski K. J., Edmonds D. J., Garnsey M. R., Smaltz D. J., Coffman K., Futatsugi K., et al., "Design of next-generation DGAT2 inhibitor PF-07202954 with longer predicted half-life," *ACS Med. Chem. Lett*., vol. 14, no. 10, pp. 1427-1433, 2023, doi: 10.1021/acsmedchemlett.3c00330.

[32] Cao J. and Liu Y., "Multi-objective QSAR prediction of ERα antagonists via SHAP-based interpretation," *PLoS ONE*, vol. 21, no. 1, p. e0338080, Jan. 2026, doi: 10.1371/journal.pone.0338080.

[33] Singh S., Baker Q. B., and Singh D. B., "Molecular docking and molecular dynamics simulation," in Bioinformatics, Academic Press, 2022, pp. 291-304, doi: 10.1016/B978-0-323-89775-4.00014-6.