scientific data

Check for updates

OPEN

DATA DESCRIPTOR

QM9star, two Million DFT-computed Equilibrium Structures for Ions and Radicals with Atomic Information

Miao-Jiong Tang[1,3], Tian-Cheng Zhu[1,3], Shuo-Qing Zhang[1] ✉ & Xin Hong[1,2] ✉

Ions and radicals serve as key intermediates in molecular transformation, with their chemical properties being essential for understanding and predicting reaction reactivity and selectivity. In this data descriptor, we report a quantum chemical dataset named QM9star, comprising cations, anions, and radicals. This dataset is derived from the molecular structures of the QM9 dataset, created by removing terminal hydrogens followed by optimization using B3LYP-D3(BJ)/6-311 + G(d,p) level of density functional theory. The QM9star dataset includes approximately 1.9 million cations, anions, and radicals, along with 120 kilo neutral molecules prior to hydrogen removal. Each entry encompasses both molecular and atomic information: representative global properties include orbital energies, vibrational frequencies, etc., while local properties cover aspects such as charges and spin densities at each atomic site. The QM9star dataset not only serves as a comprehensive source of quantum chemical information for intermediates but also offers insights into the principle of atomic property distribution. We anticipate that these data will aid in machine learning studies related to chemical intermediates and contribute to the molecular representation learning.

Background & Summary

The recent emergence of machine learning (ML), particularly deep learning, has revolutionized data-driven molecular modeling and attracted extensive attention[1–10]. Represented by message passing neural networks (MPNNs)[11], graph neural networks (GNNs) have enabled powerful AI prediction of molecular properties on both global and local levels, significantly advancing AI modeling of molecular systems[12–20]. These advancements have propelled the accuracy and efficiency of molecular property predictions to levels comparable with density functional theory (DFT) calculation, yet at orders of magnitude lower costs[21,22]. Such transformative progress has substantially empowered chemists' exploration of the molecular world, opening up new possibilities for downstream applications, such as drug screening[23,24], material design[20], and chemical simulation[16,20,25] in the AI era.

High-quality, large-scale molecular datasets are indispensable for building molecular property prediction models, directly determining their accuracy and generalization capabilities. Early efforts primarily involved the collection and organization of experimental data. Representative examples include the Cambridge Structural Database[26], the PubChem database[27], and the Internet Bond-energy Databank (iBonD)[28], *et al.* These experimental databases provide extensive data sources that not only enhance the understanding of molecular properties and serve as benchmarks for computational methods but also offer substantial support for the training and validation of molecular ML models.

Advancements in computational chemistry methods and improvements in computer hardware have provided a complementary approach for the exploration and generation of large-scale molecular datasets, making theoretical calculations an integral part of today's molecular data resources. A prime example in this regard is the QM series datasets[29–34] developed by von Lilienfeld and colleagues. The QM7[29,30] dataset includes structures

[1]Center of Chemistry for Frontier Technologies, Department of Chemistry, Zhejiang University, Hangzhou, 310027, P. R. China. [2]School of Chemistry and Chemical Engineering, Henan Normal University, Xinxiang, 453007, P. R. China. [3]These authors contributed equally: Miao-Jiong Tang, Tian-Cheng Zhu. ✉e-mail: angellasty@zju.edu.cn; hxchem@zju.edu.cn

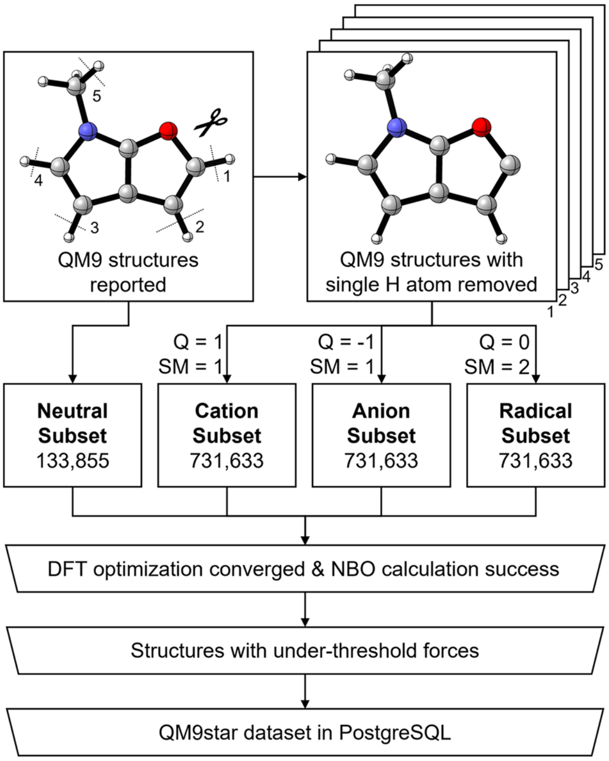

5
4
3
2
1
QM9 structures reported
QM9 structures with single H atom removed
1
2
3
4
5
Q = 1
SM = 1
Q = -1
SM = 1
Q = 0
SM = 2
Neutral Subset 133,855
Cation Subset 731,633
Anion Subset 731,633
Radical Subset 731,633
DFT optimization converged & NBO calculation success
Structures with under-threshold forces
QM9star dataset in PostgreSQL

Fig. 1 Generation Workflow of the QM9star Dataset. Starting from the three-dimensional structures of neutral molecules in QM9, each single terminal hydrogen atom is removed, with redundant structures having equivalent topologies eliminated. This results in three-dimensional structures with one fewer hydrogen atom than the neutral molecules in QM9. These structures are assigned global charges and spin multiplicities of (0, 2), (1, 1), and (−1, 1) to obtain the initial structures of radicals, cations, and anions, respectively. These initial structures, along with the original neutral structures from QM9, are optimized at the B3LYP-D3(BJ)/6-311 + G(d,p) level, retaining those that optimize successfully and pass NBO calculations. Structures with atomic forces exceeding a threshold or those with imaginary frequencies are discarded. The resulting QM9star dataset is stored as a PostgreSQL database.

and energies for 7 kilo neutral molecules at the PBE0 level; QM7b[29,32] extends QM7 with additional theoretical levels; QM8[31,34] covers electronic spectra of 20 kilo molecules computed using time-dependent DFT; and the latest QM9[31,33] dataset contains structures and properties of 134 kilo equilibrium structures of neutral molecules optimized at the B3LYP/6-31 G(2df,p) level. Based on the QM series datasets, successful efforts have further expanded their depth and breadth. Notably, the G4MP2-QM9[35] and MultiXC-QM9[36] datasets extend QM9 with calculations at higher theoretical levels. Additionally, the QM-sym[37] and QM-symex[38] datasets derive hundreds of thousands of molecules with C_nh symmetry from QM9 structures by increasing and decreasing symmetry, thereby enhancing the diversity of the chemical space.

Beyond the expansions of the QM series datasets, sampling from other chemical spaces has also produced valuable molecular datasets. Among these, theoretical calculations based on the PubChem[27] database have generated representative collections such as the PubChemQC series[39–41] and PC9[42]. PubChemQC[39] includes DFT-level ground-state electronic structures for 3 million drug-like molecules. The PubChemQC PM6[40] dataset encompasses 230 million molecules optimized using the semi-empirical PM6 method, and PubChemQC B3LYP/6-31 G*//PM6[41] includes 86 million molecules with coordinates optimized by PM6[43] and their electronic structures calculated using DFT. The PC9[42] dataset is a subset of PubChemQC, containing 99 kilo molecules that meet the QM9 element restrictions. Additionally, the Alchemy[44] dataset is sampled from the GDBMedChem[29] database and provides a collection of 200 kilo organic molecules with structures and corresponding energies. The QMugs[45] dataset, derived from the ChEMBL[46] database, includes 665 kilo drug-like molecules. These molecular geometries are optimized at the GFN2-xTB level, with corresponding quantum chemical properties computed using ωB97X-D/def2-SVP.

In addition to optimized equilibrium structures, there is a growing need for datasets containing non-equilibrium structures, particularly for molecular dynamics simulation applications. The ANI-1 series[47,48] datasets are early representatives of this effort. The ANI-1[47] dataset provides over 20 million non-equilibrium structures with corresponding energies for 57 kilo molecules, calculated using ωB97x/6-31 G(d). Subsequent datasets, ANI-1x and ANI-1ccx[48], further enhance the sampling size and theoretical levels. Building on the QM7 dataset, the QM7-X[49] dataset expands to include non-equilibrium structures generated through DFTB normal-mode displacements. The GEOM[50] dataset utilizes molecular dynamics simulations to produce 37 million snapshots for 450 kilo molecules. The SPICE[51] dataset, based on dynamical simulations, offers energy and

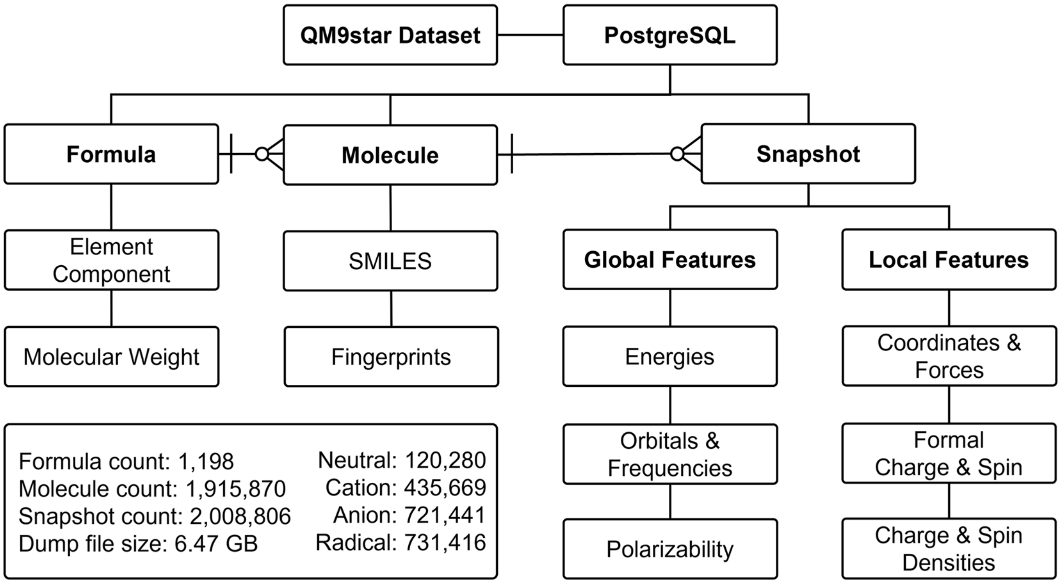


Fig. 2 Schematic of the QM9star Data Structure. The **Formula** table presents molecular formulas and facilitates the querying of information related to elements and atomic weights. This table is interconnected with the **Molecule** table, which captures the topological structure of molecules, including their SMILES representations—particularly canonical SMILES and InChI strings—as well as global charges, spin multiplicities, and derived information such as molecular fingerprints. The **Molecule** table supports both precise and fuzzy queries based on SMILES and is linked to the **Snapshot** table. The **Snapshot** table contains detailed three-dimensional structures of the molecules, along with global and local features, enabling comprehensive filtering and retrieval of all relevant characteristics.

force data at the ωB97M-D3(BJ)/def2-TZVPPD level for 2 million conformations of 114 kilo drug-like molecules, with several subsets tailored for detailed potential energy simulations in specific scenarios.

Existing molecular datasets have provided over 300 million molecular data and continue to grow, covering a broad chemical space in terms of element types, molecular sizes, and bonding types. However, since these datasets are primarily designed for the understanding and design of functional molecules, the majority focus on neutral molecules that comply with the octet rule, with very few addressing highly reactive species. In this regard, Paton[52] *et al.* reported a dataset containing 200 kilo organic radical species and their closed-shell counterparts, calculated using M06-2X/def2-TZVP. This dataset is one of the few representative efforts in this field, enabling powerful ML predictions of bond dissociation energy. However, for other reactive intermediates equally crucial in organic chemistry, such as carbocations and anions, there remains a lack of large-scale molecular datasets, to the best of our knowledge. Consequently, current datasets are still limited in their coverage of the chemical space and data accumulation for reactive intermediates.

To systematically provide information on the structures and quantitative properties of reactive intermediates, we report the QM9star dataset in this work. This dataset encompasses approximately 2 million reactive intermediates' structures and properties. The development of QM9star utilizes the three-dimensional structures from QM9[53], taking into account all possible terminal hydrogen removals. From these initial geometries, we optimized the corresponding cations, anions, and radicals at the B3LYP-D3(BJ)/6-311 + G(d,p) level, generating a series of derived reactive intermediates while benefiting from the geometric information of QM9. Through data cleaning and curation, QM9star eventually includes data on 436 kilo cations, 721 kilo anions, and 731 kilo radicals. To ensure consistency, we also re-optimized all neutral molecules from QM9 at the same theoretical level. QM9star not only contains extensive global molecular properties but also includes crucial atomic-level information, such as charges and spin densities, which are crucial for reactivity and selectivity determination. We envision that the QM9star dataset will provide useful data support for modeling organic reactions, facilitating the development and evaluation of more accurate models for reactivity and selectivity prediction. Additionally, the distribution patterns of physical organic properties within this dataset will advance representation learning in molecular science.

Methods

Starting with the three-dimensional structures of molecules from the QM9[53] dataset, we obtained the initial structures of the corresponding reactive intermediates by removing terminal hydrogen atoms and assigning charges and spin multiplicities. Subsequent optimization at the B3LYP-D3(BJ)/6-311 + G(d,p) level yielded the equilibrium three-dimensional structures, and Natural Bond Orbital (NBO) calculations provided information such as NBO bond orders and Natural Population Analysis (NPA) charges. We filtered out unreasonable data based on custom criteria, including molecular force checks and the presence of imaginary frequencies, ultimately forming the QM9star dataset. Figure 1 summarizes the data generation workflow.

No.	Property	Unit	Description
1	*Atoms		n × 1 vector of atomic numbers
2	Bonds		Formal bond orders
3	Formal charges		n × 1 vector of formal charges
4	Formal radicals		n × 1 vector of formal radical numbers
5	*Coordinates	Å	n × 3 matrix in cartesian coordinate system
6	Standard coordinates	Å	n × 3 matrix in cartesian coordinate system standardised by Gaussian 16
7	Forces	Eh/bohr	n × 3 matrix in cartesian coordinate system
8	NBO bond order		Bond orders inferred by NBO
9	NPA charges		n × 1 vector of NPA charges
10	*Mulliken charges		n × 1 vector of Mulliken charges
11	Mulliken spin densities		n × 1 vector of Mulliken spin densities
12	Single point energy	Eh	SCF energy
13	*ZPVE	Eh/particle	Zero-point vibrational energy
14	Energy correction	Eh/particle	Thermal correction to energy
15	Enthalpy correction	Eh/particle	Thermal correction to enthalpy
16	Gibbs free energy correction	Eh/particle	Thermal correction to Gibbs free energy
17	*U_0	Eh/particle	Internal energy at 0 K
18	*U_t	Eh/particle	Internal energy at 298.15 K
19	*H_t	Eh/particle	Enthalpy at 298.15 K
20	*G_t	Eh/particle	Gibbs free energy at 298.15 K
21	*S	cal/mol/K	Entropy at 298.15 K
22	*C_v	cal/mol/K	Heat capacity at 298.15 K
23	*Rotation constants	GHz	3 × 1 vector of rotational constants
24	*Isotropic polarizability	bohr^3	Isotropic polarizability (α)
25	*Electronic spatial extent	bohr^2	Electronic spatial extent ($<R^2>$)
26	*$E_{\text{alpha HOMO}}$	Eh	Energy of alpha HOMO
27	*$E_{\text{alpha LUMO}}$	Eh	Energy of alpha LUMO
28	*$E_{\text{alpha gap}}$	Eh	Energy gap between alpha LUMO and HOMO
29	$E_{\text{beta HOMO}}$	Eh	Energy of beta HOMO
30	$E_{\text{beta LUMO}}$	Eh	Energy of beta LUMO
31	$E_{\text{beta gap}}$	Eh	Energy gap between beta LUMO and HOMO
32	Frequencies	cm^{-1}	Vector of harmonic frequencies
33	Reduced masses	amu	Vector of reduced masses
34	IR intensities	km/mol	Vector of infrared intensities
35	Force constants	mdyne/Å	Vector of force constants
36	Dipole moment	debye	3 × 1 vector of dipole moment
37	Quadrupole moment	debye·Å	6 × 1 vector of quadrupole moment
38	Octapole moment	debye·Å^2	10 × 1 vector of octapole moment
39	Hexadecapole moment	debye·Å^3	15 × 1 vector of hexadecapole moment

Table 1. Properties and Descriptions in the QM9star dataset. Local features from row 1 to 11 while global features from row 12 to 39. *Also provided by QM9 dataset.

Generation of initial structures. Starting from the equilibrium geometries in the QM9 dataset, we removed a single terminal hydrogen atom from each structure while keeping the rest of the molecule unchanged. Subsequently, charges and spin multiplicities were assigned to generate initial guesses for the corresponding reactive intermediates (cations, anions, and radicals). For each structure in QM9, we examined every non-equivalent hydrogen atom and distinguished them based on their relative order in the original XYZ file. Additionally, we addressed cases where removing a hydrogen atom from different molecules in QM9 resulted in topologically equivalent structures by performing topological checks and removing redundant entries. Following these steps, a total of 731,633 distinctive structures were generated, resulting in 2,194,899 initial guesses for reactive intermediates.

DFT optimization and calculations. We used Gaussian 16 to optimize all initial guesses at the B3LYP-D3(BJ)/6-311 + G(d,p) level of theory. To maintain consistency in the dataset, we also re-optimized the neutral molecule structures from QM9 at the same level. The optimized geometries primarily correspond to the expected reactive intermediates. However, a small number of structures underwent rearrangement, resulting in different conformations of the same molecule. These conformational differences in structure, energy, and properties are chemically meaningful and were therefore retained in the dataset. Additionally, a very few structures

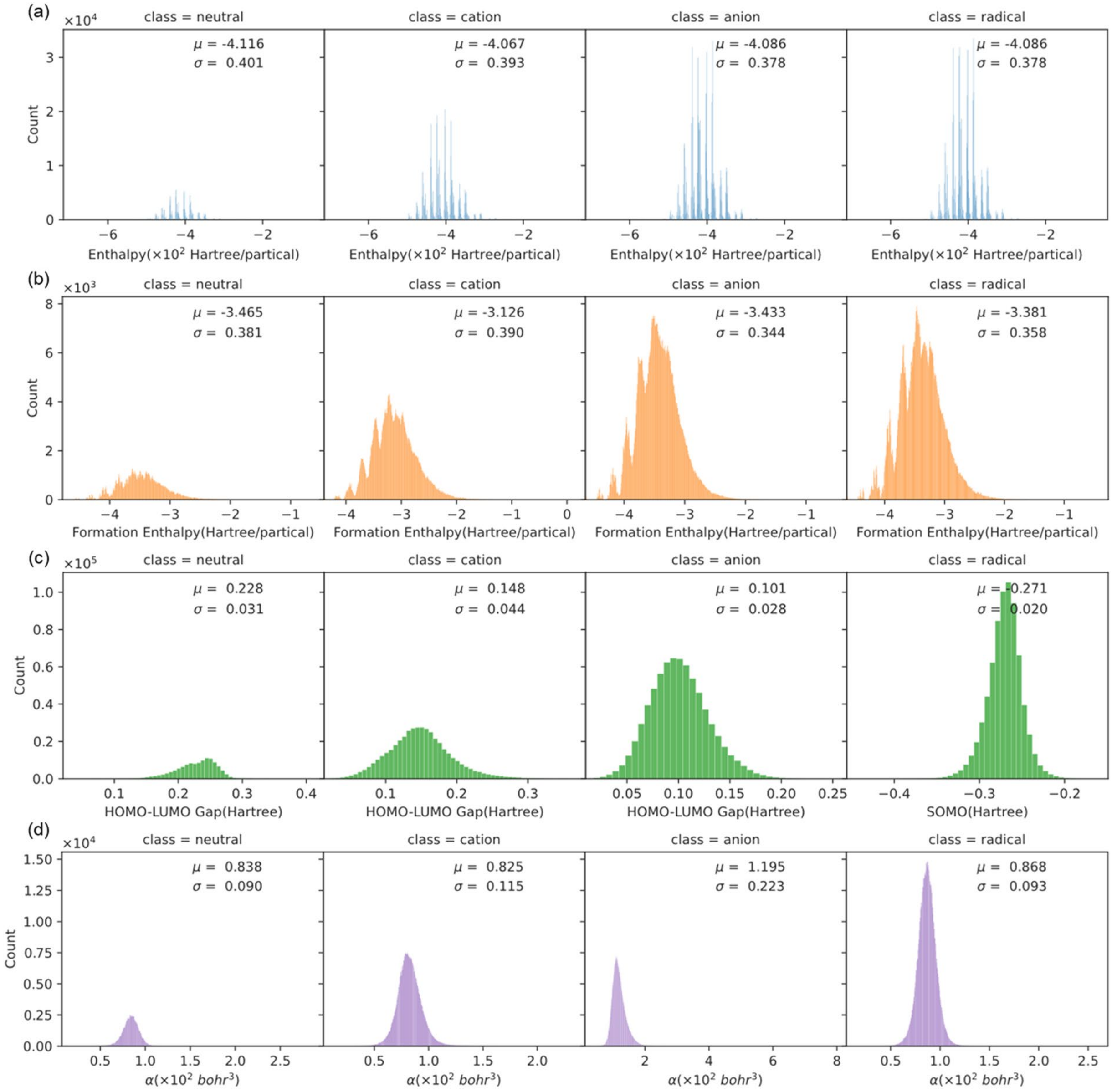


Fig. 3 Distribution of Selected Global Properties in the QM9star Dataset. (**a**) Distribution of enthalpy. (**b**) Distribution of formation enthalpy. (**c**) Distribution of energies related to the frontier molecular orbitals. (**d**) Distribution of molecular polarizability. Each range on the x-axis indicates the lower and upper limits within the respective subset.

optimized to saddle points with imaginary frequencies, which were discarded. For all successfully optimized equilibrium structures, we used NBO implemented in Gaussian 16 to calculate the corresponding NBO properties.

Data cleaning. Molecules that did not converge during geometry optimization or failed to successfully complete NBO calculations were discarded. This process led to the removal of some cases from each subset of cations, anions, and radicals, and even a few neutral molecules from the original QM9 dataset were discarded for the same reasons. For the molecules that successfully converged and obtained NBO results, we extracted atomic forces in Cartesian coordinates. Although these molecules should theoretically be local minimum with nearly zero atomic forces, we found some cases with significant outlier forces in Cartesian coordinates, likely due to differences between internal and Cartesian coordinate systems. To ensure the reliability of the dataset, molecules with maximum atomic forces exceeding 0.001 (approximately 2 kilo molecules) were removed. The remaining molecules were assigned SMILES strings for record and storage. It is noteworthy that the initial geometry guesses did not always correspond to the stable structures of reactive intermediates, leading to considerable rearrangements during optimization, particularly for cations. Due to these rearrangements, different initial structures sometimes converged to different conformations of the same molecule. Considering that the structural, energetic,

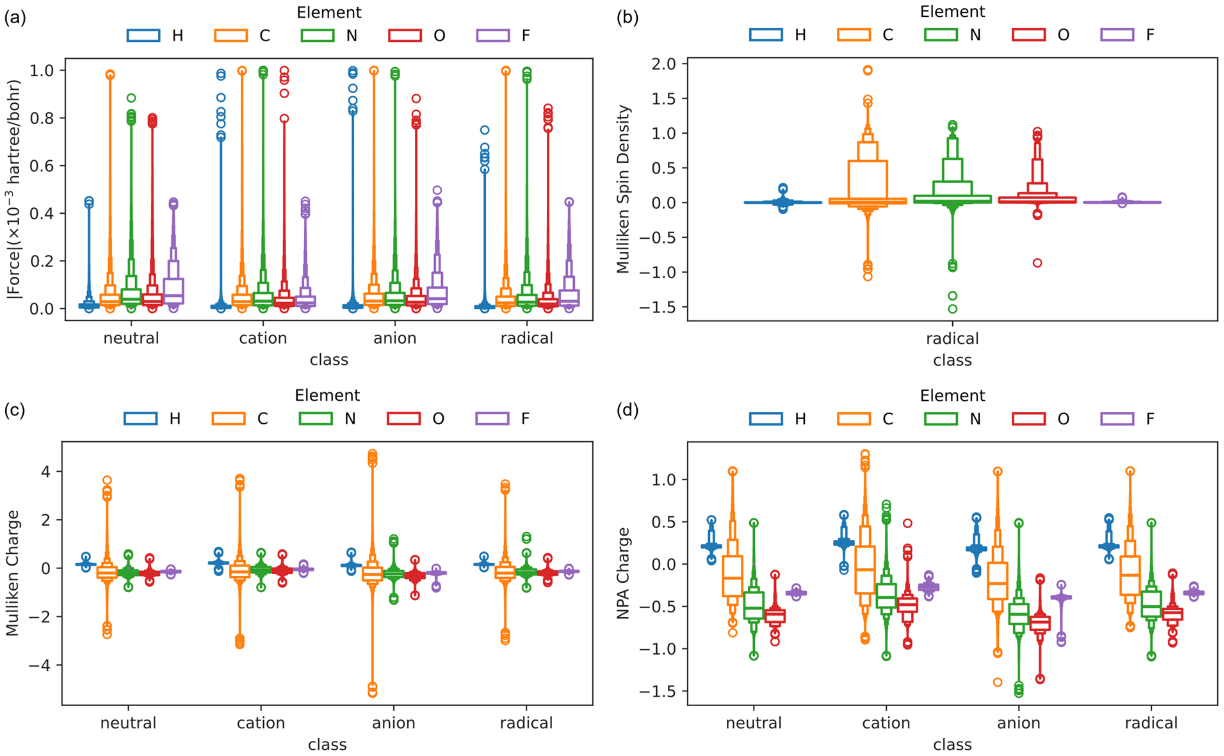


Fig. 4 Distribution of Local Features in the Dataset. (**a**) Distribution of atomic forces in Cartesian coordinates. (**b**) Distribution of Mulliken spin densities, including only the radical subset. (**c**) Distribution of Mulliken charges. (**d**) Distribution of NPA charges.

and property distributions of these conformations are chemically meaningful, we retained this data. The final QM9star dataset comprises 1,915,870 topological structures and 2,008,806 three-dimensional structures.

Data Records

QM9star data is stored in a PostgreSQL database, with a dump file available on figshare[54] to facilitate database restoration. The database structure is illustrated in Fig. 2. Each three-dimensional molecular structure includes a molecular formula and SMILES string for retrieval purposes and allows for queries based on elemental composition and molecular fingerprint similarity. We also provided canonical SMILES and InChI strings to accommodate a broader range of query requirements. For each recorded three-dimensional structure, the database offers extensive global and local quantum chemical information, including energies and forces, charges, spins, dipoles, thermodynamic energies, frequencies, vibrations, NBO bond orders, and more. Additionally, formal charges, formal spins, and bond connections are provided, enabling users to reconstruct RDKit[55] Mol objects for further processing. Table 1 lists the 39 available fields along with their respective units and descriptions. We have also provided a code repository (https://github.com/gentle1999/qm9star_query) demonstrating how to download and deploy the QM9star dataset. The repository includes examples on how to use the QM9star dataset, allowing users to access the dataset efficiently with ready-to-use query code. Furthermore, we offer a dataset class based on PyTorch Geometric[56] specifications to help users quickly apply the QM9star dataset in deep learning workflows.

Technical Validation

Dataset distribution. Figure 3 illustrates the distribution of selected global properties in the QM9star dataset. Notably, the enthalpy exhibits a multimodal distribution (Fig. 3a). This is because enthalpy (and other thermodynamic energies of the entire molecule) is highly correlated with the atomic composition of the molecule, and the discontinuous changes in atomic composition within the dataset lead to these variations. Such multimodal distributions are not conducive to machine learning modeling, but can be conveniently addressed through target transformation. Therefore, in the database, we transformed the corresponding target values by converting to the formation enthalpy. This formation enthalpy is defined as the total enthalpy of the molecule minus the enthalpy of all its constituent atoms, thereby eliminating the effect of discontinuous atomic composition (Fig. 3b). As expected, the distribution of formation enthalpy becomes nearly unimodal, and the standard deviation is significantly reduced, facilitating subsequent model training (vide infra). Figure 3c shows the data distribution related to one of the most critical properties determining molecular reactivity—the frontier molecular orbitals. The distribution reveals some interesting phenomena: the cation and anion subsets are nearly normally distributed, whereas the neutral subset exhibits a distinct bimodal distribution, indicating that global charge has a notable impact on the HOMO-LUMO gaps. For radicals, the SOMO shows a well-defined unimodal distribution. Regarding the polarizability distribution (Fig. 3d), the mean values for the neutral, cation, and radical subsets are

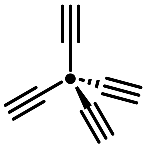
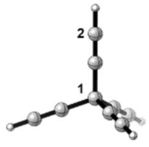
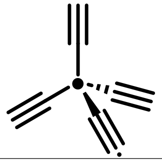
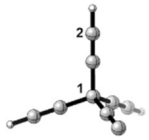
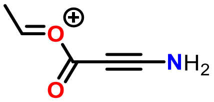
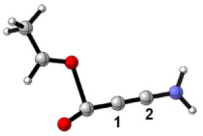
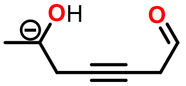
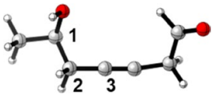
Subset	Molecule	Geometry	Index	Mulliken Charges	NPA Charges
Neutral			1	3.64	−0.31
			2	−1.14	−0.17
Radical			1	3.48	−0.31
			2	−1.13	−0.16
Cation			1	3.71	−0.28
			2	−2.85	0.45
Anion			1	1.10	−0.11
			2	−5.17	−0.54
			3	4.49	0.07

Table 2. Molecules with Outlier Mulliken Charges. This table presents molecules from the neutral, radical, cation, and anion subsets with the highest absolute Mulliken charges. Only atoms with absolute Mulliken charges greater than 1 are labeled with their corresponding Mulliken and NPA charges.

relatively similar. In contrast, the anion subset exhibits a significantly higher mean value and is characterized by a long-tail distribution.

The distributions of local properties in the QM9star dataset are presented in Fig. 4. With filtering based on atomic force, the local forces within molecules are all within the 0.001 hartree/bohr threshold (Fig. 4a), thusly the provided structures are near equilibrium. For the distribution of Mulliken spin density (Fig. 4b), the spin density on H and F atoms is almost zero, with the majority concentrated on C and N atoms, aligning with the chemical understanding that spin density is more prevalent on less electronegative atoms. We also note that the variance in Mulliken charge distribution is much greater than that of NPA charges (Fig. 4c,d), likely due to the inherent limitations of Mulliken charge calculation. When using larger basis sets with diffuse functions like 6-311 + G(d,p), Mulliken charge analysis may lack numerical stability[57]. In Table 2, we showcase several representative molecules with Mulliken charges exceeding an absolute value of 3. It is evident that these outlier Mulliken charges often occur on triple bonds, whereas the distribution of NPA charges is more consistent with chemical intuition.

Training potential energy model. We explored one potential application of the QM9star dataset by training a GNN to predict potential energy surface of molecular systems with multiple charge states. Existing mainstream 3D GNNs typically only require atomic coordinates and atom types as input, without considering the overall molecular charge and spin multiplicity[11,12,14,15,18,19]. While this setup is acceptable for datasets and applications involving only neutral molecules, it cannot distinguish the effects of charge states, especially when both ions and neutral molecules are present, resulting in challenges in downstream molecular applications. To address this issue, some models encode formal charges and radicals as distinct atom types[51], while others encode the total charge or spin multiplicity as global features[13], both achieving effective progress. In this work, we do not aim to design a new molecular modeling architecture. Instead, we used the DimeNet++[14,15] model implemented by the DIG[58] project and added two embedding layers equivalent to the atomic number embeddings, specifically for encoding formal charges and formal radicals. In simple terms, the node encoding includes atom number, formal charge, and formal radical equally, thereby differentiating various electronic states at each atom.

We randomly divided the entire QM9star dataset of 2 million data points into training, validation, and test sets in a ratio of 8:1:1. The training target was the formation energy of the molecules, defined as the single-point energy of the molecule minus the sum of the single-point energies of all its atoms. The single-point energies of all atoms were calculated under the same level and can be found in the repository. We trained the model for 500 epochs on an NVIDIA RTX4090 GPU, which took approximately 106 hours in total. The Adam optimizer was used during training, with a batch size of 192. The loss function was the mean absolute error (MAE) of the formation energies. The learning rate was initially set to 0.0001 and was halved each time the training began to converge, eventually reaching 3.125×10^{-6}. Figure 5a,b show the loss function and learning rate for the validation set during training. The final model achieved an energy error close to chemical accuracy (MAE = 0.235 kcal/mol). Figure 5c,d illustrate the model's predictive performance on the test set, demonstrating that the model accurately

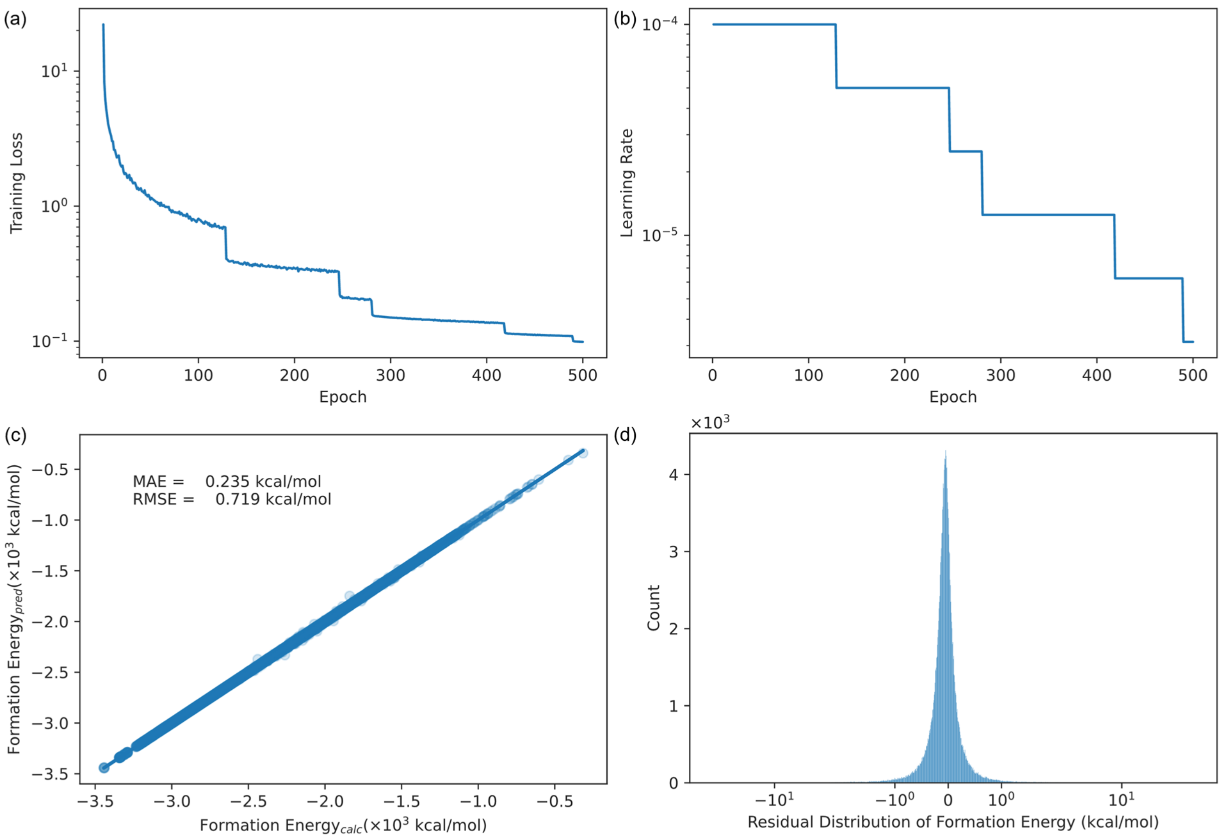


Fig. 5 Formation Energy prediction with DimeNet++ model. (**a**) The evolution of loss function on the training set. The loss function is MAE loss for energies. (**b**) The evolution of learning rate during training. (**c**) The regression and errors on the test set. (**d**) The residual energy (predicted energy minus calculated energy) distribution.

learned the formation energy distribution of molecules with different charge states. This validation indicates that the QM9star dataset contains rich structural and energetic information, enabling the modeling applications of reactive intermediates with various charge states.

Usage Notes

The database dump file for QM9star dataset has been made accessible on the figshare platform[54]. Researchers are encouraged to visit our code repository (https://github.com/gentle1999/qm9star_query) for instructions on how to download and implement the QM9star dataset for their analytical and computational chemistry studies.

Code availability

Scripts for interfacing with and utilizing the QM9star dataset, along with the code for training potential energy models using the QM9star dataset, are provided in our code repository (https://github.com/gentle1999/qm9star_query). These resources facilitate the integration of the dataset into computational frameworks and the development of predictive models within the domain of quantum chemistry.

Received: 19 June 2024; Accepted: 24 September 2024;
Published online: 21 October 2024

References

1. Delaney, J. S. ESOL: Estimating Aqueous Solubility Directly from Molecular Structure. *J. Chem. Inf. Comput. Sci.* **44**, 1000–1005 (2004).
2. Varnek, A. & Baskin, I. Machine Learning Methods for Property Prediction in Chemoinformatics: *Quo Vadis*? *J. Chem. Inf. Model.* **52**, 1413–1437 (2012).
3. Mitchell, J. B. O. Machine learning methods in chemoinformatics. *WIREs Comput Mol Sci* **4**, 468–481 (2014).
4. Watanabe, S. *et al.* High-dimensional neural network atomic potentials for examining energy materials: some recent simulations. *J. Phys. Energy* **3**, 012003 (2021).
5. Behler, J. Four Generations of High-Dimensional Neural Network Potentials. *Chem. Rev.* **121**, 10037–10072 (2021).
6. Kocer, E., Ko, T. W. & Behler, J. Neural Network Potentials: A Concise Overview of Methods. *Annu. Rev. Phys. Chem.* **73**, 163–186 (2022).
7. Yang, Y. *et al.* Quantum mechanical static dipole polarizabilities in the QM7b and AlphaML showcase databases. *Sci Data* **6**, 152 (2019).
8. Xu, L. *et al.* Towards Data-Driven Design of Asymmetric Hydrogenation of Olefins: Database and Hierarchical Learning. *Angew. Chem. Int. Ed.* **60**, 22804–22811 (2021).

9. Li, S.-W., Xu, L.-C., Zhang, C., Zhang, S.-Q. & Hong, X. Reaction performance prediction with an extrapolative and interpretable graph model based on chemical knowledge. *Nat Commun* **14**, 3569 (2023).
10. Xu, L.-C. *et al*. Enantioselectivity prediction of pallada-electrocatalysed C–H activation using transition state knowledge in machine learning. *Nat. Synth* **2**, 321–330 (2023).
11. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural Message Passing for Quantum Chemistry. *Proceedings of the 34th International Conference on Machine Learning* **70**, 1263–1272 (2017).
12. Schütt, K. *et al*. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. in *Advances in Neural Information Processing Systems* **vol. 30** (2017).
13. Unke, O. T. & Meuwly, M. PhysNet: A Neural Network for Predicting Energies, Forces, Dipole Moments, and Partial Charges. *J. Chem. Theory Comput.* **15**, 3678–3693 (2019).
14. Gasteiger, J., Groß, J. & Günnemann, S. Directional Message Passing for Molecular Graphs. in *International Conference on Learning Representations (ICLR)* (2020).
15. Gasteiger, J., Yeshwanth, C. & Günnemann, S. Directional Message Passing on Molecular Graphs via Synthetic Coordinates. *Advances in Neural Information Processing Systems* **34**, 15421–15433 (2021).
16. Park, C. W. *et al*. Accurate and scalable graph neural network force field and molecular dynamics with direct force architecture. *npj Comput Mater* **7**, 73 (2021).
17. Batatia, I., Kovács, D. P., Simm, G. N. C., Ortner, C. & Csányi, G. MACE: Higher Order Equivariant Message Passing Neural Networks for Fast and Accurate Force Fields. *Advances in Neural Information Processing Systems* **35**, 11423–11436 (2022).
18. Liu, Y. *et al*. Spherical Message Passing for 3D Molecular Graphs. in *International Conference on Learning Representations (ICLR)* (2022).
19. Wang, L., Liu, Y., Lin, Y., Liu, H. & Ji, S. ComENet: Towards Complete and Efficient Message Passing for 3D Molecular Graphs. *Advances in Neural Information Processing Systems* **35**, 650–664 (2022).
20. Batatia, I. *et al*. *A foundation model for atomistic materials chemistry*. Preprint at http://arxiv.org/abs/2401.00096 (2024).
21. Martin-Barrios, R., Navas-Conyedo, E., Zhang, X., Chen, Y. & Gulín-González, J. An overview about neural networks potentials in molecular dynamics simulation. *Int J of Quantum Chemistry* **124**, e27389 (2024).
22. Wu, Z. *et al*. MoleculeNet: a benchmark for molecular machine learning. *Chem. Sci.* **9**, 513–530 (2018).
23. Carpenter, K. A. & Huang, X. Machine Learning-based Virtual Screening and Its Applications to Alzheimer's Drug Discovery: A Review. *CPD* **24**, 3347–3358 (2018).
24. Jiang, D. *et al*. Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J Cheminform* **13**, 12 (2021).
25. Axelrod, S., Shakhnovich, E. & Gómez-Bombarelli, R. Excited state non-adiabatic dynamics of large photoswitchable molecules using a chemically transferable machine learning potential. *Nat Commun* **13**, 3440 (2022).
26. Groom, C. R., Bruno, I. J., Lightfoot, M. P. & Ward, S. C. The Cambridge Structural Database. *Acta Crystallogr B Struct Sci Cryst Eng Mater* **72**, 171–179 (2016).
27. Bolton, E. E., Wang, Y., Thiessen, P. A. & Bryant, S. H. PubChem: Integrated Platform of Small Molecules and Biological Activities. in *Annual Reports in Computational Chemistry* vol. 4 217–241 (Elsevier, 2008).
28. Yang, J. D., Xue, X. S., Ji, P., Li, X., & Cheng, J. P. Internet Bond-energy Databank (pKa and BDE): iBonD Home Page. http://ibond.chem.tsinghua.edu.cn or http://ibond.nankai.edu.cn (2022).
29. Blum, L. C. & Reymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *J. Am. Chem. Soc.* **131**, 8732–8733 (2009).
30. Rupp, M., Tkatchenko, A., Müller, K.-R. & Von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
31. Ruddigkeit, L., Van Deursen, R., Blum, L. C. & Reymond, J.-L. Enumeration of 166 Billion Organic Small Molecules in the Chemical Universe Database GDB-17. *J. Chem. Inf. Model.* **52**, 2864–2875 (2012).
32. Montavon, G. *et al*. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **15**, 095003 (2013).
33. Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci Data* **1**, 140022 (2014).
34. Ramakrishnan, R., Hartmann, M., Tapavicza, E. & Von Lilienfeld, O. A. Electronic spectra from TDDFT and machine learning in chemical space. *The Journal of Chemical Physics* **143**, 084111 (2015).
35. Kim, H., Park, J. Y. & Choi, S. Energy refinement and analysis of structures in the QM9 database via a highly accurate quantum chemical method. *Sci Data* **6**, 109 (2019).
36. Nandi, S., Vegge, T. & Bhowmik, A. MultiXC-QM9: Large dataset of molecular and reaction energies from multi-level quantum chemical methods. *Sci Data* **10**, 783 (2023).
37. Liang, J., Xu, Y., Liu, R. & Zhu, X. QM-sym, a symmetrized quantum chemistry database of 135 kilo molecules. *Sci Data* **6**, 213 (2019).
38. Liang, J. *et al*. QM-symex, update of the QM-sym database with excited state information for 173 kilo molecules. *Sci Data* **7**, 400 (2020).
39. Nakata, M. & Shimazaki, T. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry. *J. Chem. Inf. Model.* **57**, 1300–1308 (2017).
40. Nakata, M., Shimazaki, T., Hashimoto, M. & Maeda, T. PubChemQC PM6: Data Sets of 221 Million Molecules with Optimized Molecular Geometries and Electronic Properties. *J. Chem. Inf. Model.* **60**, 5891–5899 (2020).
41. Nakata, M. & Maeda, T. PubChemQC B3LYP/6-31G*//PM6 Data Set: The Electronic Structures of 86 Million Molecules Using B3LYP/6-31G* Calculations. *J. Chem. Inf. Model.* **63**, 5734–5754 (2023).
42. Glavatskikh, M., Leguy, J., Hunault, G., Cauchy, T. & Da Mota, B. Dataset's chemical diversity limits the generalizability of machine learning predictions. *J Cheminform* **11**, 69 (2019).
43. Řezáč, J., Fanfrlík, J., Salahub, D. & Hobza, P. Semiempirical Quantum Chemical PM6 Method Augmented by Dispersion and H-Bonding Correction Terms Reliably Describes Various Types of Noncovalent Complexes. *J. Chem. Theory Comput.* **5**, 1749–1760 (2009).
44. Chen, G. *et al*. Alchemy: A Quantum Chemistry Dataset for Benchmarking AI Models. Preprint at http://arxiv.org/abs/1906.09427 (2019).
45. Isert, C., Atz, K., Jiménez-Luna, J. & Schneider, G. QMugs, quantum mechanical properties of drug-like molecules. *Sci Data* **9**, 273 (2022).
46. Mendez, D. *et al*. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Research* **47**, D930–D940 (2019).
47. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1, A data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci Data* **4**, 170193 (2017).
48. Smith, J. S. *et al*. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci Data* **7**, 134 (2020).
49. Hoja, J. *et al*. QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Sci Data* **8**, 43 (2021).

50. Axelrod, S. & Gómez-Bombarelli, R. GEOM, energy-annotated molecular conformations for property prediction and molecular generation. *Sci Data* **9**, 185 (2022).
51. Eastman, P. *et al*. SPICE, A Dataset of Drug-like Molecules and Peptides for Training Machine Learning Potentials. *Sci Data* **10**, 11 (2023).
52. St. John, P. C. *et al*. Quantum chemical calculations for over 200,000 organic radical species and 40,000 associated closed-shell molecules. *Sci Data* **7**, 244 (2020).
53. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *figshare* https://doi.org/10.6084/m9.figshare.c.978904.v5 (2014).
54. Tang, M., Zhu, T., Zhang, S. & Hong, X. *QM9star, two Million DFT-computed Equilibrium Structures for Ions and Radicals with Atomic Information*. https://doi.org/10.6084/m9.figshare.27002905 (2024).
55. RDKit: Open-source cheminformatics. https://www.rdkit.org/.
56. Fey, M. & Lenssen, J. E. Fast graph representation learning with PyTorch Geometric. in *ICLR workshop on representation learning on graphs and manifolds* (2019).
57. Thompson, J. D., Xidos, J. D., Sonbuchner, T. M., Cramer, C. J. & Truhlar, D. G. More reliable partial atomic charges when using diffuse basis sets. *PhysChemComm* **5**, 117 (2002).
58. Liu, M. *et al*. DIG: A Turnkey Library for Diving into Graph Deep Learning Research. *JMLR* **22**, 1–9 (2021).

Acknowledgements

National Key R&D Program of China (2022YFA1504301, X. H.), National Natural Science Foundation of China (22122109 and 22271253, X. H.; 22103070, S.-Q. Z.), Zhejiang Provincial Natural Science Foundation of China under Grant No. LDQ23B020002 (X. H.), the Starry Night Science Fund of Zhejiang University Shanghai Institute for Advanced Study (SN-ZJU-SIAS-006, X. H.), CAS Youth Interdisciplinary Team (JCTD-2021-11, X. H.), Fundamental Research Funds for the Central Universities (226-2022-00140, 226-2022-00224, 226-2023-00115 and 226-2024-00003, X. H), the State Key Laboratory of Clean Energy Utilization (ZJUCEU2020007, X. H.), the State Key Laboratory of Physical Chemistry of Solid Surfaces (202210, X. H.), the Leading Innovation Team grant from Department of Science and Technology of Zhejiang Province (2022R01005, X. H.), Open Research Fund of School of Chemistry and Chemical Engineering of Henan Normal University (2024Z01, X. H.) are gratefully acknowledged. Calculations were performed on the high-performance computing system at Department of Chemistry, Zhejiang University.

Author contributions

Miao-jiong Tang: Data Curation and Clean, Investigation, Software, Writing – Original Draft. Tian-cheng Zhu: Calculation. Shuo-qing Zhang: Conceptualization, Writing – Review & Editing. Xin Hong: Conceptualization, Resources, Funding Acquisition, Supervision, Writing – Review & Editing.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to S.-Q.Z. or X.H.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.