Chemical Physics Letters 891 (2026) 142765

Contents lists available at ScienceDirect

Chemical Physics Letters

journal homepage: www.elsevier.com/locate/cplett

Benchmarking electronic structure methods using generative AI to construct diverse reaction databases

Amir Karton *

School of Science and Technology, University of New England, Armidale, NSW 2351, Australia.

ARTICLE INFO

Keywords:
Large language models
Density functional theory
Chemical databases
DFT benchmarking

ABSTRACT

Highly diverse chemical databases play a key role in the development of accurate and generally applicable electronic structure methods. We develop and test a workflow for autonomously generating reaction databases using large language models. As a proof-of-concept, we generate a reaction database encompassing Diels–Alder cycloaddition, electrocyclic ring opening, electrophilic addition, electrophilic aromatic substitution, and acid-base reactions. We proceed to benchmark DFT methods relative to CCSD(T) reaction energies obtained via G4 (MP2) theory. The best-performing DFT methods from each of the rungs of Jacob's Ladder (mean-absolute-deviations in parentheses) are: PBE-D4 (1.91), Skala (1.86), CAM-B3LYP-D4 and ωB97X-D (1.49), and B1B95-D4 (1.48 kcal mol^{-1}).

1. Introduction

Excluding standard wavefunction methods and a handful of density functional theory (DFT) methods, the majority of electronic structure methods are empirical in nature to some extent [1]. Examples of empirical methods include the majority of DFT [2] and double-hybrid DFT (DHDFT) methods [3], tight-binding DFT methods [4], empirically motivated wavefunction methods [5], deep-learning DFT methods, [6,7], and neural network (NN) potentials [8,9]. However, the accuracy and general applicability of these empirically motivated methods are limited by the accuracy and diversity of the chemical databases used for their parameterization. Thus, the continuous development of more accurate and generally applicable empirical electronic structure methods depends on the development of accurate and increasingly diverse chemical databases [1,10–12]. The construction of these databases is a meticulous process, demanding a delicate balance of chemical diversity, accuracy, and relevance to a specific type of chemistry. The continued development of these databases is critical for pushing the boundaries of computational chemistry.

The last decade has witnessed a paradigm shift driven by the success of data-driven machine learning, particularly deep neural networks. These models have achieved remarkable success in core chemical tasks, including predicting reaction outcomes [13], proposing novel retrosynthetic pathways [14], and generating novel molecules with desired properties [15]. The recent advent of large language models (LLMs) has introduced a new frontier: the creation of autonomous agents that can reason, plan, and integrate a diverse suite of tools to solve multi-step scientific problems [16]. This has led to the development of systems like ChemCrow, which augments an LLM with chemistry-specific tools to act as a 'chemistry intern' for synthesis planning [17]. Similarly, Coscientist is an AI agent capable of autonomously searching literature, planning, and physically executing reactions using robotic lab equipment [18]. Over the past few years, generative AI (GenAI) models have been used to create novel chemical structures with desired properties [15]. Nevertheless, the autonomous de novo generation of chemical reaction databases for benchmarking approximate electronic structure methods remains largely unexplored. A significant gap exists between using AI to predict the outcome of a given reaction and tasking it to create a diverse set of reactions from scratch. The present work addresses this gap, paving the way for generating large and highly diverse reaction databases. We explore whether GenAI can be tasked with the systematic construction of a diverse set of chemical reactions to enable the rigorous benchmarking of electronic structure methods.

The ability to automatically generate benchmark reaction databases via GenAI algorithms presents several advantages relative to human-generated benchmark databases. These include:

- Automation and scalability: GenAI can construct extensive databases in a fraction of the time required for manual curation, enabling the rapid generation of large and diverse datasets.

* Corresponding author.
E-mail address: amir.karton@une.edu.au.

https://doi.org/10.1016/j.cplett.2026.142765
Received 23 January 2026; Received in revised form 12 March 2026; Accepted 18 March 2026
Available online 23 March 2026

- Chemical diversity: LLMs trained on vast chemical data can propose a broader range of reactions and reduce the subjective selection bias of manual curation.
- Systematic curation: The automated nature of the generation and validation protocol ensures a systematic approach, which mitigates the subjective selection criteria associated with manual curation.

While our automated protocol (see the Computational Details section) rigorously validates each proposed reaction, eliminating errors in stoichiometry, charge, and spin balance, there is still room for errors to emerge. These include the risk of including irrelevant reaction types or proposing chemically plausible but mechanistically incorrect constitutional isomers, as the model may lack the nuanced, intuitive understanding of a human expert. Therefore, some degree of expert human verification or computational screening remains a valuable step in the final curation process. Furthermore, the quality of the generated database is inherently dependent on the data on which the LLM was trained. It is important to clarify that this study focuses exclusively on benchmarking reaction energies, rather than reaction barrier heights and reaction mechanisms. A central aim of the present work is to assess the balance between these advantages and disadvantages. Here, we present a series of GenAI-generated reaction databases and evaluate their chemical diversity and relevance for benchmarking electronic structure methods. Finally, we also report the first automated, GenAI-generated benchmark study. Specifically, we assess the performance of 18 conventional DFT methods and one deep-learning DFT method relative to the reaction energies calculated at the CCSD(T) level [19] by means of the G4(MP2) thermochemical protocol [20].

2. Computational details

We generate a benchmark reaction database using a standalone, general-purpose Python script designed to automate the interaction with an LLM and perform rigorous validation of a proposed set of chemical reactions. The script is available as Supporting Information. An overview of the main components of the script is as follows. We use the OpenAI GPT-4o model [21] as the core generative model, however any other open source or closed source model can be chosen. The model is prompted with user-defined criteria (e.g., "provide a set of pericyclic reactions") to propose a series of stoichiometrically balanced chemical reactions, complete with the charge and multiplicity for each species. This step is followed by a validation protocol to ensure the chemical integrity of the database. Fig. 1 illustrates the various steps in this workflow. The process begins with an LLM proposing candidate reactions, which then undergo a series of validation and curation steps. These include ensuring stoichiometric integrity (atom, spin, and charge balance), eliminating duplicate and reverse reactions, and filtering based on molecular size and elemental composition. Finally, the geometries of all species in the curated database are generated from their SMILES strings using the RDKit toolkit [22] and automatically pre-optimized using the GFN2-xTB semi-empirical tight-binding quantum chemistry method [4], providing a consistent and physically reasonable starting structure for all reactants and products. The protocol includes the following key checks:

- All chemical species are represented by canonical SMILES. This allows for the reliable identification of chemically identical species.
- Each proposed reaction is checked for atom, spin, and charge balance. Any unbalanced reactions are discarded.
- Forward and reverse reactions (e.g., $A + B \rightarrow C + D$ and $C + D \rightarrow A + B$) are treated as a single unique entry.
- An important feature of this workflow, especially for benchmarking purposes, is the ability to limit the molecular size (e.g., molecules with up to n non-hydrogen atoms) and elemental composition (e.g., only first/second-row p-block elements) of the species in the database. The Python script will confirm these limits are enforced.

Fig. 1. Schematic representation of the automated workflow for generating chemical reaction databases.

- The number of reactions of any given type can be limited to a user-defined maximum. However, it is not always possible to set a minimum number of reactions, since the minimum may be limited by the ability of the LLM to generate reactions of a given type.
- All the structures in the database are optimized using the GFN2-xTB method. These structures provide a good starting point for subsequent optimization at the DFT level.

The entire process, from the initial user prompt to the final output, is fully automated. The script generates a self-contained directory for each database, which includes the optimized XYZ coordinates for all species, alongside a log file with the final set of validated reactions.

To demonstrate the applicability of this approach, we use the above workflow to generate a database for a comprehensive DFT benchmark study. For this purpose we obtain reaction energies at the CCSD(T) level via the G4(MP2) composite ab initio method [20]. In particular, all the structures are fully optimized and confirmed to have all real harmonic frequencies at the B3LYP/6-31G(2df,p) level of theory as prescribed in the G4(MP2) thermochemical protocol. [23,24] We note that B3LYP has been shown to reproduce CCSD(T)/CBS reference equilibrium structures for a diverse set of organic and inorganic species with an RMSD of 0.0059 Å, which translates into sub-kJ mol^{-1} errors in subsequent high-level thermochemical calculations [1,25]. While the initial LLM and RDKit generation steps do not guarantee exhaustive conformational sampling, further structural sampling can be incorporated if required. Importantly, the workflow evaluates both the DFT and high-level reference methods on identical optimized geometries, eliminating any risk of comparing different isomers or conformers. All the B3LYP-optimized structures generated in the present work are provided in the Supporting Information as xyz files.

It should be pointed out that the G4(MP2) composite ab initio method has been found to produce gas-phase thermochemical properties (such as reaction energies, bond dissociation energies, and enthalpies of formation) with mean absolute deviations (MADs) approaching or below the threshold of chemical accuracy (arbitrarily defined as 1.0 kcal mol^{-1}) relative to highly accurate reference data. For example, for the 454 experimental energies of the G3/05 testset [26], G4(MP2) attains a MAD of 1.1 kcal mol^{-1} [20]. We note that the reactions in the G3/05 test set include enthalpies of formation, ionization energies, electron affinities, and proton affinities. For various reactions in the W4–11 database [27], G4(MP2) attains the following MADs: 1.5 (140 total atomization reactions), 0.8 (99 bond dissociation reactions), 1.2 (707 heavy-atom transfer reactions), 0.7 (20 isomerization reactions), and 0.5 (13 nucleophilic substitution reactions) kcal mol^{-1}. Furthermore, for challenging pericyclic reactions, a recent benchmark study by Vermeeren et al. showed that correlation contributions beyond CCSD(T) also remain below the threshold of chemical accuracy [28].

We use the G4(MP2) reaction energies to benchmark the performance of 18 conventional DFT functionals across the rungs of Jacob's Ladder [2] that are commonly used for calculating thermochemical properties (for recent review of best practices in DFT benchmark studies, see references [1,29,30]). As customary in DFT benchmark studies [1], the DFT calculations were carried out using the same geometries used for the reference G4(MP2) calculations. Table 1 lists the DFT functionals that are considered in the present work. The DFT single-point energy calculations are carried out in conjunction with the def2-TZVP basis set [31]. In nearly all cases, the D4 dispersion correction is included [32]. All the G4(MP2) and DFT calculations were carried out with the Gaussian 16 program suite (revision C.01) [33]. We also consider the performance of the recently developed deep-learning DFT Skala method [6]. We note that DHDFT functionals are not considered here since the G4(MP2) reference values are not deemed sufficiently accurate for evaluating the performance of DHDFT functionals [3].

Table 1
A representative set of DFT methods across the rungs of Jacob's Ladder considered in the present work in conjunction with the def2-TZVP basis set.

Type[a]	Functionals
GGA	BLYP [23,34], B97-D [35], PBE [36], BP86 [34,37]
MGGA	M06-L [38], TPSS [39], Skala[b,6]
HGGA	BH&HLYP [40], B3LYP [23,24], B3P86 [23,37]
	B3PW91 [23,41], PBE0 [42], CAM-B3LYP,[c] [43], ωB97X-D[c] [44]
HMGGA	M06 [45], M06-2X [45], B1B95 [46], TPSSh [47], PW6B95 [48]

[a] GGA = generalized gradient approximation, MGGA = meta-GGA, HGGA = hybrid-GGA, HMGGA = hybrid-meta-GGA.
[b] Deep-Learning DFT functional.
[c] Range-separated hybrid functional.

3. Results and discussion

Here, we construct and analyze a diverse chemical reaction database generated by our autonomous GenAI-driven workflow. We consider four reaction types: pericyclic, addition, substitution, and proton-transfer reactions. Specifically, we will generate five reaction subsets, namely Diels–Alder cycloaddition, electrocyclic ring opening/closing, electrophilic addition, electrophilic aromatic substitution, and acid-base reactions. We will test our workflow by analyzing the reactions generated in terms of their correctness and diversity for each reaction type. For each reaction we will then obtain accurate reaction energies at the CCSD(T) level via the G4(MP2) composite wavefunction protocol [20]. We will use these accurate reaction energies to benchmark the performance of DFT methods across the rungs of Jacob's Ladder [2].

Throughout this work, all the generated reactions were limited to molecules containing first and second row p-block elements and bromine (i.e., H, C–F, Si–Cl, and Br). To enable manual verification of the generated reactions as well as the ability to calculate G4(MP2) reference energies at a reasonable computational cost, we limited both the number of reactions generated and the molecular sizes of the species involved. Table 2 lists the number of unique reactions and molecular sizes found within each reaction class generated by the LLM with the specific settings used in this work. However, it is important to note that with different prompts and less stringent limitations it is possible to generate much larger reaction databases.

3.1. Pericyclic reactions

This category comprises two subsets: Diels–Alder cycloaddition and electrocyclic ring opening/closing reactions. These orbital symmetry-controlled reactions provide a stringent test of the ability of GenAI to handle complex stereochemistry rules. Using the automated workflow outlined above we generated 70 Diels–Alder and 73 electrocyclic reactions. Let us begin with analyzing the Diels–Alder reactions, which are

Table 2
Number and size of species and unique reactions generated by the GenAI workflow in each reaction subset. The range of G4(MP2) reaction energies in each reaction subset is given in kcal mol^{-1}.

Reaction class	Species	Non-H atoms[a]	Reactions	$\Delta E_{reac,G4(MP2)}$
Diels–Alder cycloaddition	88	8–20	70	−22.82 to −48.79
Electrocyclic ring opening	146	3–20	73	−22.64 to 23.20
Electrophilic addition	95	3–20	52	−20.43 to −39.77
Electrophilic aromatic substitution	112	7–20	77	−33.64 to 5.44
Acid-base reactions	72	3–10	99	106.22 to 172.10

[a] Number of non-hydrogen atoms in the largest product involved in the reaction.

one of the most important classes of reactions in organic chemistry. The subset generated by the LLM is highly systematic in terms of the diene and the dienophile. The set includes two acyclic and eight cyclic dienes. The two acyclic dienes are 1,3-butadiene (C_4H_6) and isoprene (C_5H_8). The cyclic dienes include the homologous series: 1,3-cyclopentadiene (C_5H_6), 1,3-cyclohexadiene (C_6H_8), 1,3-cycloheptadiene (C_7H_{10}), 1,3-cyclooctadiene (C_8H_{12}), 1,3-cyclononadiene (C_9H_{14}), 1,3-cyclodecadiene ($C_{10}H_{16}$), 1,3-cycloundecadiene ($C_{11}H_{18}$), and 1,3-cyclododecadiene ($C_{12}H_{20}$). This series allows for a systematic study of how electronic structure methods perform as a function of increasing ring size. This directly probes their ability to handle changes in ring strain, conformational flexibility of the diene, and steric environment of the reaction center. The GenAI-generated dienophiles include one pure hydrocarbon, one nitrogenated hydrocarbon, and six oxygenated hydrocarbons, namely: acrylonitrile (C_3H_3N), styrene (C_8H_8), acrolein (C_3H_4O), methyl acrylate ($C_4H_6O_2$), vinyl acetate ($C_4H_6O_2$), maleic anhydride ($C_4H_2O_3$), fumaric acid ($C_4H_4O_4$), and dimethyl maleate ($C_6H_8O_4$). As expected for typical Diels–Alder reactions, the majority of dienophiles are activated by electron-withdrawing groups (CHO, CN, COOH, COOMe, and OAc). However, it is important to note that GenAI also included styrene, which contains an electron-donating group. Overall, this autonomously generated set of reactions provides a good overview of both common and electronically diverse dienophiles.

The combination of the above dienes and dienophiles leads to a diverse range of important structural motifs, where acyclic dienes lead to simple monocyclic products and cyclic dienes lead to more complex bridged bicyclo[x.y.z]alkene systems. Therefore, forming the product scaffold with the correct regiochemistry is a significantly more complex challenge that tests the LLM's ability to recognize mechanistic patterns intrinsic to its training data rather than its ability to recall common examples. For example, in the Diels–Alder reaction between 1,3-cyclohexadiene and dimethyl maleate, a model could fail to generate the correct bicyclo[2.2.2]octene scaffold, or attach the two ester groups to the wrong carbon atoms, leading to the wrong regioisomer. While our workflow predicts the correct product scaffold in the current subset, the accurate prediction of the dominant stereoisomer and occasionally the precise regiochemical outcome of complex cycloadducts can still pose a challenge in some cases, highlighting the importance of subsequent human verification. Table S1 of the Supporting Information lists the CCSD(T)/CBS reaction energies for the Diels–Alder reactions. All the reactions in this subset are exothermic with G4(MP2) electronic reaction energies ranging between −22.82 and −48.79 kcal mol^{-1}.

Moving to the electrocyclic ring opening reactions, our GenAI workflow has generated 73 prototypical reactions. The generated subset is again highly systematic in terms of the reactants with varying ring sizes and degree of methylation. The core of the set is a homologous series of unsubstituted cycloalkenes, ranging from the highly strained cyclopropene up to the large and flexible cyclotetradecene. This series allows for a direct and systematic evaluation of how electronic structure methods handle the significant changes in ring strain, a key factor governing the reaction thermodynamics. Building upon this core set, the dataset includes a comprehensive and systematic array of methylated derivatives. For each parent ring system (from cyclobutene upwards), the degree of substitution is progressively increased. This begins with 1,2-dimethyl derivatives, moves to 1,2,3-trimethyl derivatives, and continues with tetramethyl, pentamethyl, and further polysubstituted analogues, up to the 1,2,3,4,5,6,7,8,9-nonamethylcyclodecene system. This methodical increase in methylation provides a systematic probe of how computational methods account for the combined influence of ring strain and the electron-donating nature of methyl groups on reaction energies. Overall, this set provides a rigorous and multi-dimensional benchmark for evaluating electronic structure methods, allowing for the effects of ring strain to be decoupled from those of substitution.

The ring opening of the above cycloalkenes leads to conjugated 1,3-dienes, which is an important structural motif. However, these reactions are equally challenging for Gen-AI, as predicting the correct stereochemical outcome requires the model to accurately reproduce complex patterns from the literature governed by the Woodward–Hoffmann rules. While our workflow successfully predicts the constitutional isomer formed upon ring opening, ensuring the correct stereochemistry of the resulting diene remains a challenge. Nevertheless, the energy difference between the possible product stereoisomers is expected to be smaller than the overall reaction energy in many cases. Table S1 of the Supporting Information lists the G4(MP2) reaction energies for the electrocyclic reactions. Unlike the uniformly exothermic Diels–Alder reactions, the thermodynamics here are highly sensitive to the structure of the cycloalkene, with electronic reaction energies ranging between −22.64 and +23.20 kcal mol^{-1}. The data reveals a strong dependency on ring strain: the ring opening of highly strained cyclopropene (−22.64 kcal mol^{-1}) and cyclobutene (−12.29 kcal mol^{-1}) is exothermic, while the reaction for less-strained systems like cyclopentene (+12.06 kcal mol^{-1}) and cyclohexene (+15.89 kcal mol^{-1}) is endothermic. Furthermore, the introduction of methyl substituents tends to make the reactions more endothermic (or less exothermic). For instance, while the ring opening of cyclobutene is exothermic with (ΔE = −12.29 kcal mol^{-1}), the reaction for 1,2-dimethylcyclobutene is less so (ΔE = −8.31 kcal mol^{-1}), and less so for 1,2,3-trimethylcyclobutene (ΔE = −6.35 kcal mol^{-1}). Similarly, for the five-membered rings, the reaction energy increases from +12.06 kcal mol^{-1} for cyclopentene to +16.72 kcal mol^{-1} for the 1,2-dimethyl derivative, to +17.96 kcal mol^{-1} for the 1,2,3-trimethyl derivative, and further to +19.46 kcal mol^{-1} for the 1,2,3,4-tetramethyl derivative. Overall, this systematic dataset provides a diverse benchmark for evaluating how well electronic structure methods capture the delicate balance between ring strain and substituent effects.

3.2. Electrophilic addition reactions

The AI-generated electrophilic addition reactions comprise 52 examples. This reaction subset includes both alkenes and alkynes with two classes of electrophiles: hydrogen halides (HCl and HBr) and dibromine (Br_2). This dataset can be divided into two subsets (i) a homologous series of terminal alkenes up to 1-octadecene, and (ii) a diverse set of alkynes including both terminal (1-ynes) and internal (2-ynes) of varying chain lengths (from butyne to nonadecyne). This diverse set of unsaturated hydrocarbons allows for a detailed probe of how electronic structure methods cope with increasing alkyl chain length and the difference in reactivity between double and triple bonds. The addition of an unsymmetrical electrophile (H–X, X = Cl, Br) to an unsymmetrical alkene or alkyne leads to more than one possible constitutional isomer. Therefore, predicting the outcome of these reactions tests the LLM's ability to learn Markovnikov selectivity patterns from the chemical literature. Our GenAI workflow consistently generated the correct, more stable constitutional isomer, placing the halogen atom on the more substituted carbon in accordance with the formation of the most stable carbocation intermediate. For example, in the reaction between propene and HCl, the model correctly predicted the formation of 2-chloropropane rather than the 1-chloro isomer. While the workflow correctly identifies the regioisomer, predicting the precise stereochemical outcome (e.g., *syn*- vs. anti-addition) remains a more complex task. Furthermore, the LLM exhibited a specific limitation regarding the electrophilic addition of a symmetric electrophile (Br_2) to alkynes. Although it correctly reduced the triple bond to a double bond, it frequently placed the bromine atoms on the wrong carbon atoms. These specific erroneous reactions were removed from the final dataset.

Table S1 of the Supporting Information lists the G4(MP2) reaction energies for the electrophilic addition reactions. All 52 reactions in this subset are exothermic, with electronic reaction energies ranging from −20.43 to −39.77 kcal mol^{-1}. Inspection of Table S1 reveals that the addition of HBr is consistently more exothermic than the addition of HCl to the same alkene. Similarly, additions to alkynes are significantly more exothermic than additions to their alkene counterparts, reflecting the

higher energy content of the C≡C triple bond. This diverse set of reactions provides an excellent benchmark for assessing the ability of electronic structure methods to capture these subtle thermodynamic differences.

3.3. Electrophilic aromatic substitution reactions

The fourth category consists of 77 electrophilic aromatic substitution (EAS) reactions. This set of EAS reactions provide a stringent test of the LLM's ability to handle the core principles of reactivity (activation vs. deactivation) and regioselectivity (ortho-, para- vs. meta-direction). The generated set is highly systematic, exploring the reactions of a diverse array of substituted benzenes. The substrates range from those containing strongly activating, electron-donating groups (e.g., phenol, aniline) to those with strongly deactivating, electron-withdrawing groups (e.g., nitrobenzene), as well as unsubstituted benzene. The dataset is further broken down into four key EAS reaction types: nitration, sulfonation, Friedel–Craft alkylation, and Friedel–Craft acylation.

The nitration subset comprises seven reactions involving nitric acid and substituted benzene rings. The substituents range from strongly electron-donating to strongly electron-withdrawing groups: aniline, phenol, anisole, toluene, benzene, chlorobenzene, and nitrobenzene. The G4(MP2) nitration reaction energies reflect the expected reactivity patterns based on the nature of the substituent. Specifically, they span a wide range, reflecting substantial differences in how different substituents influence the overall thermodynamics. Among the most exothermic reactions are the nitrations of phenol ($\Delta E = -33.64$ kcal mol^{-1}), benzene ($\Delta E = -30.84$ kcal mol^{-1}), and toluene ($\Delta E = -29.00$ kcal mol^{-1}). Nitrobenzene ($\Delta E = -19.42$ kcal mol^{-1}) is less favorable due to the strong electron-withdrawing character of the nitro group. Notably, aniline shows the least exothermic nitration energy ($\Delta E = -13.43$ kcal mol^{-1}), which reflects the strong resonance stabilization of the aniline reactant. Overall, these reaction energies do not follow a simple trend based strictly on substituent electron-donating or electron-withdrawing strength, highlighting the complex balance of resonance, inductive effects, and product stabilization in determining nitration thermochemistry. Similar to nitration, all sulfonation reactions are exothermic, with reaction energies ranging from −21.16 (chlorobenzene) to −26.21 (phenol) kcal mol^{-1}. Thus, the set of sulfonation reactions spans a much narrower energy range of only 5.05 kcal mol^{-1}, compared to the wide energetic range of 20.21 kcal mol^{-1} span by the nitration reactions.

The subset of Friedel–Craft alkylation reactions includes 34 reactions, providing a systematic overview at the introduction of alkyl groups onto the aromatic ring. The set of substituted aromatic rings is the same as for the nitration/sulfonation reactions, and the set of electrophiles systematically varies the length of the alkyl chain, from methyl chloride up to tridecyl chloride, including one branched electrophile (isopropyl chloride). The reaction energies range from moderately exothermic to thermoneutral, with most reaction energies in the range of −5 to −10 kcal mol^{-1}. As with nitration, the reaction with aniline is an exception, being slightly endothermic ($\Delta E = +0.35$ kcal mol^{-1}).

The subset of Friedel–Craft acylation reactions comprise 30 reactions. This set systematically varies the acyl chloride, from acetyl chloride up to tridecanoyl chloride, and also includes benzoyl chloride. In contrast to the EAS reactions considered so far, the Friedel–Craft acylation reactions vary from moderately exothermic to mildly endothermic. Overall, the reaction energies vary from −9.99 to +5.44 kcal mol^{-1}, demonstrating that the thermodynamics of acylation are highly sensitive to the specific combination of substrate and electrophile. Generally speaking, the acylation reaction energies exhibit two systematic trends. First, reactions involving electron-rich arenes, particularly aniline, are the most exothermic, while electron-poor systems such as nitrobenzene yield mildly endothermic reactions. Neutral and weakly activating substrates (benzene, toluene, phenol, anisole) fall between these two extremes. Second, increasing the length of the acyl chloride progressively reduces the reaction exothermicity, with small acyl groups favoring moderately exothermic reactions and long-chain acyl chlorides approaching thermoneutral or slightly endothermic reactions. Together, these trends highlight the combined influence of substituent electronics and steric bulk on the overall thermodynamics of electrophilic aromatic acylation.

3.4. Acid-base reactions

The acid-base reactions are of the form acid + base ⟶ conjugate base + conjugate acid. The acids range from simple aliphatic and aromatic carboxylic acids, while the bases span weak aromatic amines through increasingly strong aliphatic, cyclic, and polyamine bases. As expected, all the reactions are highly endothermic in the gas-phase, with reaction energies ranging between 106.22 and 172.10 kcal mol^{-1}. Generally, stronger bases (e.g., guanidine, piperazine, piperidine, morpholine) give smaller positive reaction energies, indicating more favorable proton transfer, whereas weaker or resonance-stabilized bases (aniline, pyridine) yield larger positive reaction energies, reflecting less favorable charge separation. Hydrazine-type bases fall toward the end with the largest ΔE values, consistent with substantial destabilization of the corresponding cation–anion pair in the gas phase. Hydrazine-type and hydroxylamine-derived bases consistently give the largest reaction energies, reflecting substantial destabilization of the resulting ionic pair in the gas phase. Of the acids, formic, acetic, propanoic, and benzoic acids all give broadly similar gas-phase reaction energies, with only modest variation among them. Thus, the identity of the carboxylic acid contributes less to the overall reaction energy compared to the basicity of the amine. Finally, we note that while gas-phase acid-base reactions are dominated by electrostatics and do not reflect solution-phase equilibria, they serve as a fundamental test for DFT methods.

3.5. Benchmarking DFT for the GenAI reaction database

Table 3 gives the MADs across the five reaction types for the considered DFT methods relative to the G4(MP2) reference reaction energies. Table S2 of the Supporting Information lists the root-mean square deviations (RMSDs), mean-signed deviations (MSDs), largest negative deviation (LND), and largest positive deviation (LPD) for all the reaction subsets. Inspection of Table 3 reveals that the Diels–Alder subset is the most challenging subset for the majority of DFT functionals. For example, this is the only subset for which several functionals (mostly GGAs and MGGAs) attain MADs above 5 kcal mol^{-1}. Most notably, BLYP-D4 and B97-D show remarkably high MADs of ∼12 kcal mol^{-1}. Other popular functionals, including M06L-D4 and TPSS-D4, also yield poor performance with MADs exceeding 5 kcal mol^{-1}. The poor performance of semi-local functionals is attributed to their inability to accurately describe medium-range electron correlation that becomes increasingly important in compact cycloadducts. [49,50] This deficiency persists despite the inclusion of D4 dispersion corrections. In contrast, hybrid functionals, which partially mitigate self-interaction and incorporate nonlocal exchange, generally perform better. For instance, B3P86-D4, B3PW91-D4, and ωB97X-D attain MADs of 1.83, 1.37, 1.36 kcal mol^{-1}, respectively, and the HMGGA functional B1B95-D4 attains exceptionally good performance with a MAD of 0.94 kcal mol^{-1}.

The Electrocyclic Ring Opening, Electrophilic Addition, and Electrophilic Aromatic Substitution subsets represent an intermediate level of difficulty for most DFT methods. For these subsets, most hybrid and meta-GGA functionals perform reasonably well, with MADs typically in the range of 1.5–3.0 kcal mol^{-1}. Notably, the PBE-D4 and M06-2X yield the lowest MADs of all methods for electrocyclic ring opening reactions (0.65 and 0.58 kcal mol^{-1}, respectively). For the electrophilic addition subset, several methods perform well, notably M06-2X and Skala with MADs of 1.21 and 1.22 kcal mol^{-1}, respectively. In contrast, electrophilic aromatic substitution remains more challenging, with no

Table 3
Mean-absolute deviations (MADs) in kcal mol^{-1} for the DFT methods considered in the present work in conjunction with the def2-TZVP basis set for the five reaction subsets generated by GenAI.

		Diels–Alder	Electrocyclic Ring Opening	Electrophilic Addition	Electrophilic Aromatic Substitution	Acid-Base Reactions	All Five Subsets
GGA	BLYP-D4	12.77	5.24	3.35	4.01	1.02	5.01
	BP86-D4	3.72	0.78	2.20	3.23	0.94	2.08
	PBE-D4	2.19	0.65	2.93	3.41	0.92	1.91
	B97-D	11.68	3.23	3.96	3.76	0.92	4.42
MGGA	TPSS-D4	5.83	1.75	2.71	4.91	0.90	3.08
	M06L-D4	7.46	1.56	2.70	4.24	2.71	3.70
	Skala	2.21	1.00	1.22	1.50	2.85	1.86
HGGA	B3LYP-D4	7.37	3.00	1.90	2.59	0.93	3.03
	B3P86-D4	1.83	1.77	2.71	1.81	1.15	1.76
	B3PW91-D4	1.37	1.31	2.29	1.79	1.20	1.53
	BH&HLYP-D4	2.45	1.23	1.30	1.79	2.22	1.85
	PBE0-D4	4.10	2.53	3.65	1.97	1.54	2.60
	CAM-B3LYP-D4	2.20	0.84	1.30	1.75	1.37	1.49
	ωB97X-D	1.36	1.73	1.37	2.22	0.90	1.49
HMGGA	B1B95-D4	0.94	1.42	1.58	1.59	1.74	1.48
	PW6B95-D4	3.15	0.88	1.61	2.36	1.81	1.97
	M06-D4	1.62	1.11	2.15	3.08	3.50	2.40
	M06-2X	1.78	0.58	1.21	2.56	2.18	1.73
	TPSSh-D4	3.23	0.84	2.68	4.06	0.68	2.18

considered method attaining a MAD below 1.5 kcal mol^{-1}, with the deep-learning Skala functional emerging as the best performer for this subset.

Finally, the Acid-Base Reactions subset is the least challenging for most of the considered DFT methods. A significant number of functionals, including BP86-D4, PBE-D4, B97-D, TPSS-D4, B3LYP-D4, ωB97X-D, and TPSSh-D4 achieve chemical accuracy with MADs below 1.0 kcal mol^{-1}. These reactions primarily involve local changes in bonding and charge at the protonation site, and they are less sensitive to the medium and long-range correlation effects that are more prominent in the other subsets. Overall, the performance variability across these five distinct reaction subsets underscores the need for diverse, multi-faceted benchmark databases.

Looking at the MADs across all reactions from the five subsets, we can identify more general best/worst performers from each rung of Jacob's Ladder. The best-performing GGA method is PBE-D4 with an overall MAD of 1.91 kcal mol^{-1}. Of the MGGA methods, the deep-learning Skala method emerges as the best performer with an overall MAD of 1.86 kcal mol^{-1}. Of the HGGA methods, the range-separated functionals CAM-B3LYP-D4 and ωB97X-D emerge as the best performers, both attaining a MAD of 1.49 kcal mol^{-1}, closely followed by B3PW91-D4 with a MAD of 1.53 kcal mol^{-1}. The HMGGA method B1B95-D4 also attains a similar MAD of 1.48 kcal mol^{-1}. Thus, overall, the best-performing GGA and MGGA methods attain similar MADs of just below 2 kcal mol^{-1}, and the best-performing HGGA and HMGGA methods attain similar MADs of about 1.5 kcal mol^{-1}.

4. Conclusions

We have developed and validated an autonomous workflow for generating highly diverse reaction databases using LLMs. Our workflow uses generative AI to generate reactions of a certain type and then checks for stoichiometric and spin/charge balance. We demonstrate that it is possible to rapidly construct high-quality datasets with minimal human intervention. As a proof-of-concept, we generated a database of 371 reactions spanning five distinct chemical categories: Diels–Alder cycloaddition, electrocyclic ring opening, electrophilic addition, electrophilic aromatic substitution, and acid-base reactions. This diverse set covers a broad range of chemical motifs, ring strains, and electronic environments, making it suitable for benchmarking and developing electronic structure methods. We use this GenAI-generated database to benchmark the performance of 18 DFT methods and the deep-learning Skala functional against high-level CCSD(T) reference reaction energies obtained via G4(MP2) theory. The performance of the DFT methods is highly dependent on the reaction type. The Diels–Alder subset proves to be highly challenging for many functionals, while electrophilic aromatic substitution remains moderately difficult, with no functionals attaining a mean-absolute deviation below 1.5 kcal mol^{-1}. In contrast, several functionals excel at electrophilic additions, achieving MADs approaching 1.2 kcal mol^{-1}. Overall, for the entire database of 371 reactions, the best-performing functional from each rung of Jacob's Ladder are (MAD given in parentheses):

- GGA: PBE-D4 (1.91 kcal mol^{-1})
- MGGA: Skala (1.86 kcal mol^{-1})
- HGGA: CAM-B3LYP-D4 and ωB97X-D (1.49 kcal mol^{-1})
- HMGGA: B1B95-D4 (1.48 kcal mol^{-1})

Overall, the HMGGA functional B1B95-D4 emerges as a top performer alongside the HGGAs CAM-B3LYP-D4 and ωB97X-D. All three demonstrate excellent accuracy across the diverse reaction subsets, clocking in with overall MADs of 1.48–1.49 kcal mol^{-1}, closely followed by B3PW91-D4 at 1.53 kcal mol^{-1}.

This work establishes that GenAI-driven workflows can effectively bridge the gap between ad-hoc reaction databases and large-scale, systematic database construction. By mitigating manual curation biases and automating the generation process, this approach paves the way for the rapid development of next-generation benchmark databases, essential for training and validating ML potentials and empirical electronic structure methods. Nevertheless, while the automated workflow strictly enforces stoichiometry and mass balance, expert human verification may still be required in some cases to definitively confirm the precise regiochemistry or mechanistic viability of the LLM-proposed products.

CRediT authorship contribution statement

Amir Karton: Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Investigation, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Amir Karton reports financial support was provided by University of New England, Australia. Amir Karton is an Editor of Chemical Physics Letters but was blinded from the peer-review process for this paper. Apart from that, there are no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The present work was undertaken with the assistance of resources from the National Computational Infrastructure (NCI), which is supported by the Australian Government. We gratefully acknowledge the system administration support provided by the Faculty of Science, Agriculture, Business and Law at the University of New England to the Linux cluster of the Karton group.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.cplett.2026.142765.

Data availability

The data supporting this study are available within the article and its supplementary material and from the corresponding author upon reasonable request.

References

[1] A. Karton, M.T. de Oliveira, Good practices in database generation for benchmarking DFT, WIREs Comput. Mol. Sci. 15 (2024) e1737.
[2] J.P. Perdew, K. Schmidt, Jacob's ladder of Density functional approximations for the exchange-correlation energy, AIP Conf. Proc. 577 (2001) 1.
[3] J.M.L. Martin, Golokesh Santra, empirical double-hybrid Density functional theory: A 'third way' in between WFT and DFT, Isr. J. Chem. 60 (2020) 787.
[4] C. Bannwarth, E. Caldeweyher, S. Ehlert, A. Hansen, P. Pracht, J. Seibert, S. Spicher, S. Grimme, Extended tight-binding quantum chemistry methods, WIREs Comput. Mol. Sci. 11 (2020) e01493.
[5] S. Grimme, L. Goerigk, R.F. Fink, Spin-component-scaled Electron correlation methods, WIREs Comput. Mol. Sci. 2 (2012) 886.
[6] G. Luise, C.-W. Huang, T. Vogels, D. P. Kooi, S. Ehlert, S. Lanius, K. J. H. Giesbertz, A. Karton, D. Gunceler, M. Stanley, W. P. Bruinsma, L. Huang, X. Wei, J. G. Torres, A. Katbashev, R. C. Zavaleta, B. Máté, S. Kaba, R. Sordillo, Y. Chen, D. B. Williams-Young, C. M. Bishop, J. Hermann, R. van den Berg, P. Gori-Giorgi, Accurate and scalable exchange-correlation with deep learning. arXiv:2506.14665 [physics.chem-ph] (2025).
[7] J. Kirkpatrick, B. McMorrow, D.H.P. Turban, A.L. Gaunt, J.S. Spencer, A.G.D. G. Matthews, A. Obika, L. Thiry, M. Fortunato, D. Pfau, L.R. Castellanos, S. Petersen, A.W.R. Nelson, P. Kohli, P. Mori-Sánchez, D. Hassabis, A.J. Cohen, Pushing the frontiers of density functionals by solving the fractional electron problem, Science 374 (2021) 1385.
[8] J. Behler, Four generations of high-dimensional neural network potentials, Chem. Rev. 121 (2021) 10037.
[9] N. Fedik, R. Zubatyuk, M. Kulichenko, N. Lubbers, J.S. Smith, B. Nebgen, R. Messerly, Y.W. Li, A.I. Boldyrev, K. Barros, O. Isayev, S. Tretiak, Extending machine learning beyond interatomic potentials for predicting molecular properties, Nat. Rev. Chem. 6 (2022) 653.
[10] A. Karton, Three decades of quantum science: how quantum chemistry transformed thermochemical database generation for benchmarking DFT and machine learning. Aust. J. Chem., 78 (2025) CH24130.
[11] L. Goerigk, A. Hansen, C. Bauer, S. Ehrlich, A. Najibi, S. Grimme, A look at the density functional theory zoo with the advanced GMTKN55 database for general main group thermochemistry, kinetics and noncovalent interactions, Phys. Chem. Chem. Phys. 19 (2017) 32184.
[12] J. Liang, M. Head-Gordon, Gold-Standard chemical database 137 (GSCDB137): a diverse set of accurate energy differences for assessing and developing density functionals, J. Chem. Theory Comput. 21 (2025) 12601.
[13] M.K. Schwaller, D. Probst, A.C. Vaucher, V.H. Nair, J.-L. Reymond, Mapping the space of chemical reactions using attention-based neural networks, Nat. Mach. Intell. 3 (2021) 144.
[14] M.H.S. Segler, M. Preuss, M.P. Waller, Planning chemical syntheses with deep neural networks and symbolic AI, Nature 555 (2018) 604.
[15] A. Zhavoronkov, Y.A. Ivanenkov, A. Aliper, M.S. Veselov, V.A. Aladinskiy, A. V. Aladinskaya, V.A. Terentiev, D.A. Polykovskiy, M.D. Kuznetsov, A. Asadulaev, Y. Volkov, A. Zholus, R.R. Shayakhmetov, A. Zhebrak, L.I. Minaeva, B. A. Zagribelnyy, L.H. Lee, R. Soll, D. Madge, L. Xing, T. Guo, A. Aspuru-Guzik, Deep learning enables rapid identification of potent DDR1 kinase inhibitors, Nat. Biotechnol. 37 (2019) 1038.
[16] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, ReAct: Synergizing Reasoning and Acting in Language Models, *Proceedings of the Eleventh International Conference on Learning Representations*, 2023.
[17] A.M. Bran, S. Cox, O. Schilter, C. Baldassari, A.D. White, P. Schwaller, Nat. Mach. Intell. 6 (2024) 525.
[18] D.A. Boiko, R. MacKnight, G. Kline, G. Gomes, Autonomous chemical research with large language models, Nature 624 (2023) 570.
[19] K. Raghavachari, G.W. Trucks, J.A. Pople, M. Head-Gordon, A fifth-order perturbation comparison of electron correlation theories, Chem. Phys. Lett. 157 (1989) 479.
[20] L.A. Curtiss, P.C. Redfern, K. Raghavachari, Gaussian-4 theory using reduced order perturbation theory, J. Chem. Phys. 127 (2007) 124105.
[21] A. Hurst, A. Lerer, A.P. Goucher, A. Perelman, A. Ramesh, A. Clark, A.J. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al., GPT-4o system card, arXiv:2410.21276 (2024). https://arxiv.org/abs/2410.21276.
[22] G. Landrum, *et al.*, RDKit: Open-source cheminformatics, https://zenodo.org/records/17495409 (2025), see also: https://www.rdkit.org.
[23] C. Lee, W. Yang, R.G. Parr, Development of the Colle-Salvetti correlation-energy formula into a functional of the electron density, Phys. Rev. B 37 (1988) 785.
[24] A.D. Becke, Density-functional thermochemistry. III. The role of exact exchange, J. Chem. Phys. 98 (1993) 5648.
[25] A. Karton, P.R. Spackman, Evaluation of density functional theory for a large and diverse set of organic and inorganic equilibrium structures, J. Comput. Chem. 42 (2021) 1590.
[26] L.A. Curtiss, P.C. Redfern, K. Raghavachari, Assessment of Gaussian-3 and Density-functional theories on the G3/05 test set of experimental energies, J. Chem. Phys. 123 (2005) 124107.
[27] A. Karton, S. Daon, J.M.L. Martin, W4-11: a high-confidence dataset for computational thermochemistry derived from W4 ab initio data, Chem. Phys. Lett. 510 (2011) 165.
[28] P. Vermeeren, M. Dalla Tiezza, M.E. Wolf, M.E. Lahm, W.D. Allen, H.F. Schaefer, T. A. Hamlin, F.M. Bickelhaupt, Pericyclic reaction benchmarks: hierarchical computations targeting CCSDT(Q)/CBS and analysis of DFT performance, Phys. Chem. Chem. Phys. 24 (2022) 18028.
[29] L. Goerigk, Benchmarking Modern Density Functionals for Broad Applications in Chemistry, in: R.J. Boyd, M. Yanez (Eds.), Comprehensive Computational Chemistry, 1st ed. 1, Elsevier, Amsterdam, The Netherlands, 2023, pp. 78–93.
[30] M. Bursch, J.-M. Mewes, A. Hansen, S. Grimme, Best-practice DFT protocols for basic molecular computational chemistry, Angew. Chem. Int. Ed. 61 (2022) e202205735.
[31] F. Weigend, R. Ahlrichs, Balanced basis sets of split valence, triple zeta valence and quadruple zeta valence quality for H to Rn: design and assessment of accuracy, Phys. Chem. Chem. Phys. 7 (2005) 3297.
[32] E. Caldeweyher, J.-M. Mewes, S. Ehlert, S. Grimme, Extension and evaluation of the D4 London-dispersion model for periodic systems, Phys. Chem. Chem. Phys. 22 (2020) 8499.
[33] M.J. Frisch, G.W. Trucks, H.B. Schlegel, G.E. Scuseria, M.A. Robb, J.R. Cheeseman, G. Scalmani, V. Barone, G.A. Petersson, H. Nakatsuji, et al., Gaussian 16, Revision C.01, Gaussian, Inc, Wallingford, CT, 2016.
[34] A.D. Becke, Density-functional exchange-energy approximation with correct asymptotic behavior, Phys. Rev. A 38 (1988) 3098.
[35] S. Grimme, Semiempirical GGA-type density functional constructed with a long-range dispersion correction, J. Comput. Chem. 27 (2006) 1787.
[36] J.P. Perdew, K. Burke, M. Ernzerhof, Generalized gradient approximation made simple, Phys. Rev. Lett. 77 (1996) 3865.
[37] J.P. Perdew, Density-functional approximation for the correlation energy of the inhomogeneous electron gas, Phys. Rev. B 33 (1986) 8822.
[38] Y. Zhao, D.G. Truhlar, A new local density functional for main-group thermochemistry, transition metal bonding, thermochemical kinetics, and noncovalent interactions, J. Chem. Phys. 125 (2006) 194101.
[39] J.M. Tao, J.P. Perdew, V.N. Staroverov, G.E. Scuseria, Climbing the density functional ladder: nonempirical meta-generalized gradient approximation designed for molecules and solids, Phys. Rev. Lett. 91 (2003) 146401.
[40] A.D. Becke, A new mixing of Hartree–Fock and local density-functional theories, J. Chem. Phys. 98 (1993) 1372.
[41] J.P. Perdew, J.A. Chevary, S.H. Vosko, K.A. Jackson, M.R. Pederson, D.J. Singh, C. Fiolhais, Atoms, molecules, solids, and surfaces: applications of the generalized gradient approximation for exchange and correlation, Phys. Rev. B 46 (1992) 6671.
[42] C. Adamo, V. Barone, Toward reliable density functional methods without adjustable parameters: the PBE0 model, J. Chem. Phys. 110 (1999) 6158.

[43] T. Yanai, D. Tew, N. Handy, A new hybrid exchange correlation functional using the coulomb-attenuating method (CAM-B3LYP), Chem. Phys. Lett. 393 (2004) 51.
[44] J.-D. Chai, M. Head-Gordon, Systematic optimization of long-range corrected hybrid density functionals, J. Chem. Phys. 128 (2008) 084106.
[45] Y. Zhao, D.G. Truhlar, The M06 suite of density functionals for main group thermochemistry, thermochemical kinetics, noncovalent interactions, excited states, and transition elements: two new functionals and systematic testing of four M06-class functionals and 12 other functionals, Theor. Chem. Acc. 120 (2008) 215.
[46] A.D. Becke, Density-functional thermochemistry. IV. A new dynamical correlation functional and implications for exact-exchange mixing, J. Chem. Phys. 104 (1996) 1040.
[47] V.N. Staroverov, G.E. Scuseria, J. Tao, J.P. Perdew, Comparative assessment of a new nonempirical density functional: molecules and hydrogen-bonded complexes, J. Chem. Phys. 119 (2003) 12129.
[48] Y. Zhao, D.G. Truhlar, Design of Density Functionals that are Broadly Accurate for thermochemistry, thermochemical kinetics, and nonbonded interactions, J. Phys. Chem. A 109 (2005) 5656.
[49] S. Grimme, Seemingly Simple Stereoelectronic Effects in Alkane Isomers and the Implications for Kohn–Sham Density Functional Theory, Angew. Chem. Int. Ed. 45 (2006) 4460.
[50] M.D. Wodrich, C. Corminboeuf, P. v R. Schleyer, Systematic errors in computed alkane energies using B3LYP and other popular DFT functionals, Org. Lett. 8 (2006) 3631.