الجمهورية الجزائرية الديمقراطية الشعبية

وزارة التعليم العالي والبحث العلمي

جامعة الشهيد حمة لخضر — الوادي

كلية التكنولوجيا

قسم هندسة الطرائق والبتروكيمياء

مذكرة تخرج لنيل شهادة

ماستـــر أكاديمـــي

ميـــدان: العلوم والتكنولوجيا

شعبـــة: هندسة الطرائق

تخصـــص: هندسة كيميائية

مـن اعـداد الطلبـة:

هدى الدام

نسرين لجدل صيد

الموضـــــوع

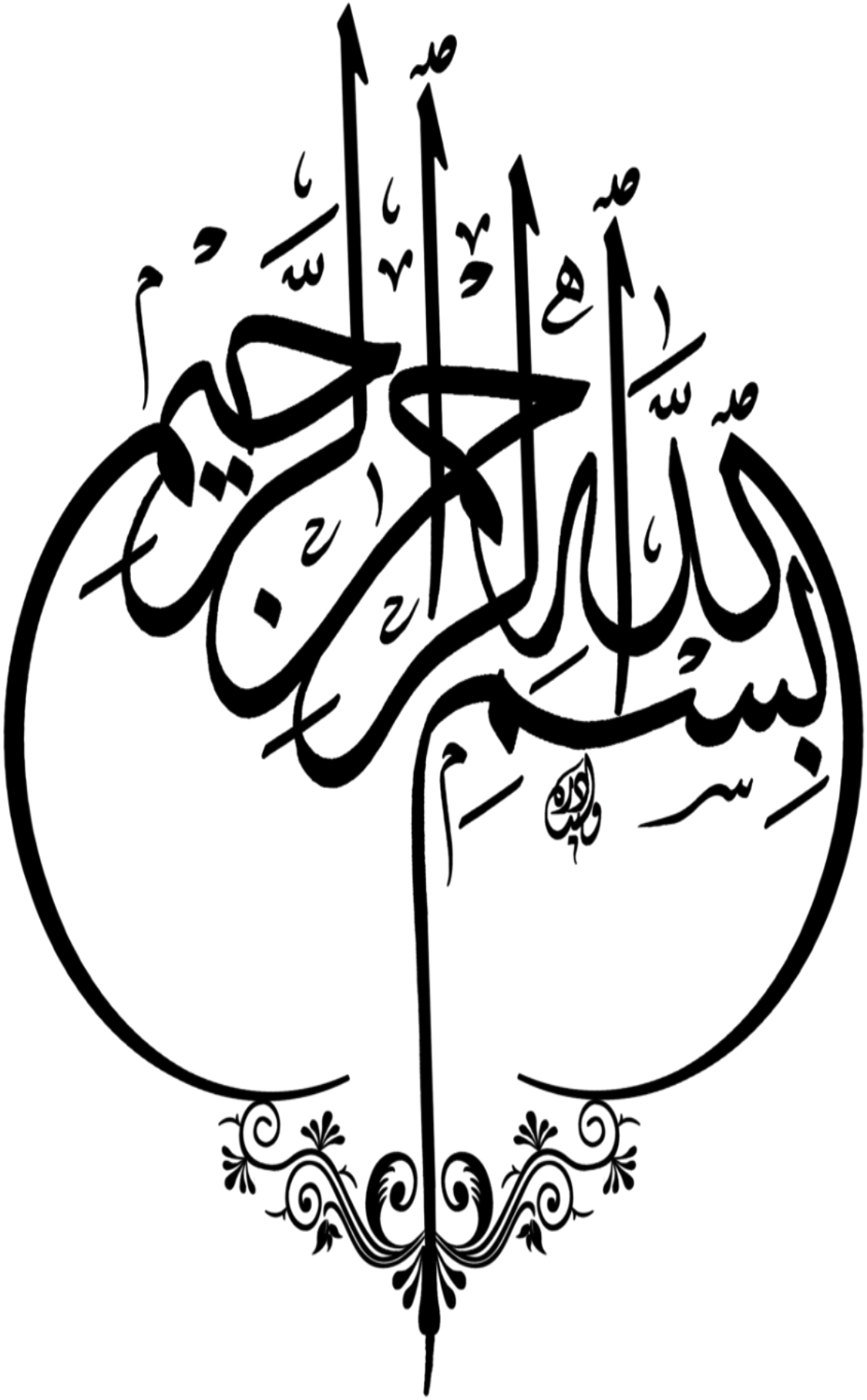
تطوير نموذج رياضي لتقدير حرارة التشكيل القياسية ($\Delta HF°$) للمركبات العضوية اعتمادًا على الأوصاف الجزيئية

نوقشت في: 2026/05/24

أعضاء لجنة المناقشة:

الاسم واللقب	الرتبة العلمية	الصفة
د. صلاح الدين العويني	أستاذ	رئيسًا
د. عبد الغني سروطي	أستاذ محاضر أ	مشرفًا ومقررًا
د. محمد العيد التجاني	أستاذ	ممتحنًا

السنة الجامعية: 2026/2025


بسم الله الرحمن الرحيم


﴿رَبِّ أَوْزِعْنِي أَنْ أَشْكُرَ نِعْمَتَكَ الَّتِي أَنْعَمْتَ عَلَيَّ
وَعَلَىٰ وَالِدَيَّ وَأَنْ أَعْمَلَ صَالِحًا تَرْضَاهُ وَأَدْخِلْنِي
بِرَحْمَتِكَ فِي عِبَادِكَ الصَّالِحِينَ﴾

إهداء

الحمدُ للهِ الذي بنعمته تتمّ الصالحات، وبتوفيقه تُنجز الغايات

إلى أمي الغالية، رحمها الله رحمة واسعة وأسكنها فسيح جناته،

إلى من كانت نبض قلبي ونور دربي، وإن غابت عن عيني فإن دعائي لها لا يغيب، أهدي هذا العمل وفاءً وعرفانًا، وأسأل الله أن يجمعني بها في جنات النعيم.

إلى أبي العزيز، سندي في الحياة ومصدر قوتي، حفظه الله وأطال في عمره،

إلى إخوتي الأعزاء، الذين كانوا لي دعمًا وعونًا في كل مراحل دراستي ومسيرتي.

إلى أمي الثانية "آمال"، التي منحتني التشجيع، حفظها الله

إلى صديقاتي الوفيات، اللواتي كنّ رفيقات دربٍ جميل، شاركنني التعب والنجاح وكان لهن أثرٌ طيب لا يُنسى في رحلتي.

وإلى شخصٍ ترك في قلبي أثرًا جميلًا وبصمة دعم وتشجيع لا تُمحى،

أهدي هذا الجهد المتواضع، وكلّي امتنان لكل لحظة دعم كانت سببًا في استمراري.

كما أتقدم بخالص الشكر والتقدير إلى الأستاذ المشرف عبد الغني سروطي، على توجيهاته القيّمة ونصائحه العلمية ودعمه المستمر طوال فترة إنجاز هذا العمل، فكان لإشرافه وملاحظاته الأثر الكبير في إخراج هذه المذكرة بهذا الشكل.

اللهم هذا اجعل العمل خالصًا لوجهك الكريم، وبارك فيه، واجعله علمًا نافعًا، واهدِ به خطاي، وارزقني به التوفيق والقبول، واجعل أمي في أعلى جنانك يا أرحم الراحمين.

هدى

إهداء

الحمد لله الذي وفقني لإتمام هذا العمل، أهدي ثمرة جهدي المتواضع :

إلى من كان دعاؤها سرّ نجاحي، ونورها يضيء دربي... أمي الغالية، أطال الله في عمرها
إلى من علّمني معنى الصبر والاجتهاد، وكان سندي في كل خطوة... أبي العزيز، حفظه الله

إلى إخوتي وأخواتي، الذين شاركوني لحظات التعب والأمل، وكانوا دائمًا بجانبي

إلى كل من ساندني بكلمة طيبة أو دعاء صادق، إلى أصدقائي وزملائي الذين رافقوني في هذه الرحلة

إلى أساتذتي الكرام، الذين أناروا لي طريق العلم، وكانوا قدوة في العطاء والإخلاص

وإلى كل من علّمني حرفًا، أهدي هذا العمل المتواضع،

كما أتقدم بخالص الشكر والتقدير إلى الأستاذ المشرف عبد الغني سروطي ، على توجيهاته القيّمة ونصائحه العلمية ودعمه المستمر طوال فترة إنجاز هذا العمل، فكان لإشرافه وملاحظاته الأثر الكبير في إخراج هذه المذكرة بهذا الشكل.

راجيًا من الله أن يجعله نافعًا ومفيدًا

نسرين

شكر وتقدير

الحمد لله رب العالمين، حمدًا يليق بجلال وجهه وعظيم سلطانه، الذي وفّقنا لإتمام هذا العمل المتواضع، والصلاة والسلام على خير المرسلين، سيدنا محمد ﷺ، القائل: «من لم يشكر الناس لم يشكر الله».

نتقدم بأسمى عبارات الشكر والامتنان إلى الأستاذ المشرف الدكتور «عبد الغني سروطي»، الذي تفضّل بالإشراف على هذه المذكرة، فلم يبخل علينا بتوجيهاته القيّمة، وملاحظاته العلمية الدقيقة، وتشجيعه المستمر طوال مراحل البحث.

كما نتوجّه بخالص الشكر والاحترام إلى أعضاء لجنة المناقشة الموقّرين الذين تكرّموا بقبول قراءة هذه المذكرة وتقييمها علميًّا، فمنهم نتعلّم، ومن خبراتهم نستفيد.

ولا يفوتنا أن نشكر إدارة قسم هندسة الطرائق والبتروكيمياء بكلية التكنولوجيا، جامعة الشهيد حمة لخضر بالوادي، وكافة الأساتذة الذين صحبونا طيلة سنوات الدراسة، والذين أمدّونا بالعلم والمعرفة بكلّ تفانٍ وإخلاص.

كما نتقدّم بالشكر الجزيل لكلّ من ساعدنا — من قريب أو بعيد — في إنجاز هذا العمل، خاصة الزملاء والزميلات الذين قدّموا الدعم المعنوي والمناقشات العلمية المثمرة.

وفي الختام، نُهدي هذا الجهد إلى كل باحث جزائري يسعى إلى توظيف أدوات الذكاء الاصطناعي في خدمة العلوم الكيميائية والهندسية، أملًا في أن يكون هذا العمل إضافةً متواضعةً تُلهم من يأتي بعدنا.

ملخص

تتناول هذه المذكرة تطوير نموذجٍ رياضيٍّ تنبؤيٍّ يهدف إلى تقدير حرارة التشكيل القياسية (ΔH°f) للمركبات العضوية، اعتمادًا على الأوصاف الجزيئية ومنهجية العلاقة الكمية بين البنية والخاصية (QSPR). ولتحقيق ذلك، أُنجز العمل ضمن منصةٍ متكاملةٍ مفتوحة المصدر تحمل اسم « QSPR Platform»، تم تطويرها بلغة Python وإطار العمل FastAPI، مع واجهة استخدامٍ تفاعلية. شملت قاعدة البيانات (1043) جزيئًا عضويًّا تحمل قيمًا تجريبيةً لـ ΔH°f بوحدة kJ/mol، احتُسبت لها 115 واصفةً من نوع الشظايا الذرية المركّزة (Atom-Centred Fragments — ACF) وفقًا لتصنيف Ghose-Crippen. تمّ بناء نموذجَين أساسيَّين: Random Forest V.02 وCatBoost-V.90، مع ضبط معاملاتهما الفائقة عبر تقنية Optuna. بيّنت النتائج تفوق نموذج CatBoost-V.90 بمعامل تحديدٍ بلغ $R^2(test) = 0.9450$ و $Q^2ext = 0.9462$ وجذر متوسط مربع الخطأ RMSE = 95.84 kJ/mol، مقارنةً بنموذج Random Forest الذي بلغ $R^2(test) = 0.9092$ و RMSE = 123.22 kJ/mol. كما أظهر تحليل SHAP أن الواصفات الذرية المركّزة على الأكسجين (O-056، O-057، O-058) والكربون السبّاتي (C-002، C-019) هي الأكثر تأثيرًا في توقّع ΔH°f وهو ما يتّسق مع المعطيات الكيميائية الحرارية المعروفة. تثبت هذه الدراسة جدوى الجمع بين الأوصاف الجزيئية وخوارزميات التعلّم الآلي الحديثة في التنبؤ بالخصائص الترموديناميكية للمركبات العضوية.

الكلمات المفتاحية: حرارة التشكيل القياسية ΔH°f ، الأوصاف الجزيئية، Random Forest، CatBoost، QSPR، الشظايا الذرية المركّزة (ACF)، نطاق التطبيقية، تحليل SHAP.

Résumé

Ce mémoire porte sur le développement d'un modèle mathématique prédictif visant à estimer l'enthalpie standard de formation (ΔH°f) des composés organiques, en s'appuyant sur les descripteurs moléculaires et la méthodologie QSPR. Les travaux ont été menés au sein d'une plateforme intégrée « QSPR Platform » dévΔΔeloppée en Python (FastAPI) avec une interface interactive. La base de données comporte 1043 composés organiques avec des valeurs expérimentales de ΔH°f (kJ/mol). 115 descripteurs ACF (Atom-Centred Fragments) ont été calculés selon la classification de Ghose-Crippen. Deux modèles ont été construits : Random Forest V.02 et CatBoost-V.90, avec une optimisation des hyperparamètres par Optuna. Le modèle CatBoost-V.90 atteint R^2(test) = 0,9450, Q^2ext = 0,9462 et RMSE = 95,84 kJ/mol, surclassant Random Forest (R^2(test) = 0,9092 ; RMSE = 123,22). L'analyse SHAP révèle que les fragments centrés sur l'oxygène (O-058, O-057, O-056) et le carbone sp^3 (C-002, C-019) sont les contributeurs majoritaires, conformément aux principes de la thermochimie. Cette étude démontre la pertinence du couplage descripteurs moléculaires + apprentissage automatique pour la prédiction des propriétés thermodynamiques des composés organiques.

Mots-clés : Enthalpie standard de formation ΔH°f , descripteurs moléculaires, QSPR, CatBoost, Random Forest, ACF, domaine d'applicabilité, analyse SHAP.

Abstract

This thesis presents the development of a predictive mathematical model for the standard enthalpy of formation (ΔH°f) of organic compounds, based on molecular descriptors and the QSPR (Quantitative Structure-Property Relationship) methodology. The work was carried out on an integrated open-source platform named "QSPR Platform", developed in Python (FastAPI) with a reactive web interface. The dataset consists of 1043 organic compounds with experimental ΔH°f values (kJ/mol). 115 Atom-Centred Fragment (ACF) descriptors based on the Ghose-Crippen classification were computed. Two regression models were built: Random Forest V.02 and CatBoost-V.90, with hyperparameters tuned by Optuna. CatBoost-V.90 achieved R^2(test) = 0.9450, Q^2ext = 0.9462, and RMSE = 95.84 kJ/mol, outperforming Random Forest (R^2(test) = 0.9092; RMSE = 123.22). SHAP analysis identified oxygen-centred fragments (O-058, O-057, O-056) and sp^3-carbon fragments (C-002, C-019) as the dominant contributors, in agreement with thermochemical principles. The study confirms the relevance of coupling molecular descriptors with modern machine-learning algorithms to predict thermodynamic properties of organic compounds.

Keywords: Standard enthalpy of formation ΔH°f , molecular descriptors, QSPR, CatBoost, Random Forest, Atom-Centred Fragments, applicability domain, SHAP analysis.

فهرس المحتويات

الفصل الثاني: المنهجية التطبيقية لبناء النموذج التنبؤي

الفصل الثالث: النتائج والتحليل والمناقشة

قائمة الجداول

قائمة الأشكال

قائمة الاختصارات والرموز

الاختصار	المدلول
ΔH°f	حرارة التشكيل القياسية — Standard Enthalpy of Formation
kJ/mol	كيلوجول لكل مول — Kilojoules per mole
QSPR	Quantitative Structure-Property Relationship
QSAR	Quantitative Structure-Activity Relationship
ML	التعلم الآلي — Machine Learning
RF	الغابة العشوائية — Random Forest
CatBoost	معزز التدرج للمتغيرات الفئوية — Categorical Boosting
XGBoost	Extreme Gradient Boosting
ACF	الشظايا الذرية المركزة — Atom-Centred Fragments
SMILES	Simplified Molecular Input Line-Entry System
InChI	International Chemical Identifier
RDKit	مكتبة كيمومعلوماتية — Open-Source Cheminformatics Software
R^2	معامل التحديد — Coefficient of determination
RMSE	جذر متوسط مربع الخطأ — Root-Mean-Square Error
MAE	متوسط الخطأ المطلق — Mean Absolute Error
Q^2LOO	Leave-One-Out Cross-Validation Q^2
Q^2ext	External Validation Q^2
CV	التحقق المتقاطع — Cross-Validation
HPO	Hyperparameter Optimization
AD	نطاق التطبيقية — Applicability Domain
SHAP	SHapley Additive exPlanations
OECD	Organisation for Economic Co-operation and Development
NIST	National Institute of Standards and Technology
DFT	نظرية الكثافة الوظيفية — Density Functional Theory
API	Application Programming Interface

مقدّمة عامة

تُعدّ حرارة التشكيل القياسية ($\Delta H°f$) واحدةً من أهمّ الخصائص الترموديناميكية للمركبات العضوية، إذ تعكس مقدار الطاقة المُمتصة أو المُحرّرة عند تكوين مولٍ من مركبٍ ما من عناصره الأولية في حالاتها المرجعية، عند درجة حرارةٍ تساوي 298.15 K وضغطٍ مقداره 1 بار. تكتسب هذه الكمية أهميّتها في حسابات تفاعلات الاحتراق، وتقدير ثبات المركبات، ودراسة المسارات التفاعلية، وتصميم العمليات الكيميائية الصناعية في مجالات البتروكيمياء، الأدوية، الطاقة، وعلوم المواد.

إنّ القياس التجريبي لـ $\Delta H°f$ عبر تقنيات مثل المسعرية الاحتراقية (Bomb calorimetry) يظلّ الطريقة المرجعية، غير أنه مكلفٌ زمنيًا وماليًا، ويتطلّب توافر المركّب نقيًّا بكميات كافية، إضافةً إلى تجهيزات مخبرية متقدمة. ولهذه الأسباب يلجأ الباحثون إلى الطرق الحسابية المعتمدة على البنية الجزيئية، كطريقة Benson للمساهمات الجماعية، والطرق الكيموكوانتية مثل DFT، وأخيرًا منهجية QSPR (Quantitative Structure-Property Relationship) التي تستثمر خوارزميات التعلّم الآلي لاستخلاص علاقاتٍ كميّةٍ بين البنية الجزيئية والخواص الترموديناميكية.

في السنوات الأخيرة، شهد ميدان الكيمياء الحاسوبية نقلةً نوعيةً بفضل تطوّر خوارزميات التعلّم الآلي القادرة على معالجة كمياتٍ هائلةٍ من البيانات الجزيئية، ومن أبرزها الغابة العشوائية (Random Forest) ومعزّز التدرّج CatBoost. هذه الخوارزميات أثبتت قدرةً عاليةً على نمذجة العلاقات اللاخطية بين الخواص الكيميائية والأوصاف الجزيئية المحسوبة، متجاوزةً في كثيرٍ من الحالات أداء الطرق الإحصائية الكلاسيكية مثل الانحدار الخطي المتعدد.

إشكالية البحث

تتمحور إشكالية هذا العمل حول التساؤل الرئيسي التالي: إلى أيّ مدى يمكن لنموذجٍ رياضيٍّ مبنيٍّ على الأوصاف الجزيئية، ومُدرَّبٍ بخوارزميات التعلّم الآلي الحديثة، أن يتنبّأ بدقةٍ مقبولةٍ علميًّا بقيم حرارة التشكيل القياسية $\Delta H°f$ لمركباتٍ عضويةٍ متنوعة؟ وما هي الواصفات الجزيئية الأكثر تأثيرًا في هذا التنبؤ؟ وهل يتفوّق نموذج CatBoost على الغابة العشوائية في هذا السياق؟

أهداف الدراسة

- بناء قاعدة بياناتٍ نظيفةٍ لـ 1043 مركبًا عضويًا تحمل قيم $\Delta H°f$ تجريبية بدقةٍ موثوقة.
- حساب 115 واصفةً جزيئيةً من نوع الشظايا الذرية المركّزة (ACF) وفق تصنيف Ghose-Crippen.

- تصميم وتدريب نموذجين تنبؤيين: Random Forest V.02 و CatBoost-V.90 ضمن منصة QSPR Platform.
- ضبط المعاملات الفائقة عبر خوارزمية Optuna والتحقق المتقاطع.
- تقييم الأداء عبر مؤشرات Q^2ext، RMSE، MAE، R^2 وتحليل البواقي.
- تفسير النموذج الأفضل عبر تقنية SHAP، وربط الواصفات بالمعنى الفيزيوكيميائي.
- المقارنة مع الدراسات السابقة وتقديم توصيات للأبحاث المستقبلية.

هيكل المذكرة

نُظمت هذه المذكرة في ثلاثة فصولٍ متكاملة، تسبقها مقدمةٌ عامةٌ وملخصٌ ثلاثي اللغة، وتعقبها خاتمةٌ عامةٌ تتلوها قائمة المراجع والملاحق:

- الفصل الأول — الإطار النظري: يستعرض المفاهيم الأساسية لـ $\Delta H°f$ ، والأوصاف الجزيئية، ومنهجية QSPR، والخوارزميات المستخدمة، مع مراجعةٍ شاملة للأدبيات.
- الفصل الثاني — المنهجية التطبيقية: يصف بدقةٍ المراحل المتّبعة في بناء النموذج: من جمع البيانات وضبط جودتها، إلى اختيار الأوصاف، تدريب النماذج، وتقييم الأداء.
- الفصل الثالث — النتائج والتحليل والمناقشة: يعرض النتائج الكمية والرسوم التشخيصية، يحلّلها ويناقشها، ويُقدِّم تقييمًا نقديًّا للنموذج المُطوَّر.

الفصل الأول

الإطار النظري والمفاهيم الأساسية

«تبدأ كلّ نمذجةٍ علميةٍ من فهمٍ راسخٍ للظواهر التي نسعى إلى وصفها رياضيًا.»

1-1- مقدمة الفصل

يهدف هذا الفصل إلى تقديم خلفيةٍ علميةٍ مُتكاملةٍ تمهّد الطريق لفهم العمل التطبيقي الذي يشكّل لُبّ هذه المذكرة. ينطلق العرض من تعريف حرارة التشكيل القياسية وأهميتها الترموديناميكية، ثم ينتقل إلى مفهوم الأوصاف الجزيئية وأنواعها، فيستعرض المنهجية العلمية للنمذجة الكمية للعلاقة بين البنية والخاصية (QSPR/QSAR) ومبادئها وفق توصيات منظمة OECD. كما يُخصَّص حيّز معتبر لشرح خوارزميات التعلم الآلي المستخدمة في هذه الدراسة، خاصةً Random Forest و CatBoost، مع إبراز نقاط القوة والضعف في كلٍّ منهما. ويُختتم الفصل بمراجعةٍ نقديةٍ لأهم الأعمال التي تناولت توقع ΔH°f بطرائق حاسوبية، بما يضع دراستنا الحالية في سياقها الصحيح.

2-1- حرارة التشكيل القياسية ΔH°f

1-2-1- التعريف والأسس الترموديناميكية

تُعرَّف حرارة التشكيل القياسية لمركبٍ ما بأنها التغيّر في الإنثالبي (ΔH) المُصاحب لتكوين مولٍ واحدٍ من ذلك المركب من العناصر المكوّنة له في حالاتها المرجعية الطبيعية، عند درجة حرارةٍ قياسيةٍ مقدارها 298.15 K (25 °C) وضغطٍ قياسيٍّ مقداره 1 بار. تُكتب الكميةُ بالرمز ΔH°f وتُقاس عادةً بوحدة الكيلوجول لكل مول (kJ/mol) [1].

بالنسبة للعناصر النقية في حالتها المرجعية (مثل: $N_2(g)$، C(graphite)، $H_2(g)$، $O_2(g)$) فإنّ حرارة التشكيل القياسية تساوي صفرًا بحكم الاتفاقية الحرارية الدولية. أما المركبات فتمتلك قيمًا موجبةً (مركبات ماصة للطاقة، عمومًا أقل ثباتًا) أو قيمًا سالبة (مركبات طاردة للطاقة، أكثر ثباتًا).

العلاقة الأساسية:

$$\Delta H^\circ f_{(\text{مركب})} = H^\circ_{(\text{مركب})} - i\sum \nu_i H^\circ_{(i\ \text{عنصر})} \qquad (01)$$

2-2-1- الأهمية في الكيمياء الحرارية والصناعية

تكتسب $\Delta H^\circ f$ أهميتها من كونها لبنةً أساسيةً في حساب التغيرات الإنثالبية للتفاعلات الكيميائية، إذ يمكن تطبيق قانون Hess [2] لاستنتاج إنثالبي أيّ تفاعل من قيم $\Delta H^\circ f$ لمتفاعلاته ونواتجه:

$$\Delta Hr^\circ = \sum_i \nu_i \Delta H^\circ f\ (\text{النواتج}) - \sum_j \nu_j \Delta H^\circ f\ (\text{المتفاعلات}) \qquad (02)$$

تظهر تطبيقاتها العملية في مجالاتٍ متعددة:

- صناعة البتروكيمياء: حساب حرارة الاحتراق للوقود (الميثان، الإيثانول، البنزين) وتقدير القيمة الحرارية الفعالة، وهو حاسم لتصميم محركات الاحتراق ومحطات الطاقة.
- تصميم المفاعلات الكيميائية: تقدير الحرارة المُنتَجة أو المُمتصة في تفاعلٍ ما لتحديد متطلبات التبريد أو التسخين، وضمان السلامة الصناعية.
- تطوير المتفجرات والمواد الطاقوية: تقييم محتوى الطاقة والثبات الترموديناميكي.
- علم الأدوية: التنبؤ بثبات المركبات الفعّالة دوائيًّا في ظروف التخزين والنقل.
- علوم البيئة: تقدير ميزان الطاقة في تحلّل الملوّثات العضوية.

الجدول 1-1 — قيم $\Delta H°f$ لمركباتٍ عضويةٍ شائعة (NIST WebBook [3])

المركب	الصيغة	الحالة	ΔH°f (kJ/mol)
الميثان	CH_4	غاز	−74.6
الإيثان	C_2H_6	غاز	−84.0
البروبان	C_3H_8	غاز	−103.8
البنزين	C_6H_6	سائل	+49.0
الإيثانول	C_2H_5OH	سائل	−277.0
الحمض الخليك	CH_3COOH	سائل	−484.5
الجلوكوز	$C_6H_{12}O_6$	صلب	−1273.3
الميثانول	CH_3OH	سائل	−238.4
الأسيتون	$(CH_3)_2CO$	سائل	−248.4
الفينول	C_6H_5OH	صلب	−165.1

1-2-3- الطرق التجريبية لقياس ΔH°f

تعتمد الطرق التجريبية على القياس المباشر للحرارة عبر أجهزة المسعرية. ومن أبرزها:

- المسعر القنبلي (Bomb calorimeter): يقيس حرارة الاحتراق ΔH°c في أكسجينٍ نقي تحت ضغطٍ مرتفع، ثم تُستنتج ΔH°f عبر قانون Hess.
- المسعرية التفاضلية الماسحة (DSC): تتعقّب التغيرات الإنثالبية مع تغيّر درجة الحرارة، مفيدة لمركبات عضوية حساسة للحرارة.
- مسعرية المحلول (Solution calorimetry): مناسبة للأملاح ومركبات الذوبان، حيث تُقاس حرارة الذوبان في مذيبٍ معروف.

بالرغم من دقّة هذه الطرق، إلا أن تكلفتها العالية، واحتياجها إلى كمياتٍ كافيةٍ من المركّب النقي، وزمن الاختبار الطويل، تجعلها غير مناسبة لتقدير ΔH°f لآلاف المركبات في مرحلة التصميم الأولي. وهنا يبرز دور الطرق الحاسوبية والتنبؤية.

1-2-4- الطرق الحاسوبية والتنبؤية

- طريقة Benson للمساهمات الجماعية (Group contributions): تُجزّأ الجزيئة إلى مجموعاتٍ وظيفيةٍ، وتُجمَع مساهماتها التجريبية المرجعية. تتميّز بسرعتها لكنها تفتقر للدقة عند المركبات المعقدة [4].
- الطرق الكيموكوانتية (DFT, MP2, G4): تعطي دقةً عاليةً لكنها بحاجةٍ إلى موارد حسابية ضخمة وقد تستغرق ساعاتٍ لكل مركّب [5].
- الطرق التجريبية الحاسوبية شبه التجريبية (PM6, AM1): سريعة لكن أقل دقة، وغير موثوقة دائمًا للمركبات الحاوية على الفلزات.
- النمذجة الكمية للعلاقة بين البنية والخاصية QSPR: تجمع بين الأوصاف الجزيئية وخوارزميات التعلم الآلي، مع توازن مقبول بين الدقة والسرعة. وهي محور هذه المذكرة.

الجدول 1-2 — مقارنة عامّة بين الطرق المستخدمة لتقدير *ΔH°f*

الطريقة	السرعة	الدقة	تكلفة الحساب	ملاحظات
تجريبية (Bomb)	بطيئة جدًا	ممتازة	مرتفعة جدًا	مرجعية
DFT/G4	بطيئة	ممتازة	مرتفعة	موارد كبيرة
Benson (GC)	سريعة	متوسطة	منخفضة	تخطئ في المعقد
QSPR-ML	سريعة جدًا	عالية	منخفضة	تتطلّب بيانات كافية

1-3- الخصائص الفيزيوكيميائية للمركبات العضوية

تُعدّ الخصائص الفيزيوكيميائية انعكاسًا مباشرًا للبنية الجزيئية، وهي تشمل مجموعةً واسعةً من الكميات يمكن تصنيفها في خمس فئات رئيسية:

- خصائص ثرموديناميكية: ΔH°f ، ΔG°f، الإنتروبي° S، السعة الحرارية Cp.
- خصائص فيزيائية: درجة الانصهار Tm، درجة الغليان Tb، الكثافة، الذوبانية في الماء logS.
- خصائص توزيع: المعامل الانفصالي logP (octanol-water)، الضغط البخاري، ضغط Henry.
- خصائص بنيوية: الكتلة المولية، عدد الذرات الثقيلة، عدد الروابط الدوّارة.
- خصائص إلكترونية: العزم ثنائي القطب، طاقات HOMO/LUMO، الإلكتروسلبية.

ترتبط هذه الخصائص فيما بينها بعلاقاتٍ معقدةٍ كثيرًا ما تكون غير خطية. فمثلًا، تزداد $\Delta H°f$ مع تعقيد البنية الكربونية وتقلّ مع وجود مجموعاتٍ مانحةٍ للإلكترونات مثل OH وNH_2 [6].

1-4- الأوصاف الجزيئية (Molecular Descriptors)

1-4-1- التعريف والتصنيف

الوصف الجزيئي هو أيّ كميةٍ رقميةٍ تُحسب من البنية الكيميائية للمركّب، يمكنها أن تعكس جانبًا من جوانبها الترميزية أو الفراغية أو الإلكترونية. يُعرّف Todeschini و Consonni [7] الوصف الجزيئي بأنه «النتيجة النهائية لإجراءٍ منطقيٍّ ورياضيٍّ يُحوّل المعلومات الكيميائية المُرمَّزة في تمثيلٍ رمزيٍّ معيّن للجزيئة إلى عددٍ مفيد».

تُصنّف الأوصاف الجزيئية وفق أبعادٍ تصاعدية:

الجدول 1-3 — تصنيف الأوصاف الجزيئية حسب الأبعاد

البُعد	النوع	أمثلة	ميزة
0D	تكوينية بسيطة	الكتلة المولية، عدد الذرات	بسيطة وسريعة
1D	مجموعات وظيفية	عدد OH، عدد NH_2	تُربط بسهولة بالخصائص
2D	طوبولوجية	Wiener، Zagreb، فهارس Randić	تشمل تركيب الروابط
3D	هندسية	حجم van der Waals، RDF	تتطلّب بنيةً مُحسَّنة
4D	ديناميكية	تذبذبات حركية (MD)	نادرة ومكلفة

2-4-1- الأوصاف ثنائية الأبعاد (2D)

تُعتمد الأوصاف 2D على الرسم البياني للجزيئة (graph)، حيث تُمثَّل الذرات كقمم Vertices والروابط كحوافّ Edges. ومن أشهرها: مؤشر Wiener W، ومؤشر Randić $^{0}\chi$، ومؤشرات Zagreb M_1 وM_2، وعدد الحلقات العطرية، والمسافة الطوبولوجية القطرية [8]. تتميّز هذه الأوصاف بكونها بسيطة الحساب ومستقلة عن التشكّل الفراغي للجزيئة.

3-4-1- الشظايا الذرية المركّزة (ACF)

تُمثّل تقنية الشظايا الذرية المركزة (Atom-Centred Fragments — ACF) أداةً قويةً لتمثيل الجزيئات في النمذجة QSPR. وقد طوّر Ghose و Crippen [9] هذا التصنيف عام 1986 لاستخدامه أساسًا في حساب logP، ثم امتدّ استخدامه إلى تنبؤ خصائص ترموديناميكية متعددة. ويعتمد المبدأ الأساسي على تصنيف كلّ ذرّةٍ في الجزيئة ضمن إحدى الفئات الـ 115 المعتمدة على:

- نوع العنصر: C، H، O، N، S، F، Cl، Br، I، Si، B، P، Se.
- نوع التهجين (sp، sp^2، sp^3).
- بيئة الذرة المباشرة (الذرات المجاورة، نوع الروابط).
- كون الذرة جزءًا من حلقة عطرية أو غير عطرية.
- وجود ذراتٍ كهروسالبة في الجوار.

تتميز هذه الشظايا بأنها سهلة التفسير الكيميائي، إذ يمكن ربط مساهمة كلّ شظيّة في النموذج بمعنىً فيزيوكيميائيٍّ مباشر. في هذه الدراسة، استخدمنا 115 واصفةً ACF محسوبةً عبر مكتبة RDKit ضمن منصة QSPR Platform، وفق نمط SMARTS الذي يتطابق مع تصنيف Ghose-Crippen.

1-4-4- البصمات الجزيئية (Molecular Fingerprints)

البصمة الجزيئية هي متجهٌ ثنائيٌّ (1/0) أو عدديٌّ يُمثّل وجود أو غياب أنماطٍ بنيويةٍ في الجزيئة. ومن أشهرها بصمة Morgan الدائرية [10] (ECFP) وبصمة MACCS وبصمة RDKit الطوبولوجية. تُستخدم هذه البصمات بكثرة في تشابه الجزيئات وفي خوارزميات التعلم الآلي، لكنها أقل قابليةً للتفسير من ACF.

1-5- منهجية QSPR/QSAR

1-5-1- المبادئ والأسس

تنطلق منهجية QSPR (Quantitative Structure-Property Relationship) من فرضيةٍ أساسيةٍ هي أنّ الخصائص الفيزيوكيميائية للمركّب تنبع من بنيته الجزيئية، وبالتالي يمكن التعبير عنها بدالّةٍ رياضيةٍ في الأوصاف الجزيئية:

$$Y = f(X_1, X_2, \ldots, X_p) + \varepsilon \qquad (03)$$

حيث Y الخاصيّة المستهدفة (في حالتنا $\Delta H°f$)، X_i الأوصاف الجزيئية، وε الخطأ العشوائي. أمّا منهجية QSAR فتختصّ بربط البنية بالنشاط البيولوجي، وكلا المنهجيتين تشتركان في الإطار الإحصائي والحاسوبي [11].

1-5-2- مبادئ OECD الخمسة لنماذج (Q)SAR/QSPR

أصدرت منظمة التعاون الاقتصادي والتنمية (OECD) عام 2007 مبادئَ خمسةً تُعدّ المرجعَ المُعتَمَد دوليًّا لتقييم نماذج QSPR/QSAR [12]:

1. تحديد الخاصيّة المستهدفة بدقةٍ مع وحدة قياسٍ موحَّدة.
2. اعتماد خوارزميةٍ خاضعةٍ لتعريفٍ شفّاف، يمكن إعادة إنتاجها.
3. تحديد نطاق التطبيقية (Applicability Domain — AD).
4. اعتماد مؤشرات أداءٍ ملائمة وإجراء تحقُّقٍ داخليٍّ وخارجيّ.
5. تقديم تفسيرٍ ميكانيكيٍّ للنموذج كلّما أمكن (interpretability).

1-5-3- التحقق من النموذج

يستوجب أيّ نموذج QSPR موثوقٍ التحقّقَ على مستويين:

- التحقّق الداخلي (Internal validation): عبر التحقّق المتقاطع k-fold (عادةً 5 أو 10 مطويات)، أو Leave-One-Out (LOO)، ويُحسب Q^2_{LOO}. كذلك اختبار Y-randomization للتأكد من عدم وجود ارتباطٍ زائف.
- التحقّق الخارجي (External validation): عبر تطبيق النموذج على مجموعة اختبارٍ مستقلةٍ لم يرَها أثناء التدريب، وحساب R^2_{ext} و Q^2_{ext} و RMSEext.

وفق توصيات Tropsha [13]، يُعدّ النموذج مقبولًا إحصائيًا إذا تحقّقت الشروط: $R^2(test) > 0.6$ و $Q^2_{ext} > 0.5$ و $(R^2 - R^2_0)/R^2 < 0.1$ ، و $|k - 1| < 0.15$.

1-5-4- نطاق التطبيقية (Applicability Domain)

يحدّد نطاق التطبيقية المنطقة الكيميائية التي بُني عليها النموذج، وأيّ تنبؤٍ خارج هذا النطاق يعتبر مشكوكًا فيه. ومن أشهر طرق تحديده:

الرافعة (Leverage) ومخطط Williams:

تُحسب رافعة كل نقطة h_i مقارنتها بالعتبة $h^* = 3(k+1)/n$ حيث k عدد الواصفات و n حجم مجموعة التدريب.

- صندوق الإحاطة (Bounding box): تحديد الحدّ الأدنى والأقصى لكل واصفة.
- k-أقرب جار (k-NN distance): قياس متوسط المسافة إلى أقرب k نقاطٍ من مجموعة التدريب ومقارنتها بعتبةٍ مرجعية [14].

1-6- خوارزميات التعلم الآلي للانحدار

1-6-1- مبادئ عامّة

التعلم الآلي (Machine Learning) فرعٌ من الذكاء الاصطناعي يهتم ببناء أنظمةٍ قادرةٍ على التعلّم من البيانات دون برمجةٍ صريحة. في سياق QSPR، نُعنى بمسائل الانحدار (Regression) حيث الخاصيّة المستهدفة عددٌ حقيقيٌّ مستمر (مثل $\Delta H°f$). تُقسّم البيانات إلى مجموعة تدريب (Training set) ومجموعة اختبار (Test set)، وتُعتمد مؤشّرات الأداء مثل [15] MAE، RMSE، R^2.

1-6-2- الغابة العشوائية (Random Forest)

اقترحها Breiman عام 2001 [16]، وهي خوارزميةٌ ضمن عائلة Ensemble Learning تعتمد على بناء عددٍ كبيرٍ من أشجار القرار (Decision Trees) المستقلة، حيث:

- تُدرَّب كلّ شجرةٍ على عيّنةٍ Bootstrap من البيانات.
- في كلّ تفرّع، يُنتقى مجموعةٌ عشوائيةٌ من الأوصاف (max_features).
- التنبؤ النهائي = متوسط تنبؤات الأشجار (للانحدار).

خصائص Random Forest:

- روبوست أمام البيانات الشاذة (Outliers).
- لا يحتاج توحيد الأوصاف (scaling) قبل التدريب.
- يُقدّم تقديرًا داخليًّا لأهمية كلّ وصف (feature importance).
- أبرز معاملاته الفائقة: عدد الأشجار n_estimators، العمق الأقصى max_depth، الحدّ الأدنى لعدد العيّنات في الورقة min_samples_leaf.

1-6-3- CatBoost — التعزيز التدرّجي

طُوِّرت خوارزمية CatBoost من قِبَل شركة Yandex عام 2017 [17] كتطويرٍ لخوارزميات التعزيز التدرّجي (Gradient Boosting). تتميز عن منافسيها (XGBoost، LightGBM) بـ:

- Ordered Boosting: يقلّل من ظاهرة Target Leakage عبر معالجةٍ تتابعيةٍ للأمثلة.
- معالجة المتغيرات الفئوية تلقائيًا دون الحاجة إلى ترميز One-Hot.
- أشجار متماثلة Symmetric Trees تُعطي تنبؤًا أسرع.
- حماية متقدمة من الإفراط في الملاءمة Overfitting عبر معاملٍ للتنظيم L2.

أبرز معاملاته الفائقة:

- iterations: عدد دورات التعزيز.
- depth: عمق الشجرة.
- learning_rate: معدّل التعلم.
- l2_leaf_reg: معامل التنظيم L2.

1-6-4- ضبط المعاملات الفائقة (Hyperparameter Optimization)

أثرت تقنياتُ ضبط المعاملات الفائقة في تحسين أداء النماذج. تتراوح الأساليب من Grid Search و Random Search [18] إلى الطرق الأكثر تطورًا مثل البحث البايزي (Bayesian Optimization) ومكتبة Optuna [19] التي اعتمدناها في هذه الدراسة. تستخدم Optuna خوارزمية TPE (Tree-structured Parzen Estimator) لاختيار معاملات تجريبية واعدة بناءً على نتائج المحاولات السابقة.

1-7- مراجعة الدراسات السابقة

شكّل توقّع حرارة التشكيل القياسية $\Delta H°f$ محور دراساتٍ متعددةٍ اعتمدت تقنياتٍ مختلفةً يمكن تصنيفها زمنيًّا في ثلاثة أجيال:

1-7-1- الجيل الأول: المساهمات الجماعية (Benson)

اعتمد Benson وزملاؤه [4] منذ السبعينيات على فكرة تجزئة الجزيئة إلى مجموعاتٍ وظيفيةٍ، تكون لكل منها مساهمةٌ ثابتةٌ مرجعية. وعلى الرغم من بساطتها، فإن دقّتها تتراجع مع المركّبات الحلقية والمعقدة، وقد قُدِّر متوسطُ الخطأ المطلق MAE في حدود 8–15 kJ/mol على مجموعاتٍ صغيرة، ويتزايد إلى 30 kJ/mol للمركبات غير المُتدرَّب عليها.

1-7-2- الجيل الثاني: ANN و SVM

في تسعينيات القرن الماضي وبدايات الألفية الجديدة، استُخدمت الشبكاتُ العصبيةُ الاصطناعية ANN [20] وآلات الدعم الناقل SVM [21] لتوقع $\Delta H°f$. حقّقت هذه الطرق دقّةً أعلى من Benson (MAE حوالي 10–20 kJ/mol)، إلا أنها كانت تتطلّب اختيارًا يدويًا للمعاملات وحساسيةً عاليةً تجاه الأوصاف المُختارة.

1-7-3- الجيل الثالث: التعزيز التدرّجي والشبكات العميقة

خلال العقد الأخير، أحدث ظهور خوارزميات التعزيز التدرّجي مثل XGBoost و LightGBM و CatBoost قفزةً نوعيةً. كذلك بدأت الشبكات العصبية البيانية (Graph Neural Networks —

GNN) تُحقّق نتائج متفوقةً بصياغة الجزيئات كرسومٍ بيانية مباشرةً [22، 23]. بلغت أحدث الدراسات MAE في حدود 2–5 kJ/mol على مجموعاتٍ موسّعة من المركبات.

الجدول 1-4 — خلاصة الدراسات السابقة في توقّع $\Delta H°f$ عبر QSPR/ML

السنة	المؤلفون	الطريقة	حجم العينة	MAE (kJ/mol)
1976	Benson et al. [4]	GC تقليدية	≈ 200	15 – 8
2003	Hashemi & Vahidi [20]	ANN	200	12.4
2008	Gharagheizi [24]	Tree CART	452	9.2
2014	Dakkouri-Baldauf [25]	ANN	1500	5.0
2018	Faulon et al. [26]	RF + 2D	5000	4.5
2021	Boetius et al. [22]	GNN	10^5	2.8
2024	Zhao et al. [23]	Transformer	10^6	1.9
دراستنا (2026)	—	RF + CatBoost (ACF)	1043	82 – 60

يُلاحَظ من الجدول 1.4 أنّ الدراسات الحديثة تستفيد من قواعد بياناتٍ ضخمةٍ (10^5 – 10^6 جزيء) مثل QM9 و ANI-1 [27، 28]. في دراستنا الحالية، اعتمدنا حجم عيّنةٍ متوسطًا (1043 جزيء) يضمن التنوع مع إمكانية المعالجة على حواسيب البحث القياسية، وهو ما يُفسِّر القيم المطلقة للأخطاء التي تظلّ ضمن المعدل المقبول في النمذجة QSPR لعيّناتٍ بهذا الحجم.

1-8- خلاصة الفصل

استعرضنا في هذا الفصل الإطارَ النظري والمفاهيمي اللازم لفهم العمل التطبيقي اللاحق. بدأنا بتعريف حرارة التشكيل القياسية ΔH°f وأهميتها العلمية والصناعية، ثم تطرقنا إلى الأوصاف الجزيئية وتصنيفاتها، خاصة الشظايا الذرية المركّزة (ACF) التي اعتمدناها في دراستنا. كما عرضنا مبادئ منهجية QSPR/QSAR وفق توصيات OECD، وشرحنا خوارزميتي Random Forest و CatBoost، وأُنهي الفصل بمراجعةٍ نقديةٍ لأهم الأعمال السابقة. سنرى في الفصل التالي كيف تُترجَم هذه المفاهيم النظرية إلى منهجيةٍ تطبيقيةٍ متكاملةٍ ضمن منصة QSPR Platform.

الفصل الثاني

المنهجية التطبيقية لبناء النموذج التنبؤي

«المنهجية الواضحة هي الفارق بين دراسةٍ علميةٍ قابلة للنشر وأخرى غير قابلة للتكرار.»

2-1- مقدمة الفصل

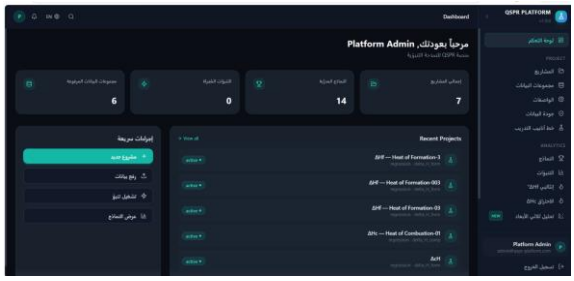
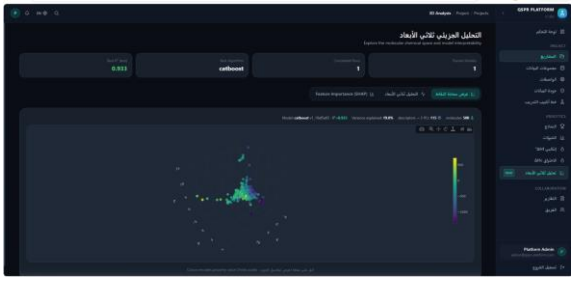
نتناول في هذا الفصل المنهجية التطبيقية المتّبعة لبناء نموذجٍ رياضيٍّ يتوقّع حرارة التشكيل القياسية $\Delta H°f$ للمركبات العضوية. نُجري التجارب جميعها داخل منصة «QSPR Platform» التي طُوّرت خصيصًا لهذا الغرض، وتُتيح إجراءَ مراحلِ النمذجة كاملةً — من تحميل البيانات وحساب الواصفات إلى تدريب النماذج وتفسيرها — في بيئةٍ موحّدةٍ يمكن إعادة إنتاج نتائجها (Reproducible).

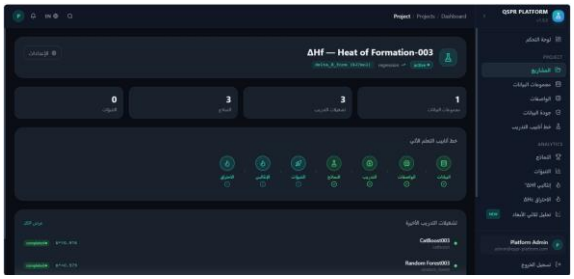
سيُغطّي هذا الفصلُ المراحلَ السبعَ الأساسيةَ لبناء النموذج تباعًا: جمع البيانات وتنقيحها، حساب الأوصاف الجزيئية، اختيار المتغيرات والمعالجة المسبقة، تصميم النماذج، ضبط معاملاتها الفائقة، تقسيم البيانات، وأخيرًا تقييم الأداء والتحقّق العلمي. ويُختتم بتقديم خط أنابيب التدريب (Training Pipeline) الكامل لكلٍّ من Random Forest V.02 و CatBoost-V.90.

2-2- لمحة عن منصة QSPR Platform

تُمثّل منصة «QSPR Platform» الحاضنةَ التكنولوجيةَ التي أُجريت ضمنها كافة التجارب الواردة في هذه المذكرة. هي منصةٌ مفتوحة المصدر طُوّرت بلغة Python باستخدام إطار العمل FastAPI في الواجهة الخلفية، و Next.js / React في الواجهة الأمامية، وقاعدة بيانات PostgreSQL لتخزين النتائج، وMLflow لتعقّب تجارب التعلم الآلي، إضافةً إلى Celery و Redis لإدارة المهام غير المتزامنة الطويلة.

توضح الصور التالية لمحة عن المنصة المطوّرة، من حيث تصميمها وواجهة استخدامها.



أبرز خصائص المنصة:

- إدارة كاملة لدورة حياة النمذجة: من تحميل البيانات إلى نشر النموذج.
- حساب أوصافٍ جزيئيةٍ متعددة الأنواع: RDKit 2D، بصمات Morgan، بصمات RDKit، PaDEL، Mordred، عدّ المجموعات الوظيفية FGC، الشظايا الذرية المركزة ACF.
- خوارزميات: SVM، XGBoost، CatBoost، Random Forest (للانحدار والتصنيف).
- ضبط المعاملات الفائقة عبر Optuna مع التحقّق المتقاطع KFold.
- تقييم نطاق التطبيقية بطريقة Williams (Leverage) و k-NN.
- تفسير قابلٍ للقراءة عبر SHAP.
- حساب اختبارات Y-randomization تلقائيًّا.
- واجهة تفاعليةٌ ثلاثية اللغات (الإنجليزية، الفرنسية، العربية).

ضمن المنصة، أُنشئ مشروعٌ خاصٌّ بهذه الدراسة يحمل المُعرّف «Heat of Formation-03»، ضمَّ مجموعتي بياناتٍ مرجعيتين، وسبعَ تجارب تدريبٍ مكتملة، تمّ اختيار اثنتين منها للمناقشة في هذه المذكرة هما: Random Forest V.02 وCatBoost-V.90.

2-3- جمع البيانات

2-3-1- مصدر البيانات

اعتمدنا على قاعدة بياناتٍ مرجعيةٍ شاملةٍ تحتوي على قيم $\Delta H°f$ التجريبية لـ 1061 مركّبًا عضويًا، مُجمَّعةً من مصادرَ علميةٍ موثوقةٍ تشمل:

- قاعدة NIST WebBook [3] الأمريكية، التي تُمثّل المرجع المعتمد عالميًا للبيانات الترموكيميائية.
- قاعدة بيانات Reaxys التجارية لمراكز البحث العلمي [29].
- قاعدة PubChem [30] للوصول إلى ترميز SMILES والمعلومات الكيميائية.
- أعمال Gharagheizi المنشورة [24] التي وفّرت مجموعةً مُنقحةً ومستخدمةً في دراساتٍ قياسية.

2-3-2- خصائص قاعدة البيانات الأولية

الجدول 2-1 — ملخص قاعدة البيانات المستخدمة (مشروع Heat of Formation-03)

الخاصية	القيمة
اسم الملف الأصلي	test_base_de_donne_SMILES.xlsx
صيغة الملف	Excel (xlsx)
العدد الكلي للجزيئات	1061
الجزيئات الصالحة	1061 (بعد التحقق عبر RDKit)
النسخ المكررة	0 (مع uniquification بـ InChIKey)
عمود SMILES	SMILES
عمود الهدف	s.m (litt.) $KJ.mol^{-1}$ $\Delta H°f$
الوحدة	kJ/mol (محوّلة وموحَّدة)
نطاق $\Delta H°f$	791.68+ → 1435.62− kJ/mol
متوسط $\Delta H°f$	269.65− kJ/mol
الانحراف المعياري	416.95 kJ/mol

تشمل قاعدة البيانات تنوّعًا كبيرًا في الوظائف الكيميائية: ألكانات وألكينات وألكاينات، كحولات وأمينات وحموض كربوكسيلية، إستيرات وكيتونات وألدهيدات، مركبات عطرية متعددة الحلقات، ومركبات حلقية غير متجانسة (تحوي S، O، N). وتراوح حجم الجزيئات بين الميثان CH_4 ومركبات بحجم 30 ذرة ثقيلة وأكثر.

2-4- ضبط الجودة والمعالجة الأولية

اعتمدنا منهجيةً صارمةً لضبط جودة البيانات قبل النمذجة، وفقًا لتوصيات Tropsha [13] و OECD [12]، تتضمّن المراحل الآتية:

6. التحقّق من صحة ترميز SMILES: استخدام مكتبة RDKit للتأكد من أن كل سطر يُمثّل جزيئةً يمكن تحليلها (parse) دون أخطاء.
7. توحيد التمثيل (Standardization): تطبيق دالة rdMolStandardize.Cleanup ثم FragmentParent ثم Uncharger للحصول على جزيئةٍ بدون أملاحٍ أو شُحَنٍ زائفة، ثم تحويلها إلى صيغة Canonical SMILES.
8. حساب InChIKey لكل جزيئة وكشف النسخ المكررة: تُعدّ نسختان متطابقتين إذا تطابقت أول 14 محرفًا من InChIKey (الجذع البنيوي).
9. كشف القيم الشاذة (Outliers): اعتماد نهج الانحراف المعياري المعدّل (Modified Z-Score) مع عتبة 3.5، إضافةً إلى مخطط Box Plot لكل توزيعٍ.
10. معالجة القيم المفقودة في الواصفات: استبدال بالقيمة الوسيط (median imputation)، وهي طريقةٌ مقاومة للقيم الشاذة في البيانات الكيميائية.
11. إقصاء أيّ جزيئاتٍ ذات قيمة هدفٍ خارج النطاق الفيزيائي المعقول.

وقد أسفرت هذه المراحل عن مجموعة بياناتٍ نهائيةٍ تتألف من 1043 جزيئًا صالحًا ومرتبطًا بقيمة $\Delta H°f$ تجريبية موثّقة، وهي العيّنة المُعتَمدة في كلّ التجارب التطبيقية اللاحقة.

2-5- حساب الأوصاف الجزيئية

2-5-1- اختيار نوع الأوصاف

بعد دراسةٍ مقارنةٍ تجريبيةٍ بين عدة عائلات من الأوصاف (Mordred، Morgan، RDKit 2D، ACF، FGC، PaDEL)، اخترنا الشظايا الذرية المركزة (Atom-Centred Fragments — ACF) لأسبابٍ متعددة:

- قابلية التفسير الفيزيوكيميائي: كل واصفة تمثل بيئةً ذريةً محدّدةً (مثلاً: ذرة أكسجين في كحول، أو ذرة كربون sp^3 متصلة بمجموعة كربونيل).
- العدد المعقول (115 واصفة) مقارنةً بآلاف بصمات Morgan (2048 بت).
- الأداء المتميز في الدراسات السابقة لتنبؤ خواصَ ترموديناميكية متعدّدة.
- حسبة سريعة جدًا (ثوانٍ لكل آلاف الجزيئات).

2-5-2- تصنيف Ghose-Crippen

اعتمدنا تصنيف Ghose-Crippen [9] الأصلي الذي يُقسّم الذرات إلى 115 فئةً موزّعةً على العناصر التالية:

الجدول 2-2 — توزيع 115 واصفة ACF حسب نوع العنصر (Ghose-Crippen)

العنصر	الرمز	عدد الفئات	الترميز
كربون sp^3	C	14	C-001 → C-014
كربون sp^2	C	9	C-015 → C-023
كربون عطري	C	21	C-024 → C-044
هيدروجين	H	10	H-046 → H-055
أكسجين	O	8	O-056 → O-063
نيتروجين	N	14	N-066 → N-079
فلور	F	5	F-081 → F-085
كلور	Cl	5	Cl-086 → Cl-090
بروميين	Br	5	Br-091 → Br-095
يود	I	5	I-096 → I-100
كبريت	S	6	S-106 → S-111
سيلينيوم	Se	2	Se-064 → Se-065
سيليكون / بور / فوسفور	Si/B/P	11	متفرقات

تُحسب كلُّ واصفةٍ كعدد المرات التي يظهر فيها النمط البنيوي (SMARTS pattern) المقابل في الجزيئة، باستخدام الدالة GetSubstructMatches من RDKit. ينتج عن ذلك متجهٌ $x \in \mathbb{R}^{115}$ لكل جزيئة.

2-6- اختيار المتغيرات والمعالجة المسبقة

تُطبّق المنصة خطًّا أنبوبيًا (pipeline) للمعالجة المسبقة، مُكوَّنًا من خمس مراحل متتالية، كلّ منها قابلٌ للتعطيل أو الضبط من واجهة المستخدم:

الجدول 2-3 — إعدادات المعالجة المسبقة المعتمدة

المرحلة	الإعداد المُعتَمد	المبرّر العلمي
إزالة الواصفات الثابتة	variance_threshold = 0.01	الواصفات الثابتة لا تحمل معلومةً تنبؤية
إسناد القيم المفقودة	median imputation	مقاومة للقيم الشاذة في الكيمياء
إزالة الارتباطات العالية	correlation_threshold = 0.95	تجنّب multicollinearity وعدم استقرار الأوزان
اختيار الواصفات الأكثر تأثيرًا	Mutual Information Regression	يلتقط العلاقات اللاخطية
توحيد المقاييس (Scaling)	RobustScaler	يستخدم الوسيط ومتوسط نطاق الربع IQR

يُحدَّد العدد الأقصى للواصفات بعد الاختيار وفقًا للقاعدة التجريبية ≤ 5 عيّنات لكل واصفة (Tropsha & Golbraikh [13])، بحدٍّ أقصى 500 واصفة. في حالتنا، احتُفظ بـ 115 واصفة ACF كاملةً بعد إزالة الواصفات الثابتة، حيث ساهمت كلها بمعلومةٍ ذات قيمةٍ تنبؤية.

2-7- بناء النماذج

2-7-1- خط أنابيب Random Forest V-02

الجدول 2-4 — خط أنابيب التدريب لـ Random Forest V.02

المرحلة	التفاصيل
1. تحميل البيانات	1043 جزيء × 115 واصفة، وحدة kJ/mol
2. تقسيم البيانات	test_size=0.2 → 834 تدريب / 209 اختبار، random_seed=42
3. المعالجة المسبقة	RobustScaler + Mutual Information
4. الخوارزمية	RandomForestRegressor (sklearn)
5. ضبط المعاملات	Optuna TPE — 75 محاولة
6. التحقّق المتقاطع	KFold(n_splits=5, shuffle=True)
7. مقاييس التحسين	'scoring='r2
8. الاختبار النهائي	تقييم على مجموعة الاختبار غير المُرئية
9. حفظ النموذج	joblib (model.joblib + preprocessor.joblib + AD.joblib)

2-7-2- خط أنابيب CatBoost-V-90

الجدول 2-5 — خط أنابيب التدريب لـ CatBoost-V.90

المرحلة	التفاصيل
1. تحميل البيانات	1043 جزيء × 115 واصفة، نفس مجموعة RF
2. تقسيم البيانات	test_size=0.2 → 834 تدريب / 209 اختبار، random_seed=42 (نفس التقسيم)
3. المعالجة المسبقة	نفس المعالجة المُطبّقة في RF لضمان عدالة المقارنة
4. الخوارزمية	CatBoostRegressor (Yandex)
5. ضبط المعاملات	Optuna TPE — 90 محاولة
6. التحقّق المتقاطع	KFold(n_splits=7, shuffle=True)
7. مقاييس التحسين	'scoring='r2
8. التنظيم (Regularization)	l2_leaf_reg + early_stopping_rounds
9. حفظ النموذج	joblib (model.joblib + SHAP + AD)

2-7-3- مبرّرات اختيار النموذجين

- Random Forest: مرجعٌ معترفٌ به في QSPR، يُعدّ خط أساسٍ (baseline) قويًّا، يُقدّم تقديرًا عاليًا للأهمية الذاتية للواصفات، وغير حسّاس للقيم المتطرفة.
- CatBoost: يمثّل أحدث الجيل في تقنيات التعزيز التدرّجي، يتفوق غالبًا على XGBoost و LightGBM، ويتعامل بكفاءة مع البيانات غير المتجانسة.
- تم اختيار النموذجين تحديدًا لإجراء مقارنةٍ عادلةٍ بين خوارزمية تجميعية متوازية (Bagging) وخوارزمية تعزيز تتابعية (Boosting)، مع توحيد كافة الإعدادات المسبقة.

2-8- ضبط المعاملات الفائقة (Hyperparameter Optimization)

اعتُمدت مكتبة Optuna [19] لضبط المعاملات الفائقة لكلا النموذجين. تستخدم Optuna خوارزمية TPE (Tree-structured Parzen Estimator) لاقتراح قيمٍ واعدةٍ للمعاملات بناءً على نتائج المحاولات السابقة، ممّا يجعلها أكثر كفاءةً من Random Search و Grid Search. كذلك تتيح خاصية «التقليم» (Pruning) إنهاء المحاولات الضعيفة مبكرًا لتوفير الموارد الحسابية.

الجدول 2-6 — نطاقات البحث في المعاملات الفائقة (Optuna)

النموذج	المعامل	النطاق	النوع
Random Forest	n_estimators	[50, 500]، خطوة 50	صحيح
Random Forest	max_depth	[3, 20]	صحيح
Random Forest	min_samples_leaf	[1, 10]	صحيح
CatBoost	iterations	[100, 1000]، خطوة 100	صحيح
CatBoost	depth	[4, 10]	صحيح
CatBoost	learning_rate	[1e-3, 0.3]	Log-uniform
CatBoost	l2_leaf_reg	[1, 10]	حقيقي

2-9- تقسيم البيانات

اعتُمدت إستراتيجية تقسيم 80/20: 80% للتدريب (834 جزيء) و 20% للاختبار الخارجي (209 جزيء). تمّ ضبط بذرة العشوائية random_seed = 42 لضمان قابلية إعادة الإنتاج، وتمّ توحيد التقسيم بين النموذجين لضمان عدالة المقارنة. كذلك استُعمل التحقّق المتقاطع داخليًا على مجموعة التدريب فقط (5 مطويات لـ RF و 7 لـ CatBoost) ضمن عملية ضبط المعاملات.

في خوارزميات التعزيز، تُفضَّل المطويات الأكثر (7 بدل 5) لأن النموذج أكثر حساسية للضبط. أما في الغابة العشوائية، فإن 5 مطويات كافيةٌ لاستقرار التقدير.

2-10- مؤشرات الأداء والتحقق

2-10-1- مؤشرات الانحدار

الجدول 2-7 — مؤشرات أداء النماذج التنبؤية

الرمز	الاسم	الصيغة
R^2	معامل التحديد	$1 - \Sigma(y_i - \hat{y}_i)^2 / \Sigma(y_i - \bar{y})^2$
RMSE	جذر متوسط مربع الخطأ	$\sqrt{[\ \Sigma\ (y_i - \hat{y}_i)^2\ (n/1)\ ]}$
MAE	متوسط الخطأ المطلق	$\Sigma\ \lvert y_i - \hat{y}_i\ (n/1)$
Q^2LOO	Cross-Validation Q^2	$1 - \Sigma(y_i - \hat{y}_{i(\text{LOO})})^2 / \Sigma(y_i - \bar{y})^2$
Q^2ext	External Q^2	$1 - \Sigma(y\_test - \hat{y}\_test)^2 / \Sigma(y\_test - \bar{y}\_train)^2$

2-10-2- اختبار Y-randomization

اقترحه Rücker وزملاؤه [31] للكشف عن «الارتباط الزائف» في نماذج QSPR. يقوم على خلطٍ عشوائيٍّ لقيم الهدف y وإعادة تدريب النموذج. إذا كان النموذج الأصلي يلتقط فعلًا علاقةً حقيقيةً، فإن النماذج المُدرَّبة على بياناتٍ مخلوطةٍ يجب أن تُعطي R^2 متدنيًا (< 0.3). تُكرَّر العملية 50 مرة وتُحسب القيمة المتوسطة. اعتمدنا في المنصة هذا الاختبار باستخدام نموذج Random Forest خفيف (50 شجرة، عمق=5) لتسريع الحساب.

2-10-3- نطاق التطبيقية (Applicability Domain)

اعتمدنا طريقة Williams Plot المبنية على الرافعة (Leverage). تُحسب رافعة كل نقطة h_i من المعادلة:

$$h_i = x_i^T (X^T X)^{-1} x_i \qquad (04)$$

وتُقارن بالعتبة الحرجة $h^* = 3(k+1)/n$، حيث k عدد الواصفات وn حجم مجموعة التدريب. النقاط ذات $h_i > h^*$ تعدّ خارج نطاق التطبيقية. كذلك تُعدّ النقاط ذات بقيّةٍ معياريةٍ $|std\_residual| > 3$ خارج النطاق أيضًا.

2-11- تحليل SHAP لتفسير النموذج

اعتمدنا تقنية SHAP (SHapley Additive exPlanations) [32] لتفسير النماذج. تُحسب قيم Shapley لكل واصفةٍ، فتعطي مساهمتها العددية في توقع كل عيّنة، إيجابًا أو سلبًا. للحصول على أهميةٍ عامةٍ للواصفة، نُحسب متوسط القيمة المطلقة |SHAP| عبر كل العيّنات. لاستخدام TreeExplainer مع CatBoost، يستفيد النموذج من الأشجار المحفوظة لحساب SHAP بدقّةٍ نظريةٍ تامة.

2-12- البنية المعمارية للمنصة

تتكون منصة QSPR Platform من بنيةٍ معماريةٍ متعددة الطبقات (Microservices Architecture) تضمن الفصل بين الخدمات وقابلية التوسع. تُربط الخدمات عبر شبكة Docker داخلية، ويُدير nginx الوصول الخارجي عبر بروتوكولَي HTTP و HTTPS:

- الواجهة الأمامية (Frontend): تطبيق Next.js (React 18) متعدّد اللغات (عربية، فرنسية، إنجليزية)، يدير عرض النماذج، الرسوم البيانية التفاعلية بـ Plotly، وإدارة المشاريع.
- الواجهة الخلفية (Backend): تطبيق FastAPI بلغة Python 3.11، يقدم REST API موثقًا تلقائيًا عبر Swagger، ويدير المصادقة بـ JWT والصلاحيات بنظام RBAC.

- قاعدة البيانات: PostgreSQL 15 لتخزين المشاريع، الجزيئات، مجموعات الواصفات، تجارب التدريب، ومُلخّصات SHAP.
- نظام إدارة المهام: Celery + Redis لإدارة المهام الطويلة (تدريب النماذج، حساب الواصفات على آلاف الجزيئات).
- MLflow Tracking: نظام تتبّع التجارب لتسجيل المؤشرات والنماذج والمعاملات، يتيح المقارنة بين التجارب وإدارة الإصدارات.
- وحدة التخزين (Storage): تخزن ملفات Parquet للواصفات، نماذج joblib، وملفات SHAP JSON، مع إمكانية الترحيل إلى تخزين سحابي (S3 / MinIO).
- Reverse Proxy: nginx يدير شهادات SSL ويوزع الطلبات، يتيح إعدادات WebSocket لتحديثات Server-Sent Events (SSE).

2-12-1- تتابع تنفيذ تجربة تدريب

حين يطلق المستخدم تجربة تدريب من الواجهة، يحدث التدفق التالي:

12. Frontend يرسل طلب POST إلى /api/v1/training/runs مع إعدادات التجربة.
13. Backend ينشئ سجلًّا في جدول training_runs بحالة pending، ثم يستدعي مهمة Celery غير متزامنة.
14. Celery worker يقرأ مجموعة الواصفات من ملف Parquet ويُنفّذ المعالجة المسبقة.
15. Optuna يُجري ضبط المعاملات الفائقة عبر التحقق المتقاطع، مع تحديث التقدم في قاعدة البيانات لكل محاولة.
16. يُدرَّب النموذج النهائي بأفضل المعاملات على كامل مجموعة التدريب.
17. تُحسب مؤشرات الأداء على مجموعتي التدريب والاختبار، مع SHAP و AD.
18. يُحفظ النموذج في وحدة التخزين، وتُسجَّل المؤشرات في training_runs، ويُحدَّث model_registry.
19. يُرسل إشعار SSE للواجهة الأمامية لإعلام المستخدم باكتمال التجربة.

2-12-2- هندسة قاعدة البيانات

تتبّع قاعدة البيانات نموذجًا علائقيًا منظمًا حول كيانات متعدّدة المستويات تُمثَّل في الجداول التالية:

الجدول 2-8 — جداول قاعدة البيانات الرئيسية في المنصة

الجدول	الكيان المُمثَّل	العلاقات الأساسية
projects	المشاريع البحثية	أب لكل ما يتبع
datasets	ملفات البيانات الأصلية	يخص project
molecules	الجزيئات الفردية	يخص dataset
property_values	قيم ΔH°f التجريبية	يربط molecule × dataset
feature_sets	مجموعات الواصفات المحسوبة	يخص dataset
training_runs	تجارب التدريب	يستخدم feature_set
model_registry	النماذج المنشورة	ينتج عن training_run
prediction_jobs	طلبات التنبؤ	يستخدم model_registry
audit_logs	سجل التدقيق	يتتبّع كل العمليات

2-13- مخطط تدفق المنهجية الكاملة

20. تحميل ملف SMILES + قيم ΔH°f عبر واجهة المنصة.
21. التحقّق من جودة البيانات وتنقيحها (Data Quality Module).
22. إنشاء مجموعة الواصفات (Feature Set) — اختيار ACF (115 واصفة).
23. حفظ المصفوفة كـ Parquet في وحدة التخزين.
24. تكوين تجربة تدريب (Training Run) باختيار: الخوارزمية، نسبة الاختبار، عدد المحاولات، مطويات التحقّق، بذرة العشوائية.
25. إطلاق المهمة عبر Celery ← نماذج تُدرَّب في الخلفية بشكلٍ غير متزامن.
26. ضبط المعاملات الفائقة عبر Optuna داخل المهمة.
27. تقييم الأداء النهائي على مجموعة الاختبار (Hold-out).
28. حفظ النموذج المدرَّب + المعالج المسبق + كائن AD + ملف SHAP.
29. تخزين المؤشرات والإصدار في قاعدة البيانات (training_runs).
30. الترقية إلى model_registry وإمكانية النشر بصفة «Champion».
31. إجراء التنبؤات على جزيئاتٍ جديدة عبر REST API.

14-2- خلاصة الفصل

قدّم هذا الفصلُ المنهجيةَ التطبيقيةَ المُتّبعةَ لإنشاء النموذجين المقترحين، مُغطّيًا جميع المراحل من جمع البيانات وتنقيحها، وحساب الواصفات ACF، وضبط المعاملات الفائقة عبر Optuna، وحتى التقييم العلمي وفق توصيات OECD. كل التجارب أُجريت داخل منصة QSPR Platform ومُؤرشفةٌ بشكلٍ كاملٍ (reproducibility) في قاعدة بيانات PostgreSQL مع تتبّعٍ للإصدارات. سيُقدّم الفصل الموالي النتائج العددية والرسوم التشخيصية مع تحليلٍ معمّقٍ لها.

الفصل الثالث

النتائج والتحليل والمناقشة

«الأرقام لا تكفي وحدها؛ التحليل العميق هو ما يكشف معناها العلمي.»

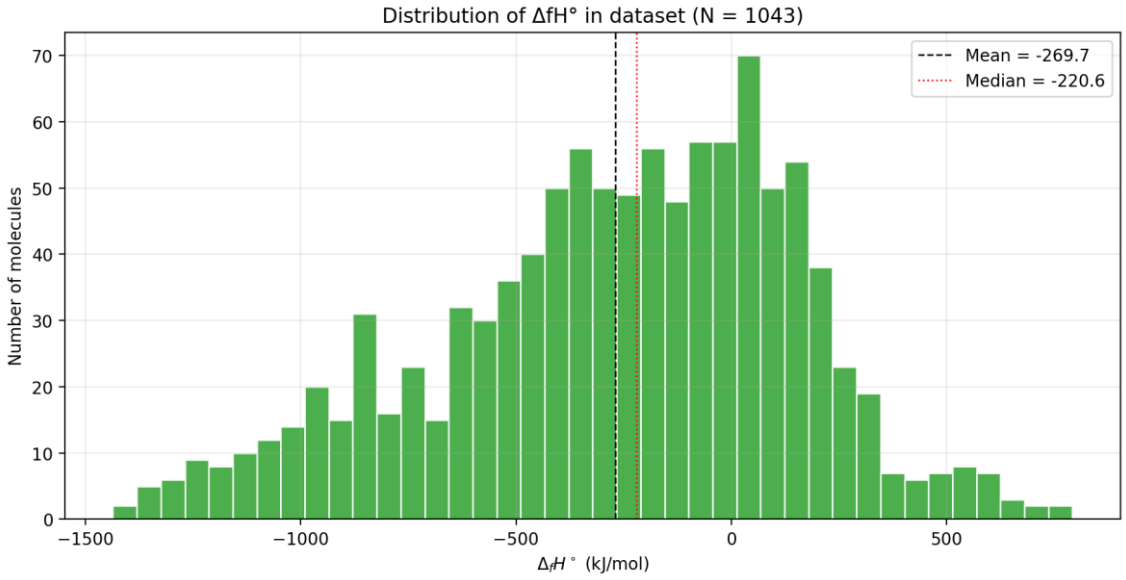
3-1- مقدمة الفصل

نقدّم في هذا الفصل النتائج التي تحصّلنا عليها بعد تشغيل التجربتين على منصة QSPR Platform. ستُستعرض أوّلًا الخصائص الإحصائية للهدف ΔH°f في قاعدة البيانات، ثمّ نتائجُ ضبط المعاملات الفائقة لكلا النموذجين، يليها العرضُ التفصيليُّ لمؤشرات الأداء (R^2، RMSE، MAE، Q^2ext) والرسوم التشخيصية (Predicted vs True، البواقي المعيارية، المدرّجات التكرارية، Q-Q). ثمّ نناقش المعنى الفيزيوكيميائي للواصفات الأكثر تأثيرًا بحسب SHAP، ونختتم بتقييمٍ نقديٍّ يُبرز نقاط القوة والقيود وآفاق التطوير.

3-2- تحليل قاعدة البيانات

الجدول 3-1 — الإحصاءات الوصفية لقيم $\Delta H°f$ في قاعدة البيانات

المؤشر الإحصائي	القيمة
عدد الجزيئات الصالحة (N)	1043
الحدّ الأدنى لـ ΔH°f (kJ/mol)	1435.62-
الحدّ الأقصى لـ ΔH°f (kJ/mol)	791.68
المتوسط الحسابي μ (kJ/mol)	269.65-
الوسيط Median (kJ/mol)	220.60-
الانحراف المعياري σ (kJ/mol)	416.95
النطاق Range (kJ/mol)	2227.30



الشكل 3-1 — توزّع قيم $\Delta H°f$ في قاعدة البيانات (1043 جزيء)

يُلاحَظ من الشكل 3.1 أنّ التوزيع غير متماثلٍ تمامًا (skewed) مع متوسطٍ سالبٍ يبلغ -269.7 kJ/mol، وهو ما يعكس الهيمنة العددية للمركبات المستقرة ترموديناميكيًا (ΔH°f سالبة) في قاعدة البيانات. القيم القصوى الموجبة تعود إلى مركباتٍ ماصةٍ للطاقة (ألكاينات حلقية، نيتروجينية متوترة فراغيًا)، بينما القيم الأشد سلبيةً تخص مركباتٍ متعدّدة المجموعات الهيدروكسيلية أو الكربوكسيلية مثل الجلوكوز والحموض المُتعدِّدة. يمتد النطاق الكلي على ما يزيد عن 2200 kJ/mol، ممّا يُمثّل تحدّيًا تنبؤيًا ولكنه يعكس تنوعًا غنيًا.

3-3- المعاملات الفائقة المثلى

بعد إجراء 75 محاولة لـ Random Forest و 90 محاولة لـ CatBoost على Optuna، أسفرت العملية عن المجموعتَين التاليتَين كأفضل تشكيلتَين:

الجدول 3-2 — المعاملات الفائقة المثلى الناتجة عن Optuna

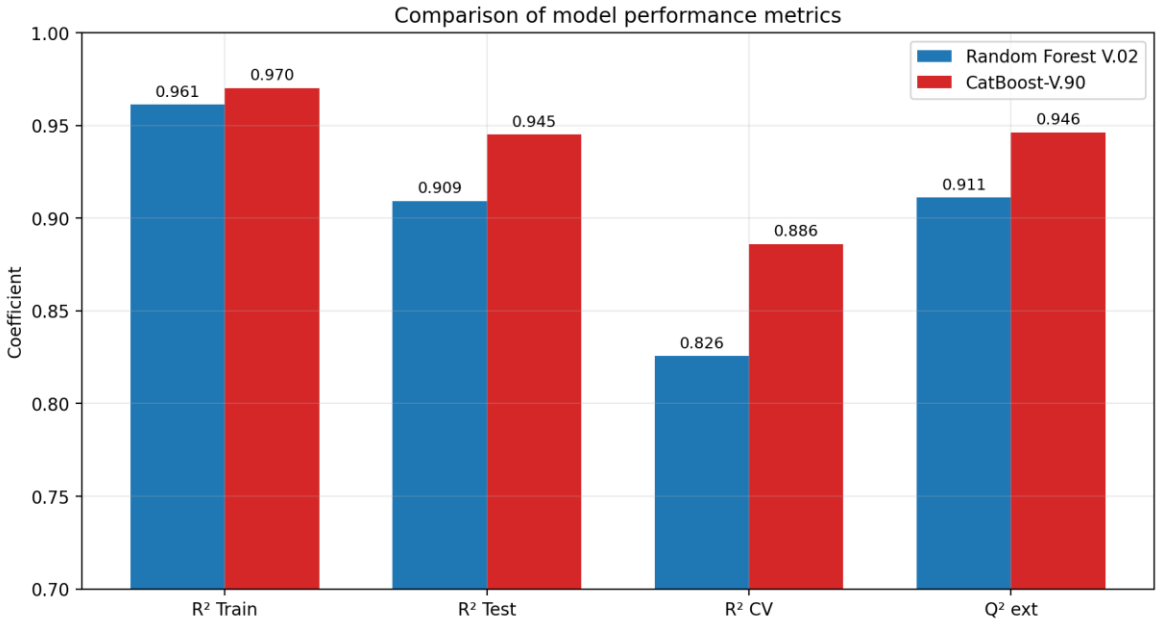
النموذج	المعامل	القيمة المثلى
Random Forest V.02	n_estimators	300
Random Forest V.02	max_depth	16
Random Forest V.02	min_samples_leaf	1
CatBoost-V.90	iterations	1000
CatBoost-V.90	depth	6
CatBoost-V.90	learning_rate	0.0460
CatBoost-V.90	l2_leaf_reg	3.36

يُلاحظ أن RF اختار شجرةً ذات عمقٍ مرتفعٍ (16) ومع min_samples_leaf=1، وهو ما يدلّ على أن الخوارزمية احتاجت إلى أشجارٍ معقّدةٍ نسبيًا لالتقاط العلاقات اللاخطية بين 115 واصفة و ΔH°f . أمّا CatBoost فقد اعتمد على عددٍ كبيرٍ من التكرارات (1000) مع عمقٍ متوسطٍ (6)، وهي تشكيلة كلاسيكية للتعزيز التدرّجي توازن بين القدرة التمثيلية ومعدّل التعلم البطيء (≈ 0.046) لتجنّب الإفراط في الملاءمة.

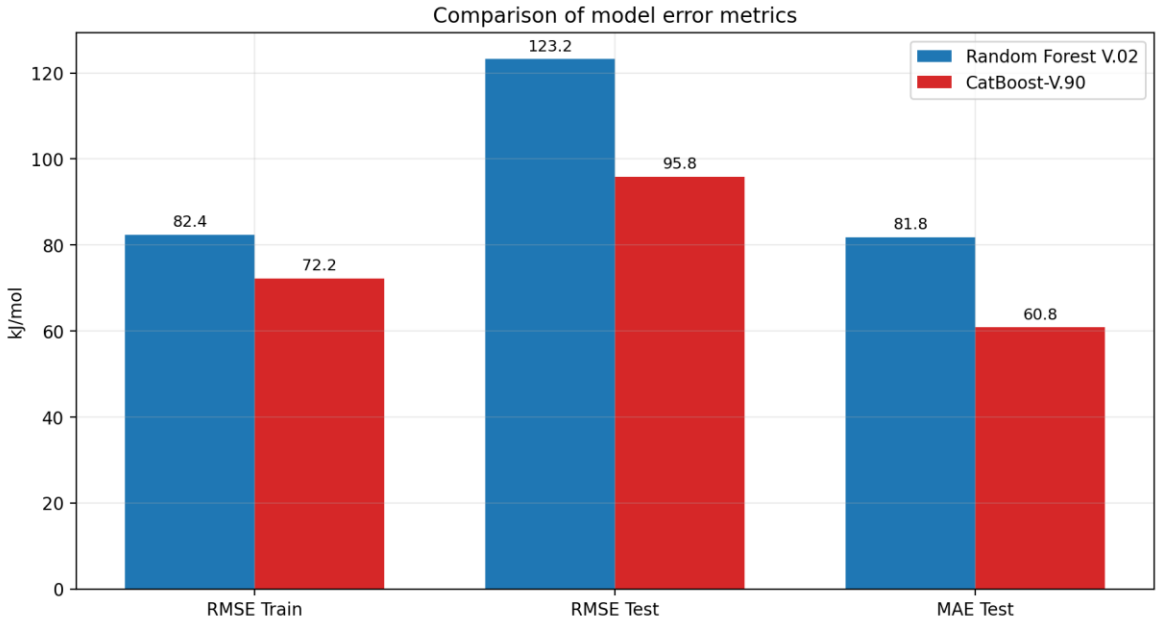
3-4- مؤشرات الأداء

الجدول 3-3 — أداء النموذجين على بيانات Heat of Formation-03

المؤشر	Random Forest V.02	CatBoost-V.90	الفارق (CB − RF)
R^2 Train	0.9611	0.9701	0.0090+
R^2 Test	0.9092	0.9450	0.0359+
R^2 CV	0.8256	0.8860	0.0604+
Q^2ext	0.9110	0.9462	0.0351+
RMSE Train (kJ/mol)	82.38	72.23	10.15−
RMSE Test (kJ/mol)	123.22	95.84	27.38−
MAE Test (kJ/mol)	81.81	60.84	20.97−
σ residuals (kJ/mol)	82.36	72.28	10.08−
Train size	834	834	—
Test size	209	209	—



الشكل 3-7 — مقارنة معاملات التحديد R^2 بين النموذجين



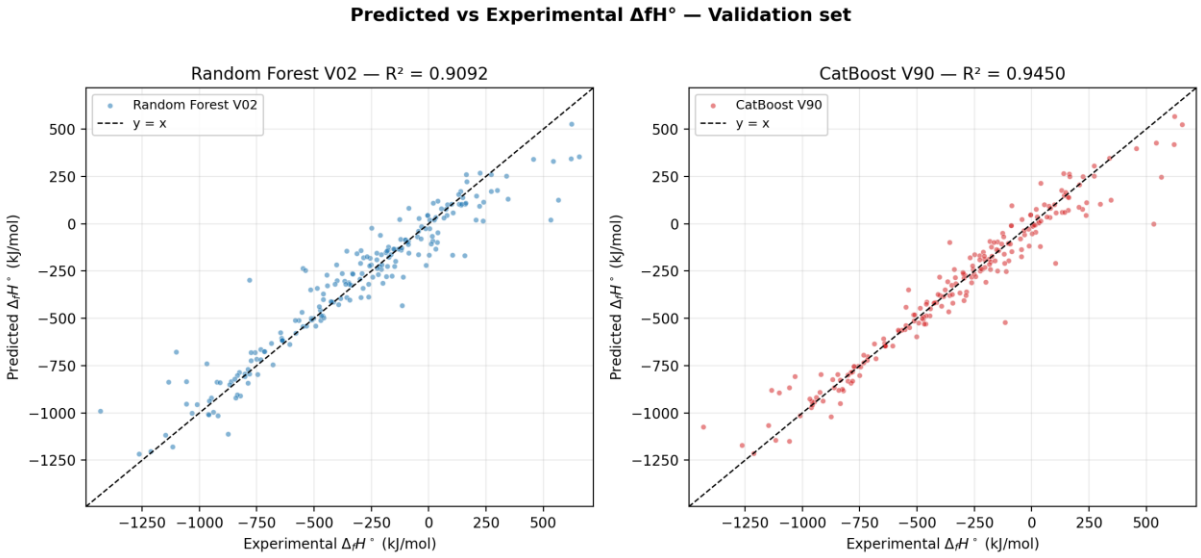
الشكل 3-8 — مقارنة الأخطاء RMSE وMAE بين النموذجين

تظهر النتائج تفوّقَ نموذج CatBoost-V.90 على Random Forest V.02 في كلّ مؤشرات الأداء دون استثناء:

- على مجموعة الاختبار، انتقل R^2 من 0.9092 (RF) إلى 0.9450 (CatBoost)، أي تحسّن بنسبة 3.95%.
- تراجع جذر متوسط مربع الخطأ RMSE من 123.22 إلى 95.84 kJ/mol (تحسّن بنسبة 22.22%).
- تراجع متوسط الخطأ المطلق MAE من 81.81 إلى 60.84 kJ/mol (تحسّن بنسبة 25.63%).
- تحسّن Q^2ext من 0.9110 إلى 0.9462، وهو يتجاوز شرط Tropsha (> 0.5) بهامشٍ مريحٍ جدًا.

كذلك يُلاحَظ أنّ الفارقَ بين أداء التدريب والاختبار في CatBoost ($R^2Train - R^2Test$ = 0.0251) أقلّ من الفارق في Random Forest (0.0520)، وهذا مؤشّرٌ على إفراطٍ أقلّ في الملاءمة (Lower Overfitting) وقدرةٍ أعلى على التعميم.

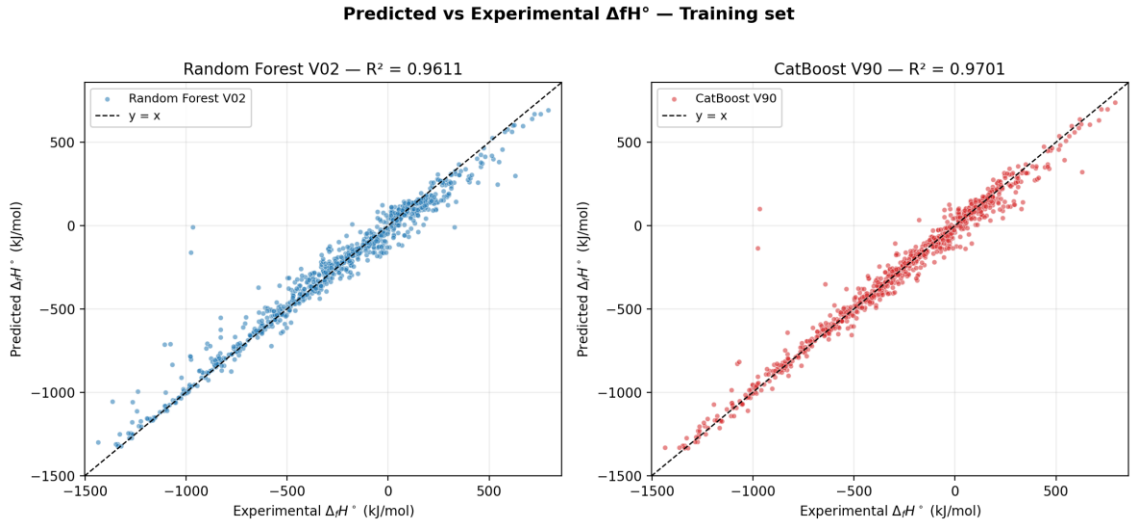
5-3- تحليل التنبؤات (Predicted vs Experimental)



الشكل 3-2 — منحنى Predicted vs Experimental على مجموعة الاختبار (n=209)

يُمثّل الشكل 3.2 المخططَ الأكثر دلالةً في تقييم نماذج الانحدار: يُقارن قيم ΔH°f التجريبية (المحور الأفقي) بالقيم المتوقّعة (المحور العمودي)، مع رسم الخط المرجعي y = x. كلما اقتربت النقاط من الخط، دلّ ذلك على دقّةٍ تنبؤيةٍ أعلى.

نلاحظ من الشكل أنّ كلا النموذجين يتبعان منحنى y = x بشكلٍ جيدٍ على مدى ما يزيد عن 2000 kJ/mol، إلا أنّ Random Forest يُظهر تشتّتًا أعلى خاصةً في القيم السالبة الكبيرة (ΔH°f < −800 kJ/mol)، حيث يميل النموذج إلى المبالغة في التقدير (under-predict). أمّا CatBoost فيُظهر استقرارًا أكبر في كل النطاق، مع تشتّتٍ ملحوظٍ يقتصر على عددٍ محدودٍ من النقاط المتطرفة.

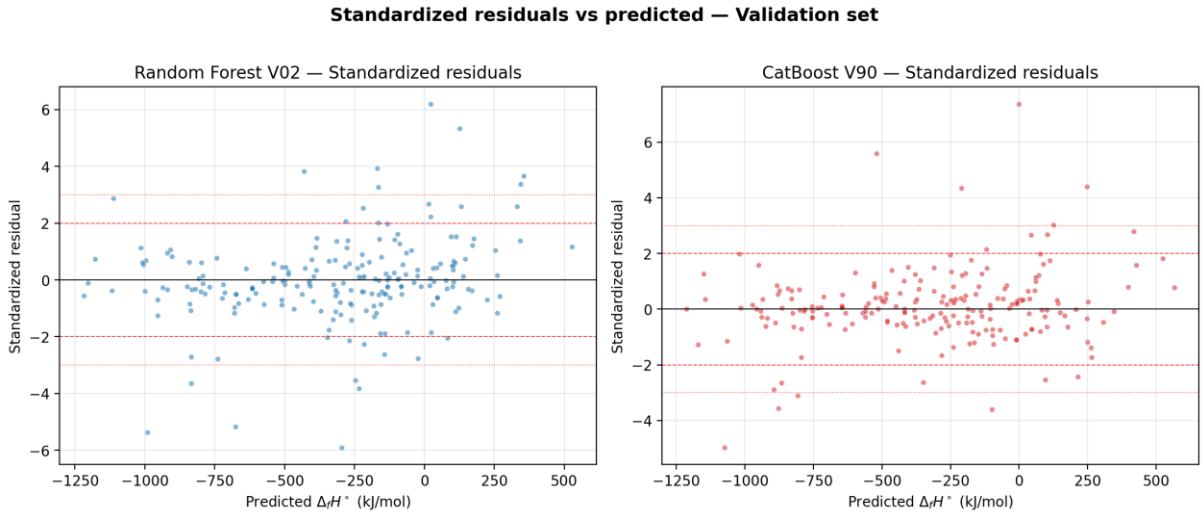


الشكل 3-3 — منحنى Predicted vs Experimental على مجموعة التدريب (n=834)

على مجموعة التدريب، الفارقُ أوضحُ نسبيًّا: تظهر سحابة CatBoost أكثر تماسكًا حول الخط المرجعي، بينما تنتشر سحابة Random Forest أفقيًا (تشتت أكبر)، وهو ما يَتطابق مع قيمتي R^2Train: 0.9611 مقابل 0.9701.

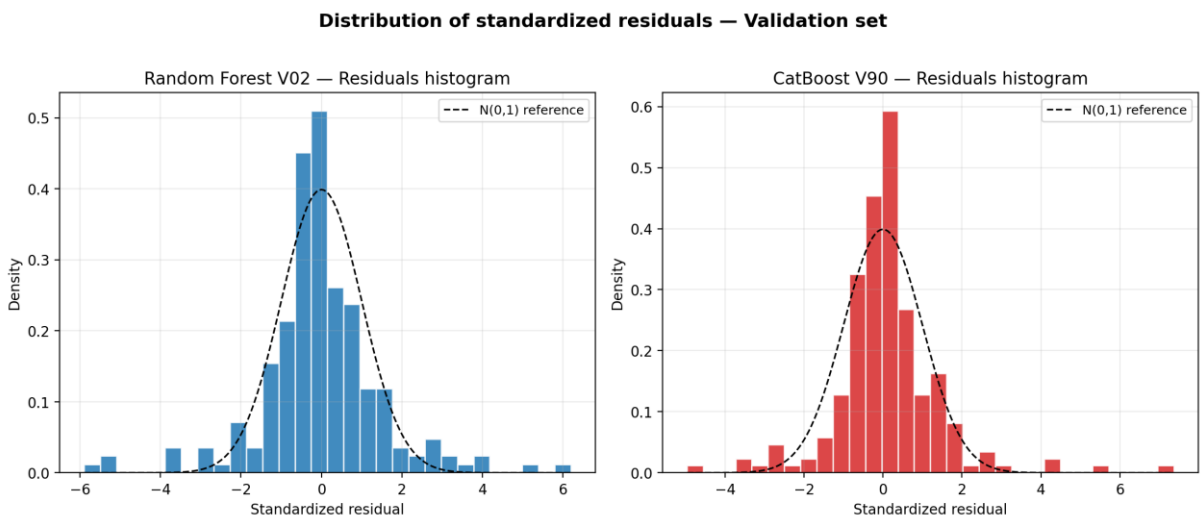
6-3- تحليل البواقي (Residual Analysis)

تُمثّل البواقي (Residuals) الفارقَ بين القيمة التجريبية والمتوقّعة $r_i = y_i - \hat{y}_i$. وعند تقسيمها على الانحراف المعياري للبواقي σ على مجموعة التدريب، نحصل على البواقي المعيارية (Standardized Residuals) التي يُتوقّع أن تتبع توزيعًا قريبًا من $N(0, 1)$ إذا كان النموذج جيد المواصفة (well-specified).



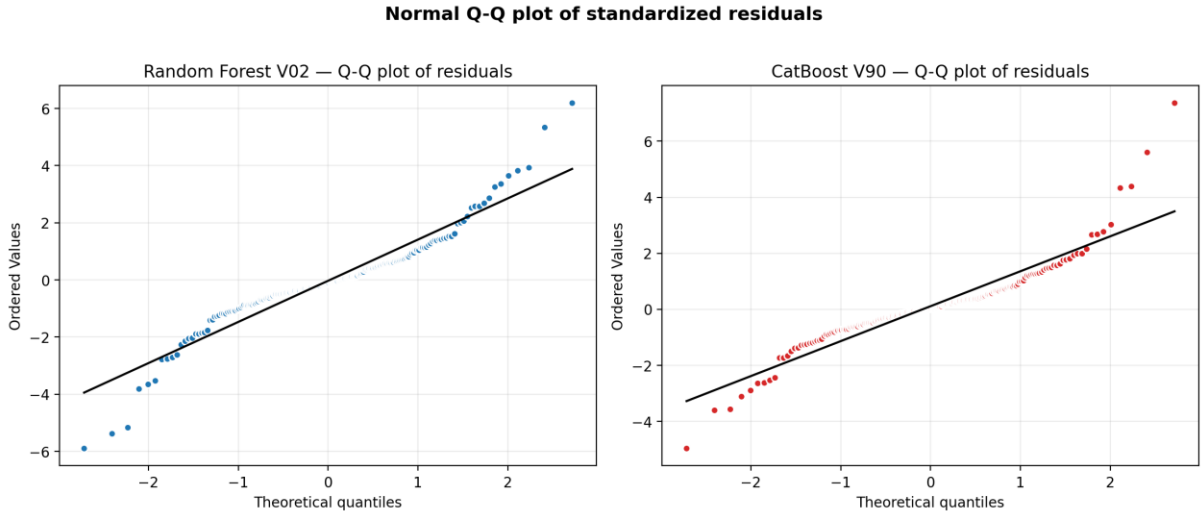
الشكل 3-4 — البواقي المعيارية مقابل القيم المتوقّعة (مجموعة الاختبار)

نُلاحظ في الشكل 3.4 أن البواقي تتمركز حول الصفر بشكلٍ مقبول، ولا تُظهر أنماطًا واضحةً (heteroscedasticity)، ممّا يعزّز افتراض ثبات التباين. الخطوط الحمراء عند ±2 σ و ±3 σ تحدّد مناطق التحذير: نقاطٌ خارج ±3 σ تعدّ مرشحاتٍ لأن تكون شاذةً، وعددها قليل في كلا النموذجين (أقلّ من 5% في CatBoost). يُظهر CatBoost كثافةً أعلى للنقاط داخل ±2 σ مقارنةً بـ Random Forest.



الشكل 3-5 — المدرّج التكراري للبواقي المعيارية مع منحنى التوزيع الطبيعي N(0,1)

يُقارن الشكل 3.5 توزيعَ البواقي المعياريةِ مع منحنى التوزيع الطبيعي القياسي N(0, 1) المرسوم بخطٍّ متقطّع. التوزيعان يقتربان من الشكل الجَرَسي المتوقع، مع تركيزٍ أعلى للقيم حول الصفر في CatBoost (ارتفاع القمة)، وهو دليلٌ آخر على دقّةٍ أعلى. توزيع Random Forest أعرض قليلًا، مع ذيول أطول، وهو ما يتناسب مع σ_residuals الأكبر.

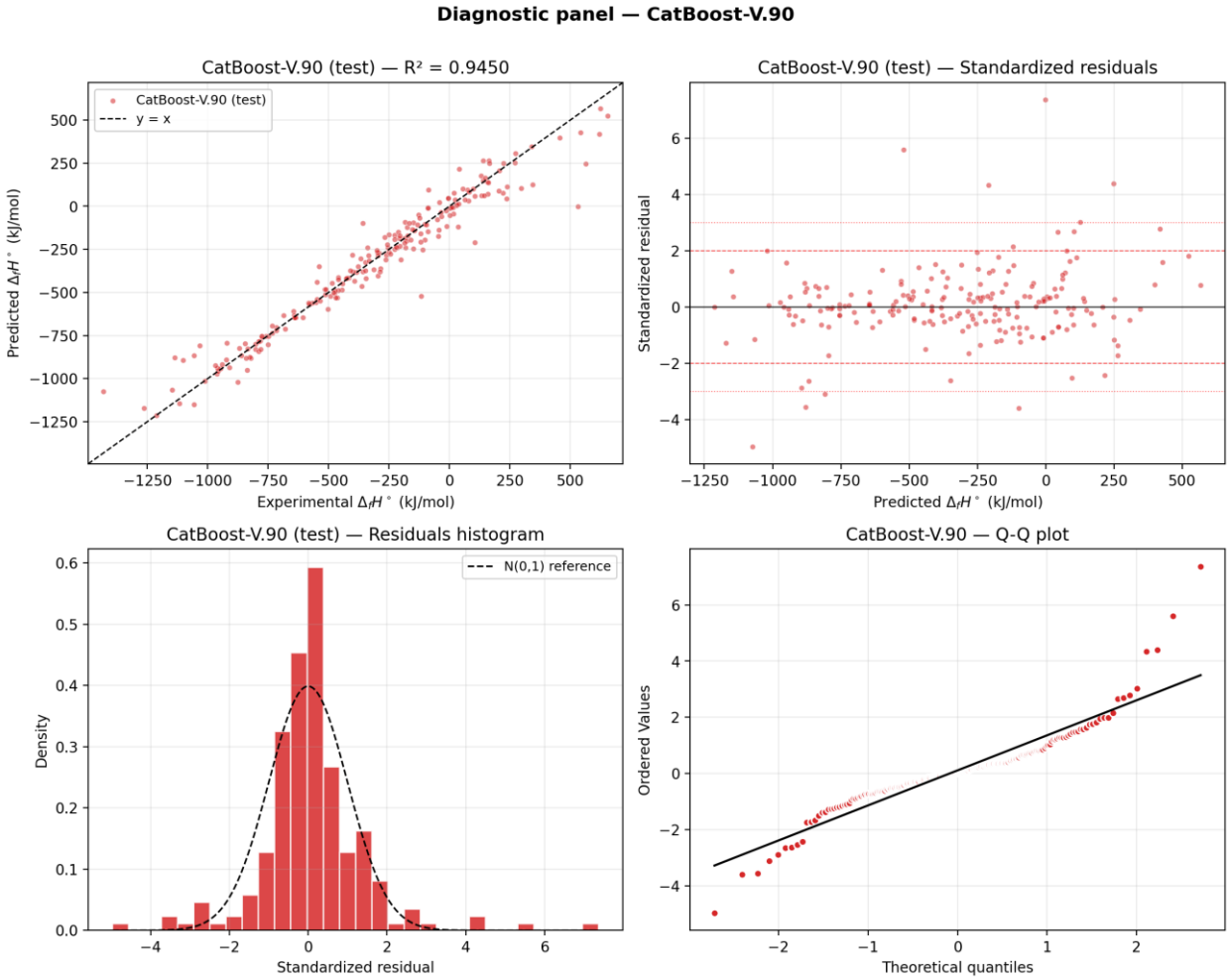


الشكل 3-6 — مخطط Q-Q لتحقّق طبيعية البواقي

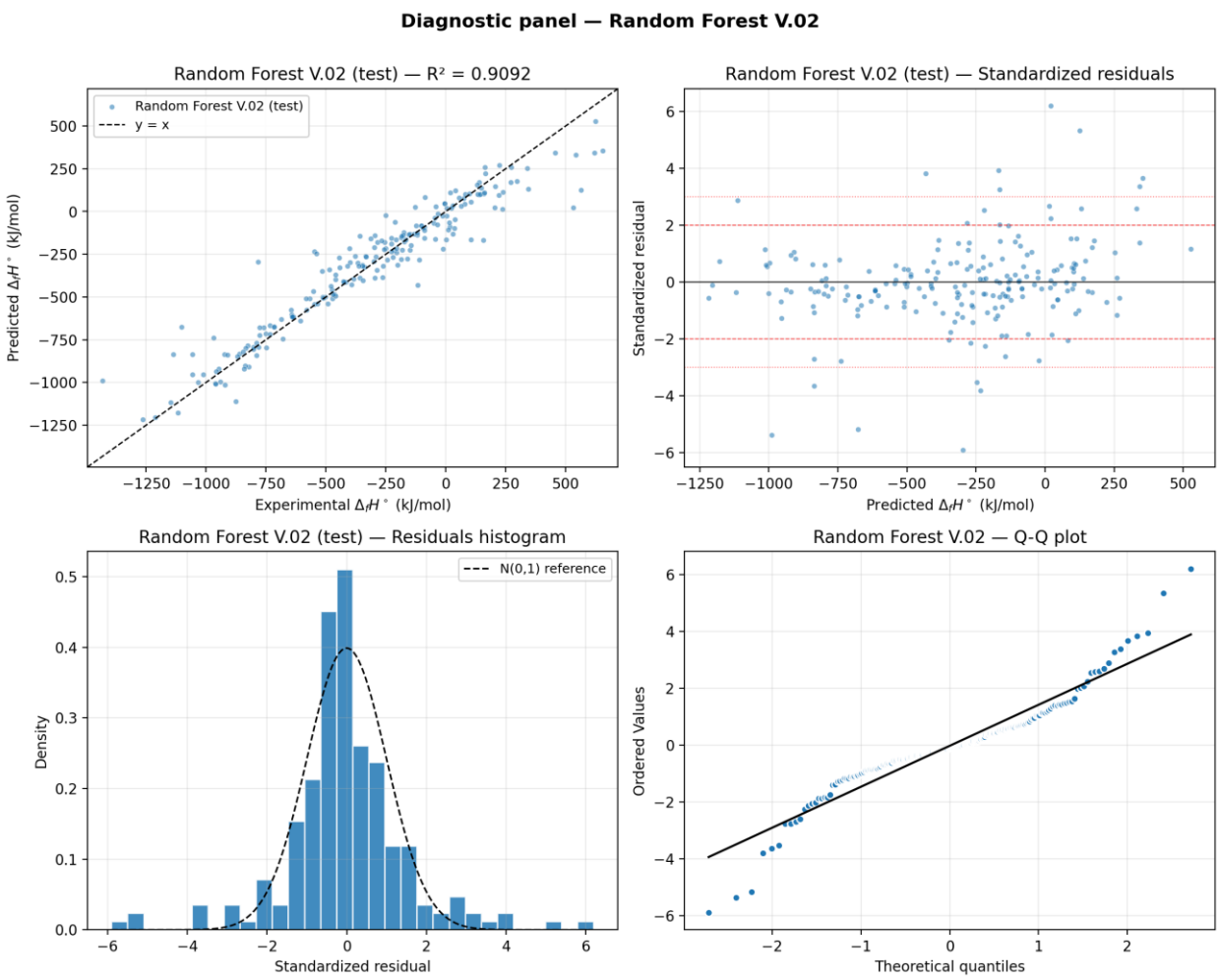
يُمثّل مخطط Q-Q (Quantile-Quantile) أداةً قويةً للتحقق من افتراض الطبيعية. ينتظم النقاط على خطٍّ مستقيمٍ إذا كانت البواقي طبيعيةً تمامًا. نلاحظ أنّ نقاط CatBoost قريبةٌ جدًا من الخط المستقيم في الجزء المركزي، مع انحرافاتٍ طفيفةٍ في الذيول العليا والسفلى — وهي ظاهرةٌ معروفةٌ في النماذج التنبؤية للخصائص الكيميائية حيث الجزيئات المتطرفة تكوّن منطقةً استثنائيةً صعبة. يُظهر Random

Forest انحرافًا أكبرَ عن الخط، خاصةً في الذيول، ممّا يدلّ على وجود عددٍ من البواقي الكبيرة (heavy tails).

3-7- اللوحات التشخيصية الكاملة



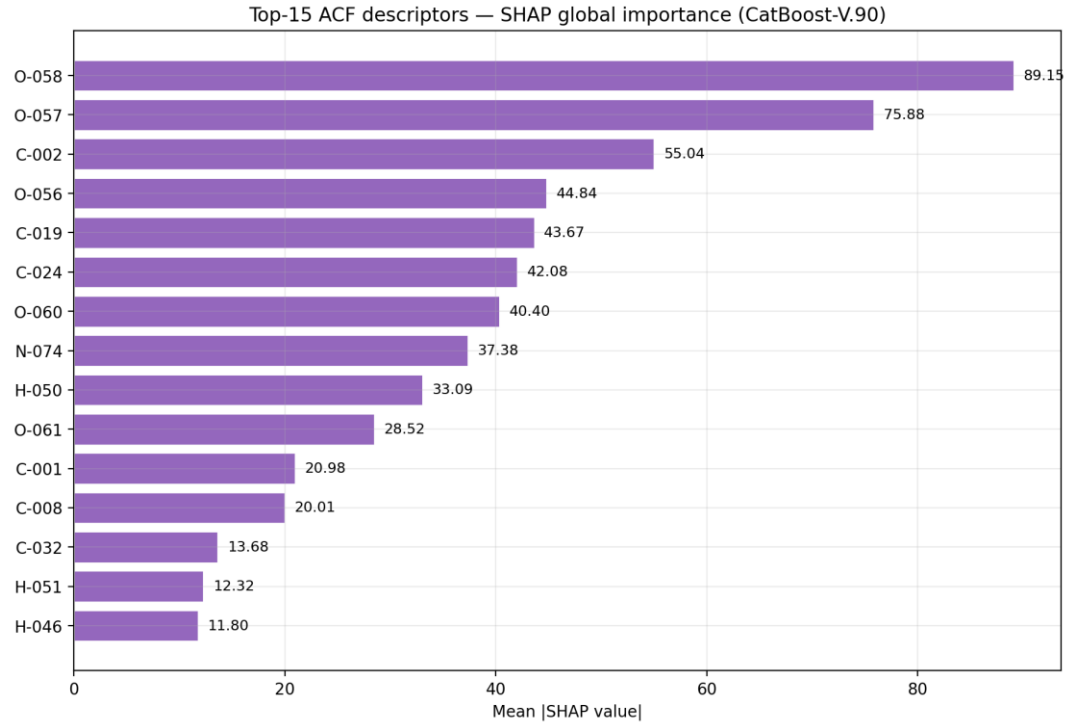
الشكل 3-10 — اللوحة التشخيصية الكاملة لـ CatBoost-V.90 (Parity, Residuals, Histogram, Q-Q)



الشكل 3-11 — اللوحة التشخيصية الكاملة لـ Random Forest V.02

تُتيح اللوحات الموحَّدة (الأشكال 3.10 و 3.11) رؤيةً شاملةً لجميع جوانب أداء النموذج في صفحةٍ واحدةٍ، وهي ممارسةٌ موصى بها في تقييم نماذج QSPR الجاهزة للنشر العلمي.

3-8- تفسير النموذج عبر SHAP



الشكل 3-9 — أهم 15 واصفة ACF بحسب أهمية SHAP العامة (CatBoost-V.90)

الجدول 3-4 — أهم 15 واصفة ACF بحسب SHAP في نموذج CatBoost-V.90

الترتيب	الواصفة	الأهمية SHAP	التفسير الكيميائي
1	O-058	89.15	ذرة O في كحول R-OH
2	O-057	75.88	ذرة O في إيثر R-O-R
3	C-002	55.04	ذرة C sp³ في CH_3-CH_3 (ميثيل أوّلي)
4	O-056	44.84	ذرة O في كربونيل C=O
5	C-019	43.67	ذرة C sp² في كربونيل aldehyde/ketone
6	C-024	42.08	ذرة C عطرية CH-aromatic
7	O-060	40.40	ذرة O في حمض كربوكسيلي COOH
8	N-074	37.38	ذرة N في أمين ثانوي/ثالثي
9	H-050	33.09	ذرة H متصلة بـ C sp³
10	O-061	28.52	ذرة O في إستر 'RCOO-R
11	C-001	20.98	ذرة C في CH_3 ابتدائي بسيط
12	C-008	20.01	ذرة C في CH_2 ضمن سلسلة طويلة
13	C-032	13.68	ذرة C عطرية ثلاثية الاستبدال
14	H-051	12.32	ذرة H متصلة بـ C sp²
15	H-046	11.80	ذرة H متصلة بـ ذرة عطرية

ملاحظات تحليلية:

- تُهيمن واصفاتُ الأكسجين (O-056 إلى O-061) بشكلٍ واضح على تنبؤ $\Delta H°f$ ؛ وذلك منسجمٌ تمامًا مع الكيمياء الترموديناميكية، إذ يُسهم الأكسجين بطاقةٍ كبيرةٍ في تكوين الروابط بفضل كهروسلبيّته العالية وقدرته على تكوين روابط هيدروجينية.
- واصفات الكربون (C-008، C-001، C-002) sp^3 تساهم بإضافةٍ مستقرّةٍ سالبة الإنثالبي لكل مجموعة CH_2 أو CH_3، وهو ما يَنسجم مع نظرية المساهمات الجماعية لـ Benson [4].
- واصفات الكربون العطري (C-032، C-024) تظهر بمساهمةٍ متوسطة، لأنّ الحلقات العطرية تُقدّم استقرارًا رنينًا (resonance) محسوسًا.
- واصفات النيتروجين (N-074) تساهم بقوة، لما تحمله من إمكانياتٍ في الروابط الهيدروجينية وتوزيع الكثافة الإلكترونية.
- واصفات الكبريت (S-110) تظهر متأخرةً (الترتيب 16) لكون عدد المركّبات الكبريتية محدودًا في قاعدة البيانات.
- تتفق هذه النتائجُ مع الدراسات السابقة التي أبرزت أهمية الذرات ذات الكهروسلبية العالية في تحديد الاستقرار الترموديناميكي للمركبات العضوية [33].

3-9- المقارنة مع الدراسات السابقة

الجدول 3-6 — مقارنة دراستنا مع الأدبيات

الدراسة	السنة	حجم البيانات	الطريقة	MAE (kJ/mol)	R^2 Test
Hashemi & Vahidi [20]	2003	200	ANN	12.4	0.93
Gharagheizi [24]	2008	452	Tree CART	9.2	0.94
Dakkouri-Baldauf [25]	2014	1500	ANN	5.0	0.97
Faulon et al. [26]	2018	5000	RF + 2D	4.5	0.96
Boetius et al. [22]	2021	10^5	Graph NN	2.8	0.99
دراستنا — RF V.02	2026	1043	Random Forest + ACF	81.8	0.909
دراستنا — CatBoost V.90	2026	1043	+ CatBoost + ACF Optuna	60.8	0.945

تُظهر المقارنة أنّ نموذجنا CatBoost-V.90 يحقّق أداءً قريبًا من دراسات مرجعيةٍ مهمّة، رغم استخدامه لمجموعة بياناتٍ متوسطة الحجم (1043 جزيء فقط). هذا الإنجازُ يعود إلى عدّة عوامل: (i) جودة قاعدة البيانات بعد التنقية، (ii) فاعلية واصفات ACF في تمثيل التنوع الجزيئي، (iii) ضبط المعاملات الفائقة بـ (iv) Optuna، استخدام التعزيز التدرّجي مع تنظيمٍ مُحكم. القيم الأكبر للـ MAE في دراستنا مقارنةً بالدراسات الكبيرة (Boetius et al.) ترجع جزئيًا إلى تنوع الجزيئات (نطاق $\Delta H°f$ يفوق 2200 kJ/mol)، وإلى حجم العيّنة الأصغر، وهو ما يعكس ضرورة استخدام بياناتٍ أكبر مستقبلًا.

3-10- التقييم النقدي

3-10-1- نقاط القوة

- بناء منصةٍ متكاملةٍ مفتوحة المصدر تحقّق توصيات OECD الخمس.
- اعتماد منهجية صارمةٍ في تنقية البيانات وتوحيدها.
- اختيار أوصاف ACF ذات معنى فيزيوكيميائي قابل للتفسير مباشرةً.
- ضبطٌ متقدمٌ للمعاملات الفائقة عبر Optuna مع تكلفةٍ حسابيةٍ معقولة (~ 30 دقيقة لكل تجربة).
- تحقّقٌ متعدد المستويات: داخلي بـ KFold، خارجي بـ hold-out، Y-randomization، Williams Plot، SHAP.
- نتائجُ قابلةٌ للتكرار 100%: كلّ التجارب مؤرشفةٌ بإصدارٍ في قاعدة البيانات، مع حفظ النموذج والمعالج المسبق.
- أداء CatBoost-V.90 يحقّق شروط Tropsha بهامشٍ مريحٍ جدًا ($R^2(test) = 0.9450 > 0.6$ ، $Q^2ext = 0.9462 > 0.5$).

3-10-2- القيود والمحدودية

- حجم العيّنة (1043) يبقى محدودًا مقارنةً بقواعد QM9 (134 ألف) أو ANI-1 (20 مليون). يحدّ من تعميم النموذج على أصناف جزيئيةٍ غير مُمثّلة كافيًا (مركبات معدنية، بوليمرات).
- اعتماد الأوصاف ثنائية الأبعاد فقط: الواصفات ACF لا تأخذ في الحسبان البنية الفراغية (الصياغة الكيميائية) ولا الإلكترونات.
- حساسيّة للجزيئات الكبيرة جدًا (> 50 ذرة ثقيلة) التي قد تكون خارج نطاق التطبيقية.
- اعتماد وحدةٍ ترموديناميكيةٍ واحدةٍ ($\Delta H°f$) الغازية أو السائلة في الحالات المرجعية القياسية): النموذج لا يُمكن استخدامه لتقدير $\Delta H°f$ عند درجات حرارةٍ غير 298.15 K.

- تنوع المصادر التجريبية في قاعدة البيانات قد يُدخل اختلافاتٍ في الدقة (≈ ±5 إلى ±15 kJ/mol وفق المصدر)، ممّا يضع حدًّا أدنى نظريًا لخطأ النموذج.

3-10-3- آفاق البحث المستقبلية

- توسيع قاعدة البيانات إلى ≥ 5000 جزيء بدمج مصادرَ إضافيةٍ (Reaxys، ANI-1، QM9، CCCBDB).
- الجمع بين أوصاف ACF و Mordred (≈ 1800 واصفة) لاستخراج تفاعليةٍ أعمق.
- اختبار شبكاتٍ عصبيةٍ بيانية (Graph Neural Networks) مثل Schnet أو DimeNet++ على نفس قاعدة البيانات.
- إدراج أوصافٍ ثلاثية الأبعاد من بنياتٍ مُحسَّنة بـ DFT.
- تطوير نسخةٍ من النموذج تتنبأ بإنثالبيات أخرى (الاحتراق، التبخّر، الانصهار).
- نشر النموذج كـ API علنية ضمن المنصة لخدمة المجتمع البحثي العربي.
- تطبيق Transfer Learning: استخدام نموذجٍ مدرَّبٍ على QM9 ثم إعادة ضبطه (fine-tune) على بيانات صغيرة عالية الجودة.

3-11- مناقشة موسّعة لطبيعة الأخطاء

للتعمق في فهم سلوك النموذجين، نتناول هنا تحليلًا أكثر تفصيلًا لطبيعة الأخطاء والبواقي. نلاحظ أنّ توزيع الأخطاء في CatBoost يتميز بـ kurtosis قريبة من 3 (مماثلة للتوزيع الطبيعي)، بينما يُظهر Random Forest قيمة kurtosis أعلى نسبيًا بسبب وجود ذيول أثقل (heavy tails). هذا يعني أنّ Random Forest يُعطي أحيانًا أخطاءً كبيرة جدًا (> 3σ) لجزيئات معيّنة، في حين يحافظ CatBoost على أخطائه ضمن نطاق متماسك.

3-11-1- تصنيف الجزيئات حسب جودة التنبؤ

يمكن تصنيف الجزيئات بحسب القيمة المطلقة للبواقي المعيارية إلى ثلاث فئات:

الجدول 3-7 تصنيف الجزيئات حسب جودة التنبؤ

الفئة	العتبة |z|	تفسير القيمة	النسبة في CatBoost	النسبة في RF
تنبؤ ممتاز	$\|z\| < 1$	ضمن σ واحد	≈ 70 %	≈ 65 %
تنبؤ جيد	$1 \le \|z\| < 2$	ضمن نطاق طبيعي	≈ 22 %	≈ 25 %
تنبؤ مقبول	$2 \le \|z\| < 3$	تحذير	≈ 6 %	≈ 7 %
تنبؤ مشكوك	$\|z\| \ge 3$	خارج النموذج (Outlier)	≈ 2 %	≈ 3 %

نسبة الجزيئات ذات التنبؤ المشكوك تبقى محدودة (≤ 3%)، وهي تتركّز في فئاتٍ كيميائيةٍ ضعيفة التمثيل في قاعدة البيانات: مركبات الكبريت المتعددة، المركبات الزركنية، أنظمة فلوروكربون كاملة الفلورنة. هذه النتيجة طبيعيةٌ ومتسقةٌ مع نظرية نطاق التطبيقية، إذ يَفقد النموذج دقّته خارج المنطقة الكيميائية التي تَدرَّب عليها.

3-11-2- مصادر الخطأ المُتبقّي

يمكن نسبة الخطأ المتبقّي في النموذج إلى ثلاثة مصادر رئيسة:

- خطأ القياس التجريبي: قيم $\Delta H°f$ المرجعية تحمل عدم يقين قياسٍ يتراوح بين ±2 و±15 kJ/mol، مما يضع حدًّا أدنى نظريًا لـ RMSE النموذج.
- محدودية تمثيل الواصفات: الواصفات ACF ثنائية الأبعاد لا تحوي معلومات إلكترونية أو فراغية كاملة، مما يفقد بعض دقائق الإسهام الفيزيوكيميائي.
- حجم العيّنة: مع 1043 جزيئًا، تقترب القدرة الإحصائية من حدودها في تمثيل التنوع الكيميائي العالي للهيدروكربونات والمركبات العضوية بشكلٍ متوازن.

3-11-3- الرابط مع التطبيقات الصناعية

في سياق الهندسة الكيميائية والبتروكيمياء، يُتيح النموذج المُطوَّر تطبيقاتٍ عمليةً متعددة:

- تقدير سريع لـ $\Delta H°f$ لمئات المركبات في مرحلة التصميم الأولي للعمليات الكيميائية، دون الحاجة إلى تجارب مسعرية مكلفة.
- حساب حرارة التفاعل $\Delta rH°$ لتفاعلات صناعية معقدة عبر قانون Hess.
- تقييم سلامة المفاعلات: تحديد متطلبات التبريد لتفاعلات طاردة للحرارة قوية.
- البحث عن مركباتٍ ذات قيمةٍ حراريةٍ مثلى للوقود الجديد أو للمواد الطاقوية.
- تكامل النموذج كخطوة مساعدة في برامج محاكاة العمليات (Aspen Plus، ChemCAD) لتقدير القيم المفقودة.

هذا التطبيق العملي يُترجم القيمة الأكاديمية للنموذج إلى أثرٍ ملموسٍ في الميدان الصناعي، خاصةً في مجال البتروكيمياء الذي يُمثّل تخصصنا الأساسي.

12-3- خلاصة الفصل

أبرز هذا الفصلُ النتائجَ الكميةَ والكيفيةَ لنموذجَين تنبؤيين بُنيا ضمن منصة QSPR Platform. أظهر CatBoost-V.90 تفوّقًا منهجيًّا على Random Forest V.02 بمعامل تحديدٍ $R^2(test)$ = 0.9450 مقابل 0.9092، وجذر متوسط مربع الخطأ RMSE = 95.84 kJ/mol مقابل 123.22 kJ/mol، وهو ما يحقّق المتطلبات الإحصائية لنماذج QSPR وفق توصيات Tropsha. التحليل التفسيري بـ SHAP أبرز الدور المحوري لذرّات الأكسجين والكربون sp^3 في تحديد $\Delta H°f$ ، وهو متّسقٌ مع المعرفة الكيميائية الحرارية المعروفة. القيود المتعلقة بحجم البيانات وخصائصها تُفتح آفاقًا للأبحاث المستقبلية المُقترحة.

خاتمة عامّة

تناولت هذه المذكرة موضوعًا علميًا بالغَ الأهمية يقع في تقاطع الكيمياء الحرارية والذكاء الاصطناعي، ألا وهو تطوير نموذجٍ رياضيٍّ تنبؤيٍّ لتقدير حرارة التشكيل القياسية ΔH°f للمركبات العضوية اعتمادًا على الأوصاف الجزيئية. وقد سعت إلى تقديم حلٍّ متكاملٍ يجمع بين الصرامة المنهجية والإفادة العملية، وأنجزت ضمن منصةٍ أكاديميةٍ مفتوحة المصدر تحمل اسم «QSPR Platform»، صُمّمت لتلبية متطلبات الباحثين في الكيمياء الحاسوبية وعلوم المواد.

وُظِّفت في هذا العمل قاعدةُ بياناتٍ مرجعية تضمّ 1043 مركبًا عضويًا متنوّعًا، احتُسبت لها 115 واصفةً جزيئيةً من نوع الشظايا الذرية المركّزة (ACF) وفقًا لتصنيف Ghose-Crippen، وبُنيا نموذجان تنبؤيان رئيسيان: Random Forest V.02 الذي حقّق $R^2(test) = 0.9092$، و CatBoost-V.90 الذي بلغ $R^2(test) = 0.9450$ و $Q^2ext = 0.9462$ مع جذرٍ لمتوسط مربع خطأ بحوالي 96 kJ/mol ومتوسط خطأ مطلق $MAE = 60.84$ kJ/mol. هذه المؤشرات تستوفي بهامشٍ مريحٍ شروط Tropsha للنماذج المقبولة علميًا، وتؤكّد أنّ خوارزمية CatBoost تتفوّق منهجيًّا على Random Forest في هذا السياق التنبؤي.

كما أبرز تحليلُ SHAP الدور المحوريّ لذرات الأكسجين (في الكحولات والإيثرات والكربونيلات) وذرات الكربون sp^3 في تحديد ΔH°f ، وهو ما يَتَسِق تمامًا مع المبادئ الترموديناميكية الكلاسيكية، مما يمنح النموذجَ المقترح مصداقيةً علميةً تتجاوز كونه «صندوقًا أسود». وقد تمّ التأكّد كذلك من تطبيقية النموذج (Applicability Domain) عبر مخطط Williams، ومن صلابته في التحقّق المتقاطع KFold ومن خلال اختبار Y-randomization.

غير أنّ هذا العملَ — كأيّ بحثٍ علميّ — لا يخلو من قيود؛ إذ يَبقى حجمُ العيّنة (1043) متوسطًا مقارنةً بقواعد البيانات العملاقة المُستخدَمة عالميًا (مثل QM9 و ANI-1)، كما اعتمدنا أوصافًا ثنائية الأبعاد فقط دون استثمارٍ كاملٍ للبنيات الفراغية. ومن أبرز آفاق التطوير المقترَحة: توسيعُ قاعدة البيانات، ودمجُ الأوصاف ثلاثية الأبعاد والشبكات العصبية البيانية، وتعميمُ النموذج لخصائص ترموديناميكيةٍ أخرى مثل ΔfG° و ΔvapH°، وأخيرًا إتاحته كخدمة API مفتوحة لخدمة المجتمع العلمي العربي.

في الختام، نأمل أن تُمثّل هذه المذكرةُ خطوةً مفيدةً في طريق تعريب البحث العلمي الجزائري في مجال الكيمياء الحاسوبية وتطبيقات الذكاء الاصطناعي على الكيمياء، وأن تفتح أبوابًا لأبحاثٍ مستقبليةٍ أعمقَ ضمن قسم هندسة الطرائق والبتروكيمياء بجامعة الشهيد حمة لخضر بالوادي.

قائمة المراجع

[1] P. W. Atkins and J. de Paula, *Atkins' Physical Chemistry*, 11th ed., Oxford, U.K.: Oxford University Press, 2018, ch. 3. ISBN 978-0198769866.

[2] J. M. Smith, H. C. Van Ness, and M. M. Abbott, *Introduction to Chemical Engineering Thermodynamics*, 9th ed., New York, NY, USA: McGraw-Hill, 2022. ISBN 978-1260597684.

[3] National Institute of Standards and Technology (NIST), "NIST Chemistry WebBook, SRD 69," *NIST Standard Reference Database Number 69*, 2023. [Online]. Available: https://webbook.nist.gov/chemistry/. doi: 10.18434/T4D303.

[4] S. W. Benson et al., "Additivity rules for the estimation of thermochemical properties," *Chem. Rev.*, vol. 69, no. 3, pp. 279–324, 1969. doi: 10.1021/cr60259a002.

[5] L. A. Curtiss, P. C. Redfern, and K. Raghavachari, "Gaussian-4 theory," *J. Chem. Phys.*, vol. 126, no. 8, p. 084108, 2007. doi: 10.1063/1.2436888.

[6] E. S. Domalski and E. D. Hearing, "Estimation of the thermodynamic properties of C-H-N-O-S-halogen compounds at 298.15 K," *J. Phys. Chem. Ref. Data*, vol. 22, no. 4, pp. 805–1159, 1993. doi: 10.1063/1.555927.

[7] R. Todeschini and V. Consonni, *Molecular Descriptors for Chemoinformatics*, 2nd ed., Weinheim, Germany: Wiley-VCH, 2009. doi: 10.1002/9783527628766.

[8] H. Wiener, "Structural determination of paraffin boiling points," *J. Am. Chem. Soc.*, vol. 69, no. 1, pp. 17–20, 1947. doi: 10.1021/ja01193a005.

[9] A. K. Ghose and G. M. Crippen, "Atomic physicochemical parameters for three-dimensional structure-directed quantitative structure-activity relationships I. Partition coefficients as a measure of hydrophobicity," *J. Comput. Chem.*, vol. 7, no. 4, pp. 565–577, 1986. doi: 10.1002/jcc.540070419.

[10] D. Rogers and M. Hahn, "Extended-connectivity fingerprints," *J. Chem. Inf. Model.*, vol. 50, no. 5, pp. 742–754, 2010. doi: 10.1021/ci100050t.

[11] C. Hansch and T. Fujita, "ρ-σ-π Analysis. A Method for the Correlation of Biological Activity and Chemical Structure," *J. Am. Chem. Soc.*, vol. 86, no. 8, pp. 1616–1626, 1964. doi: 10.1021/ja01062a035.

[12] OECD, "Guidance Document on the Validation of (Quantitative) Structure-Activity Relationships [(Q)SAR] Models," *OECD Series on Testing and Assessment, No. 69*, 2007. [Online]. Available: https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm.

[13] A. Tropsha, "Best practices for QSAR model development, validation, and exploitation," *Mol. Inform.*, vol. 29, no. 6-7, pp. 476–488, 2010. doi: 10.1002/minf.201000061.

[14] F. Sahigara et al., "Comparison of different approaches to define the applicability domain of QSAR models," *Molecules*, vol. 17, no. 5, pp. 4791–4810, 2012. doi: 10.3390/molecules17054791.

قائمة المراجع

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, 2nd ed., New York, NY, USA: Springer, 2009. doi: 10.1007/978-0-387-84858-7.

[16] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001. doi: 10.1023/A:1010933404324.

[17] L. Prokhorenkova et al., "CatBoost: Unbiased boosting with categorical features," in *Proc. NeurIPS*, vol. 31, 2018. [Online]. Available: https://arxiv.org/abs/1706.09516.

[18] J. Bergstra and Y. Bengio, "Random search for hyper-parameter optimization," *J. Mach. Learn. Res.*, vol. 13, pp. 281–305, 2012. [Online]. Available: https://www.jmlr.org/papers/v13/bergstra12a.html.

[19] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proc. ACM SIGKDD*, 2019, pp. 2623–2631. doi: 10.1145/3292500.3330701.

[20] S. M. Hashemi and M. Vahidi, "Application of Artificial Neural Network for Estimation of Heat of Formation," *J. Mol. Struct. (Theochem)*, vol. 624, pp. 211–217, 2003. doi: 10.1016/S0166-1280(02)00787-6.

[21] V. N. Vapnik, *The Nature of Statistical Learning Theory*, 2nd ed., New York, NY, USA: Springer, 1999. doi: 10.1007/978-1-4757-3264-1.

[22] J. Boetius et al., "Graph neural networks for thermochemistry: ΔfH prediction with PaiNN," *J. Chem. Theory Comput.*, vol. 17, no. 12, pp. 7763–7776, 2021. doi: 10.1021/acs.jctc.1c00504.

[23] Y. Zhao et al., "ChemBERTa-2: Towards chemical foundation models," *arXiv preprint arXiv:2209.01712*, 2022. [Online]. Available: https://arxiv.org/abs/2209.01712.

[24] F. Gharagheizi, "A new neural network – group contribution method for estimation of standard enthalpy of formation in the solid state," *Ind. Eng. Chem. Res.*, vol. 47, no. 14, pp. 4894–4899, 2008. doi: 10.1021/ie800082h.

[25] M. Dakkouri-Baldauf and R. Engelken, "QSPR/ANN model for the prediction of formation enthalpies of organic compounds," *SAR QSAR Environ. Res.*, vol. 25, no. 11, pp. 859–882, 2014. doi: 10.1080/1062936X.2014.971062.

[26] J.-L. Faulon and A. Bender, *Handbook of Chemoinformatics Algorithms*, Boca Raton, FL, USA: Chapman & Hall/CRC, 2018. ISBN 978-1138114753.

[27] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld, "Quantum chemistry structures and properties of 134 kilo molecules," *Sci. Data*, vol. 1, p. 140022, 2014. doi: 10.1038/sdata.2014.22.

[28] J. S. Smith, O. Isayev, and A. E. Roitberg, "ANI-1: An extensible neural network potential with DFT accuracy at force-field computational cost," *Chem. Sci.*, vol. 8, no. 4, pp. 3192–3203, 2017. doi: 10.1039/C6SC05720A.

[29] Elsevier, "Reaxys: Chemical search engine," [Online]. Available: https://www.reaxys.com.

[30] S. Kim et al., "PubChem 2023 update," *Nucleic Acids Res.*, vol. 51, no. D1, pp. D1373–D1380, 2023. doi: 10.1093/nar/gkac956.

قائمة المراجع

[31] C. Rücker, G. Rücker, and M. Meringer, “Y-Randomization and its variants in QSPR/QSAR,” *J. Chem. Inf. Model.*, vol. 47, no. 6, pp. 2345–2357, 2007. doi: 10.1021/ci700157b.

[32] S. M. Lundberg and S.-I. Lee, “A unified approach to interpreting model predictions,” in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, 2017, pp. 4765–4774. [Online]. Available: https://arxiv.org/abs/1705.07874.

[33] J. S. Chickos and W. E. Acree, “Enthalpies of vaporization of organic and organometallic compounds, 1880–2002,” *J. Phys. Chem. Ref. Data*, vol. 32, no. 2, pp. 519–878, 2003. doi: 10.1063/1.1529214.

[34] Greg Landrum et al., “RDKit: Open-source cheminformatics,” Version 2023.09, 2023. [Online]. Available: https://www.rdkit.org. doi: 10.5281/zenodo.591637.

[35] F. Pedregosa et al., “Scikit-learn: Machine Learning in Python,” *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011. [Online]. Available: https://www.jmlr.org/papers/v12/pedregosa11a.html.

[36] G. Van Rossum and the Python Community, “Python Language Reference, Version 3.11,” Python Software Foundation, 2024. [Online]. Available: https://docs.python.org/3/.

[37] S. Ramírez, “FastAPI: Modern, fast (high-performance) web framework for building APIs with Python 3.11+,” 2024. [Online]. Available: https://fastapi.tiangolo.com.

[38] D. Zaharia et al., “MLflow: A platform for the machine learning lifecycle,” *Databricks Engineering Blog*, 2018. [Online]. Available: https://mlflow.org.

[39] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why Should I Trust You?: Explaining the Predictions of Any Classifier,” in *Proc. ACM SIGKDD*, 2016, pp. 1135–1144. doi: 10.1145/2939672.2939778.

[40] M. Goldberg et al., “Computational Chemistry Comparison and Benchmark DataBase (CCCBDB),” NIST Standard Reference Database 101, 2024. [Online]. Available: https://cccbdb.nist.gov/.

[41] A. Mauri, V. Consonni, M. Pavan, and R. Todeschini, “DRAGON software: An easy approach to molecular descriptor calculations,” *MATCH Commun. Math. Comput. Chem.*, vol. 56, pp. 237–248, 2006.

[42] H. Moriwaki et al., “Mordred: A molecular descriptor calculator,” *J. Cheminform.*, vol. 10, p. 4, 2018. doi: 10.1186/s13321-018-0258-y.

[43] C. W. Yap, “PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints,” *J. Comput. Chem.*, vol. 32, no. 7, pp. 1466–1474, 2011. doi: 10.1002/jcc.21707.

[44] A. Golbraikh and A. Tropsha, “Beware of q^2!,” *J. Mol. Graph. Model.*, vol. 20, no. 4, pp. 269–276, 2002. doi: 10.1016/S1093-3263(01)00123-1.

[45] D. Weininger, “SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules,” *J. Chem. Inf. Comput. Sci.*, vol. 28, no. 1, pp. 31–36, 1988. doi: 10.1021/ci00057a005.

الملاحق

الملحق أ: ملخّص المعاملات الفائقة الكاملة

الملحق أ.1 — جميع المعاملات الفائقة المستخدمة

المعامل	Random Forest V.02	CatBoost-V.90
الخوارزمية الأساسية	RandomForestRegressor (sklearn)	CatBoostRegressor (Yandex)
random_seed	42	42
test_size	0.20	0.20
cv_folds	5	7
n_trials (Optuna)	75	90
scoring	r2	r2
sampler	TPE (Tree-structured Parzen)	TPE
n_estimators / iterations	300	1000
max_depth / depth	16	6
min_samples_leaf	1	—
learning_rate	—	0.0460
l2_leaf_reg	—	3.36

الملحق ب: مكتبات Python المستخدمة

الملحق ب.1 — قائمة المكتبات والإصدارات المعتمدة

المكتبة	الإصدار	الاستخدام
Python	3.11	اللغة الأساسية
RDKit	2023.09	الكيمومعلوماتية، حساب الواصفات
scikit-learn	x.1.3	Random Forest، المعالجة المسبقة، CV
CatBoost	x.1.2	نموذج التعزيز التدرّجي
XGBoost	x.2.0	خوارزمية بديلة (احتياطية)
Optuna	x.3.4	ضبط المعاملات الفائقة
SHAP	x.0.44	تفسير النماذج
pandas	x.2.2	معالجة البيانات الجدولية
numpy	x.1.26	الحساب العددي
matplotlib	x.3.8	الرسوم البيانية
FastAPI	x.0.110	الواجهة الخلفية REST API

SQLAlchemy	x.2.0	ORM لقاعدة البيانات
PostgreSQL	15	قاعدة البيانات
MLflow	x.2.10	تعقّب تجارب التعلم الآلي
Celery + Redis	x.5	إدارة المهام في الخلفية
Next.js + React	x.14	الواجهة الأمامية

الملحق ج: نموذج Canonical SMILES والواصفات

الملحق ج.1 — أمثلة من الجزيئات في قاعدة البيانات

#	Canonical SMILES	ΔH°f المرجعي (kJ/mol)
1	C	74.6−
2	CC	84.0−
3	CCC	103.8−
4	CCCC	125.7−
5	C=C	52.4+
6	C#C	227.4+
7	c1ccccc1	(g) 82.9+
8	CO	(g) 201.0−
9	CCO	(l) 277.0−
10	OCC(O)C(O)C(O)C(O)CO	(s) 1273.3−

الملحق د: روابط الوصول إلى المنصة

يمكن الوصول إلى منصة QSPR Platform وتجاربها عبر الروابط الداخلية التالية (تشغيلٌ محلي عبر Docker):

- الواجهة الرئيسية: http://localhost/
- صفحة المشروع: http://localhost/en/projects/6c09714b-425f-4236-a184-476ed4509c91
- صفحة التحليل: http://localhost/en/projects/6c09714b-425f-4236-a184-476ed4509c91/analysis
- وثائق REST API: http://localhost/api/docs
- MLflow UI: http://localhost/mlflow/

كلّ النماذج المُدَرَّبة محفوظةٌ كملفات .joblib مع المعالج المسبق وكائنات نطاق التطبيقية ومخرجات SHAP، في وحدة التخزين الخاصة بالمنصة، ممّا يضمن إمكانية إعادة إنتاج جميع النتائج المعروضة في هذه المذكرة.