scientific data

OPEN

DATA DESCRIPTOR

Quantum Topological Atomic Properties of 44K molecules

Brandon Meza-González[1], David I. Ramírez-Palma[2], Pablo Carpio-Martínez[3], David Vázquez-Cuevas[2], Karina Martínez-Mayorga[2] & Fernando Cortés-Guzmán[1]✉

We present a data set of quantum topological properties of atoms of 44K randomly selected molecules from the GDB-9 data set. These atomic properties were obtained as defined by the quantum theory of atoms in molecules (QTAIM) within an atomic basin, a region of real space bounded by zero-flux surfaces in the electron density gradient vector field. The wave function files were generated through DFT static calculations (B3LYP/6-31G), and the atomic properties were calculated using QTAIM. The calculated atomic properties include the energy of the atomic basin, the electronic population, the magnitude of the total dipole moment, and the magnitude of the total quadrupole moment. The atomic properties allow one to understand the chemical structure, reactivity, and molecular recognition. They can be incorporated into force fields for molecular dynamics or for predicting reactive sites. We believe that this data set could facilitate new studies in chemical informatics, machine learning applied to chemistry, and computational molecular design.

Background & Summary

The current use of artificial intelligence (AI) results from the combined availability of databases, algorithms, and computing power. Interest has emerged from every corner of scientific research. However, knowing the strengths and limitations of the data and methods gives a proper perspective of what to expect. Molecular databases have been constructed and utilized in Chemistry and Biology for several decades. Thousands of descriptors can be automatically calculated with commercial and open-source software packages. Descriptors based on two-dimensional representations of the molecules are independent of geometry and conformation and are easy to calculate. In turn, descriptors obtained from the three-dimensional representations of the molecules are commonly obtained from energy-minimized structures using molecular mechanics (MM) or quantum mechanics (QM) methods. The descriptors obtained from the MM and QM methods have also been developed for decades[1]. The structures given from the QM calculations allow us to obtain the distribution of electrons within the molecules, which can be used to develop models to predict reactivity[2]. The description of molecules (and atoms) with quantum chemical topological properties allows for the analysis of the reactivity and spectral properties

The quantum theory of atoms in molecules (QTAIM)[3], developed by Bader, defines an atom within a molecule and allows one to calculate atomic properties based on the electron density topology. In QTAIM, an atom is defined as a region of real space bounded by zero-flux surfaces in the electron density gradient vector field called an atomic basin[4]. An atomic property $P(\Omega)$ of an atom, i.e., atomic property (AP), is defined as the expectation value of an effective single-particle density equation (1) in its atomic basin Ω. Two essential properties of the AP are additivity and transferability due to the recoverability of the functional group and the total molecular quantities. One important feature of the QTAIM partition is the recovery of the harpoon mechanism of electron transfer in forming molecules like LiF[5].

$$P(\Omega) = \int_{\Omega} \hat{P}_{\Omega} \rho\ (\mathbf{r}) d\tau \tag{1}$$

The partition of molecular properties into their atomic contributions helps us to understand the chemical structure, reactivity, and molecular recognition. Furthermore, physical and chemical molecular properties can be predicted using AP as descriptors. Several groups around the world have used molecular features and machine learning methods to predict AP, mostly atomic charges and moments, using QTAIM, Hirshfeld, or molecular

[1]Facultad de Química, Universidad Nacional Autónoma de México, Ciudad de Méxinclude thexico, Mexico City, Mexico. [2]Instituto de Química, Unidad Mérida, Universidad Nacional Autónoma de México, Mérida, Yucatán, Mexico. [3]Centro Conjunto de Investigación en Química Sustentable UAEM-UNAM, Carretera Toluca-Atlacomulco, km. 14.5, Toluca, Estado de México, C.P. 50200, Mexico. ✉e-mail: fercor@unam.mx

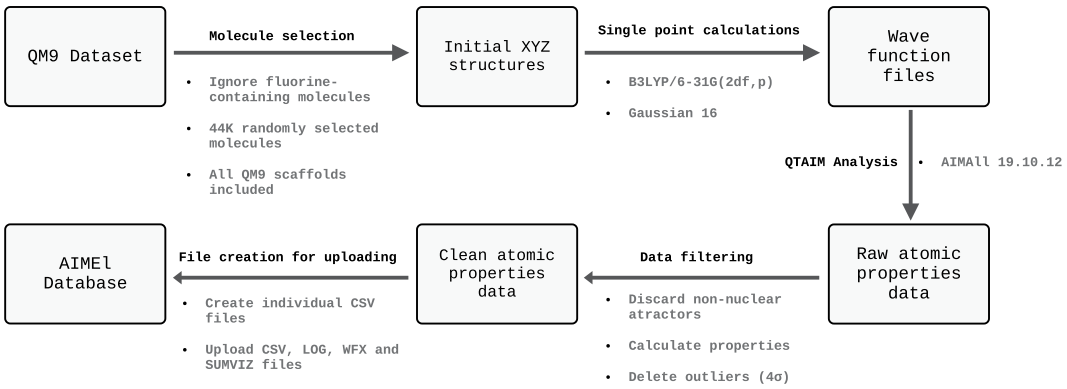

Fig. 1 Overview of the data generation process. We carefully selected 44K molecules from the GDB-9 (QM9) dataset. These chosen molecules were randomly extracted, and we purposely excluded any containing fluorine, and included all QM9 scaffolds. We then put these molecules through a thorough process that involved Cartesian coordinate extraction, single–point calculations, and meticulous refinement. This included the integration of atomic properties using the AIMALL package and utilizing Python scripts to organize them into CSV files to create the AIMEl database.

orbital partitions[6–12]. Hirshfeld and Molecular orbital partitioning schemes provide useful atomic information with moderate computational demands. As they are cheaper methods, they can be more easily used for faster data generation and ML processing. On the other hand, the QTAIM partitioning method represents the most detailed scheme. Detailed electron density and atomic interaction analysis offer meaningful information that can be used in more specific ML procedures, but at a higher computational cost. This makes it less feasible for very large datasets. It is for this reason that the AIMEl dataset represents a significant impact in ML model and force field generation processes. In this way, predicted AP can be included in the force fields used in molecular dynamics simulations[13] or to predict reactive sites[14]. This is especially important since several studies on predicting chemical reactivity use machine learning tools. Examples include predicting physical properties by considering molecules as graphs in convolutional neural networks[15,16], property prediction using deep neural networks[6], as well as predicting the products of chemical reactions[17]. In these works, the representation of molecules is derived from structural information: the chemical environment, atomic bonds in a molecule, and other physicochemical properties. Consequently, a database containing AP information based on electronic density is invaluable. Moreover, the understanding that these properties indicate an atom's reactivity within a molecule further enriches the data for training cutting-edge machine learning algorithms. However, the training data sets associated with the earlier studies are not publicly available. Calculating public data sets of atomic properties is essential to train models using machine learning approaches. This paper presents a public data set of the quantum topological properties of atoms within 44K molecules randomly selected from the GDB-9 data set[18]. The data set includes electrostatic atomic properties, such as atomic charge and moments, based on our working hypothesis that the reactivity of an atom depends on its electron population and the way it polarizes within the atomic basin, described by atomic electrostatic moments[14]. There are diverse datasets widely used in benchmarking, molecular discovery, and reactivity studies. Table 1 shows some of these databases. Although some sets, such as PubChemQC[19] and ANI-1[20], provide quantum chemical calculations for organic molecules, there is no public data set that contains data on atomic properties. On the other hand, data sets such as Pistachio[21] and ORD[22] contain many reactions published in patent databases. Nevertheless, they do not present reactivity descriptors based on atomic properties. This highlights the relevance of AIMEl as a valuable data set for advancing the understanding of chemical reactivity.

Methods

Figure 1 illustrates the data acquisition process. We subtracted 44K molecules from the 134K molecules contained in the GDB-9 data set, also called QM9[18], (https://springernature.figshare.com/collections/Quantum_chemistry_structures_and_properties_of_134_kilo_molecules9789044), each molecule contains up to nine atoms (CONF) without considering hydrogen atoms. This subset of 44K molecules was randomly chosen, excluding those containing fluorine atoms. Furthermore, a scaffold analysis of the QM9 database was conducted to include at least one molecule from each Murcko scaffold[23]. This approach ensures that the AIMEl database encompasses the structural diversity of the QM9 set.

The Cartesian coordinates of the molecules were extracted as *XYZ* files. Single-point energy calculations were performed at the theoretical level of B3LYP / 6-31G (2df, p) to obtain molecular orbitals using Gaussian 16[24]. The base set and the functional combination are the same as those used to build the GDB-9 database. The atomic properties were then integrated using the AIMAll package[25]. Molecules that presented spurious non-nuclear attractors were discarded. Outliers were removed on the basis of four atomic properties (See Table 2). The atomic properties were extracted from the AIMAll output files and collected in CSV files. The characterization of the database was performed using in-house-written Python scripts.

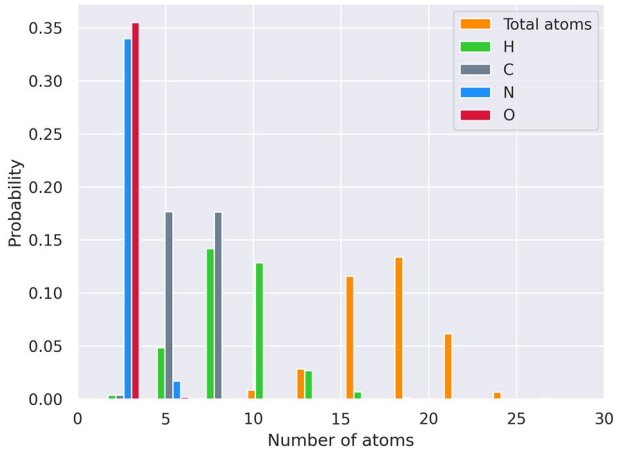

Fig. 2 Distribution of the chemical elements constituting the database. The bins have been adjusted to ensure that the total sum of bar heights equals 1.

Atomic Properties. We focussed on four APs, notably the energy $E(\Omega)$, population ($N(\Omega)$), dipole moment ($\mu(\Omega)$) and quadrupole moment ($Q(\Omega)$). $E(\Omega)$ is obtained within the virial approach as a scaled atomic kinetic energy, $E(\Omega) = (1-\gamma)T(\Omega)$, where γ is the virial ratio between the molecular potential energy (V) and the kinetic energy (T), that is, $\gamma = V/T$[26]. The atomic population is simply the average number of electrons obtained by integration of $\rho(\mathbf{r})d\boldsymbol{\tau}$ into the atomic basin Ω (see Equation (2)). The atomic dipole moment $\mu(\Omega)$ measures the extent and direction of displacement between the centroid of the negative charge and the position of the nucleus. Its components along the coordinates x, y, and z are represented as $\mu_\alpha(\Omega)$ with directions $\alpha = x/y/z$ (see Equation (3)) while the corresponding magnitude is denoted as $|\mu(\Omega)|$ (see Equation (4)). In turn, the quadrupole moment tensor comprises the matrix elements $Q_{\alpha\beta}(\Omega)$ (see equation (5)) with directions $\alpha = x/y/z$ and $\beta = x/y/z$. Its subsequent diagonalization leads to quadrupole moments along a specific direction, i.e., $Q_{xx}(\Omega)$, $Q_{yy}(\Omega)$, and $Q_{zz}(\Omega)$, which measure the increase or decrease of the electronic density along a specific axis. Thus, if any of the diagonal elements of the quadrupole moment tensor is lower than zero, it implies a concentration of the electronic density on that axis. Equation (6) shows the magnitude of the quadrupole moment. Both quantities, $|\mu_\alpha(\Omega)|$ and $|Q_{\alpha\beta}(\Omega)|$, describe the deviation of the electron density relative to a spherically symmetric electron distribution.

$$N(\Omega) = \int_\Omega \rho(\mathbf{r})d\boldsymbol{\tau} \tag{2}$$

$$\mu_\alpha(\Omega) = -\int_\Omega \mathbf{r}_\Omega^\alpha \rho(\mathbf{r})d\boldsymbol{\tau} \tag{3}$$

$$|\mu(\Omega)| = \sqrt{\mu_x^2 + \mu_y^2 + \mu_z^2} \tag{4}$$

$$Q_{\alpha\beta}(\Omega) = -\int_\Omega \mathbf{r}_\Omega^\alpha \mathbf{r}_\Omega^\beta \rho(\mathbf{r})d\boldsymbol{\tau} \tag{5}$$

$$|Q(\Omega)| = \sqrt{\frac{2}{3}\left(Q_{xx}^2 + Q_{yy}^2 + Q_{zz}^2\right)} \tag{6}$$

Data Records

The entire AIMEl database, including atomic coordinates and properties of each molecule, is publicly accessible in a Zenodo repository[27]. Input and output files for Gaussian 16 as well as output files from AIMAll software are provided. Data set `aimel_merged_44k.csv` comprises 802,870 rows representing individual atoms, collectively forming 44,470 molecules. Each row is identified by the file column, which matches the index in the GDB-9 dataset[18]. Additionally, the database provides Cartesian coordinates for all atoms within each molecule, along with their respective atomic properties. The median number of atoms for each molecule in the database is 18, while the medians for the other atoms are $H = 9$, $C = 7$, $N = 1$, and $O = 1$. In particular, the predominant atomic composition in most molecules includes one nitrogen atom and one oxygen atom. The general distribution is visualized in Fig. 2.

File format. The `aimel_merged_44k.csv` file contains column names as headers. The columns included in this file are presented in Table 2. The subsequent rows following the header contain the properties of the atoms within the molecule. The first column is reserved for the molecule index, aligned with the GDB-9 data set. The

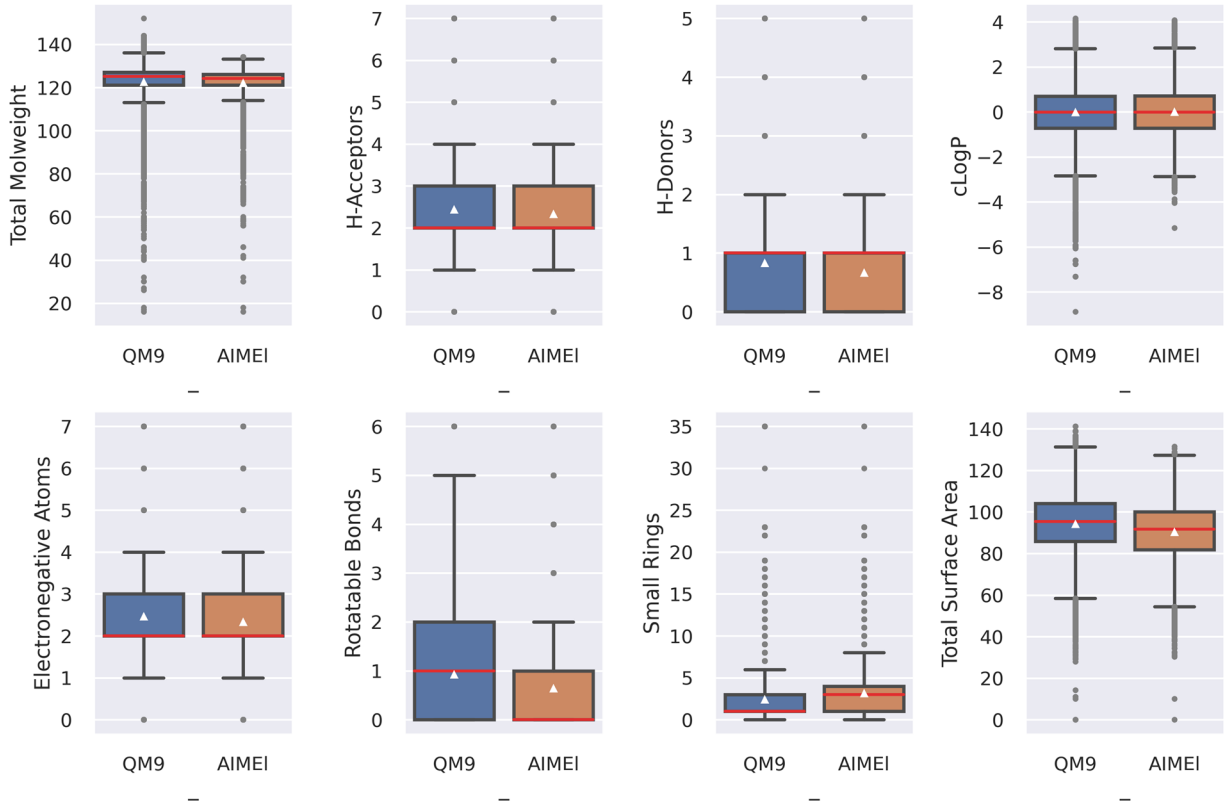

Fig. 3 Comparison of molecular properties between QM9 and AIMEl databases.

Database	Number of molecules	Molecules type	Method	Annotations	Goal
QM9[32]	134 K	Diverse	B3LYP/6-31G(2df,p)	Includes 15 physical chemistry properties	Benchmarking
ANI-1[20]	20M	Diverse	ωB97X/6-31G(d)	Off-equilibrum conformations for organic molecules	Fitting ML potentials
PubChemQC[19]	86M	Diverse	PM6//B3LYP/6-31G*	Includes electronic structure outputs	Molecular discovery
QMugs[33]	665K	Druglike	ωB97X-D/def2-SVP	Includes electronic structure and spectroscopical data	Bioactivity
Pistachio[21]	13.3M	Diverse	Experimental	Focused on reactions extracted from USPTO, EPO and WIPO patents.	Reactivity
ORD[22]	2M	Diverse	Experimental	Open-acces project, focused on reactions from USPTO and contributions from users	Reactivity
USPTO-50K[34]	50K	Diverse	Experimental	50K randomly selected reactions from USPTO, classified into 10 reaction classes	Reactivity
AIMEl-DB[27]	40K	Diverse	B3LYP/6-31G(2df,p)//QTAIM	Includes atomic properties using QTAIM theory	Reactivity

Table 1. Comparison of some popular databases.

second column denotes the name of each atom in the format "*Xn*," where *X* represents the element name and *n* represents the sequential index of the atom within the molecule. Following these columns, three columns present Cartesian coordinates for each atom. Subsequently, four columns encompass electronic properties, including population ($N(\Omega)$), dipole moment magnitude ($|\mu(\Omega)|$), quadrupole moment magnitude ($|Q(\Omega)|$), and atomic energy ($E(\Omega)$). (See Table 2 for more details). Furthermore, for each of the 44,470 molecules, four types of files are provided: G16 input `.com`, output `.log`, and wave function `.wfx` files; and AIMAll output `.sumviz` files. These files were no longer processed.

Technical Validation

Comparison to QM9 database. Since the AIMEl database described here is a subset of the QM9 database, validation is carried out by comparing typical molecular properties: total mole weight, count of H acceptors, H donors, electronegative atoms, rotatable bonds and small rings, partition coefficient between n-octanol and water cLogP, and total surface area. The analysis is presented in Fig. 3. The median and mean values are represented in red lines and white triangles. Since the AIMEl subset was randomly selected, substantial overlap was expected compared to the QM9 property set. Interestingly, the data sets present nearly identical median and mean values

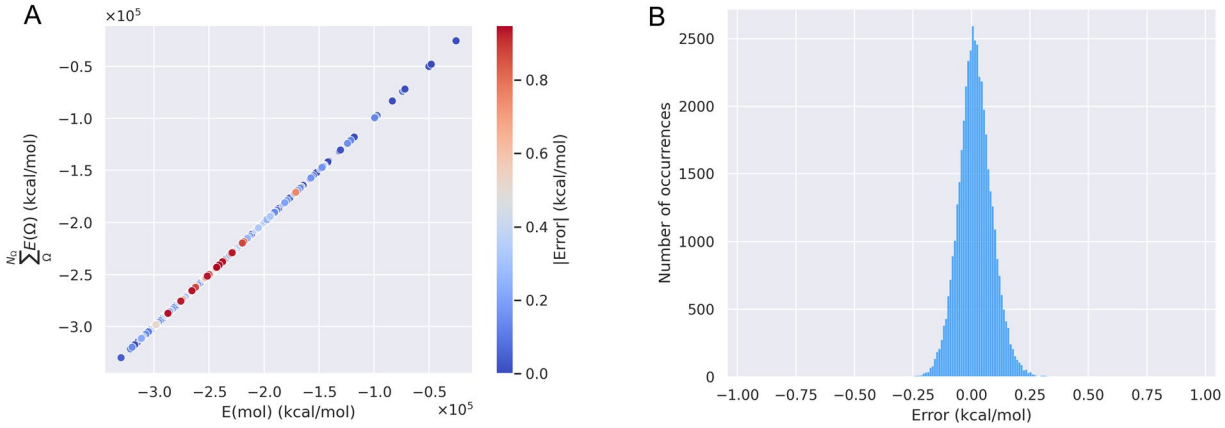

Fig. 4 Error distributions for the molecules in the AIMEl dataset. The absolute error is presented in (**A**). Here, molecules with larger energy differences appear, ranging from −3000.0 to −2000.0 kcal/mol. In (**B**), the distribution shows that the highest number of occurrences oscillates around 0.00 kcal/mol.

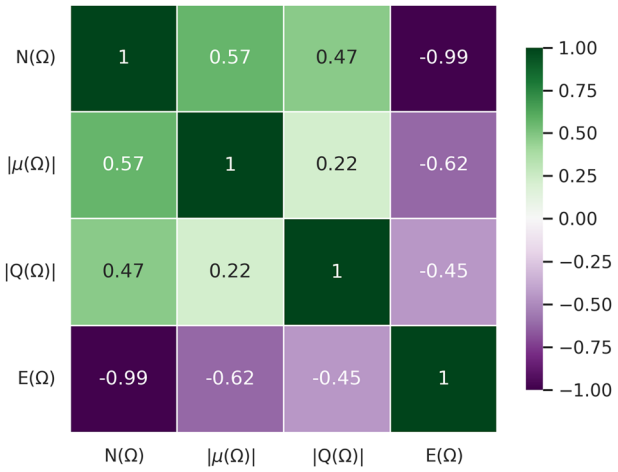

Fig. 5 Correlation matrix between atomic properties presented in the database.

Column name	Units	Content
file	—	Molecule index sourced from the GDB-9 dataset
a_name	—	Atom name in Xn format
position_x	Å	Cartesian *x* coordinate
position_y	Å	Cartesian *y* coordinate
position_z	Å	Cartesian *z* coordinate
N	*e*	Atomic Population
|mu|	ea_0	Magnitude of the total dipole moment
|Q|	ea_0^2	Magnitude of the total quadrupole moment
E	E_h	Atomic Energy

Table 2. Atomic parameters and properties in the `aimel_merged_44k.csv` file.

for most properties. In turn, the number of rotatable bonds is the most notable discrepancy. Although the mean values are very close, the median in AIMEl is zero, while a value of one is presented in QM9. However, the AIMEl data set shows diverse structures that feature molecules with two or more rotatable bonds. The original QM9 database included unstable structures. As a first filter, only chemically sound molecules were maintained. For instance, non-bonded atoms and overly strained ring structures were eliminated. This refinement led to the elimination of 13, 402 molecules, leaving a total of 45, 900. However, after eliminating the molecules that included NNA, the data set comprises 44,470 molecules.

Property	MAE	RMSE
$N(\Omega)$	0.011	0.029
$\lvert\mu(\Omega)\rvert$	0.076	0.215
$\lvert Q(\Omega)\rvert$	0.014	0.319
$E(\Omega)$	0.007	0.033

Table 3. Comparison of Mean Absolute Error (MAE) and Root Mean Squared Error (RMSE) of QTAIM properties between the wB97XD/6–31G(2df, p) and B3LYP/6–31G(2df, p) levels of theory for a subset of 4,397 molecules.

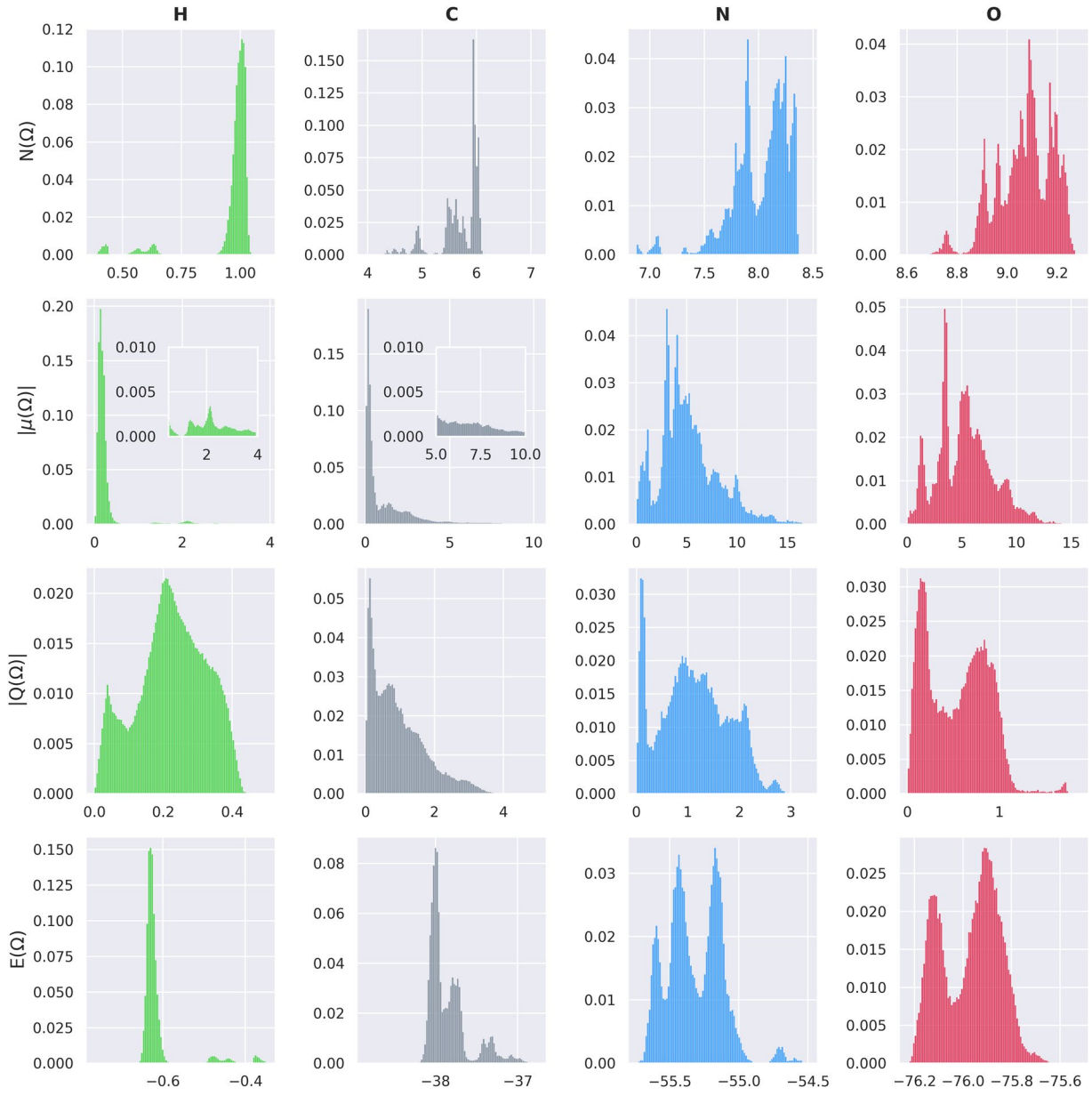

Fig. 6 Histograms of properties. Atomic population, $N(\Omega)$; Magnitude of the total dipole moment, $|\mu(\Omega)|$; Magnitude of the total quadrupole moment, $|Q(\Omega)|$; d) Atomic Energy, $E(\Omega)$ The bins have been adjusted to ensure that the total sum of bar heights equals 1.

Validation of QTAIM calculations. To check the integrity of the data generated using AIMAll, we performed an error analysis for the energy of the system. We compare the total molecular energy E(mol) obtained from the calculation of the electronic molecular structure with the sum of the atomic energies for each molecule, $\sum_{\Omega}^{N_\Omega} E(\Omega)$. N_Ω represents the total number of atoms. The difference between E(mol) and $\sum_{\Omega}^{N_\Omega} E(\Omega)$ shows the quality of the atomic integration process. For this reason, the error $(\mathrm{E(mol)} - \sum_{\Omega}^{N_\Omega} E(\Omega))$ is a useful quantity to validate the

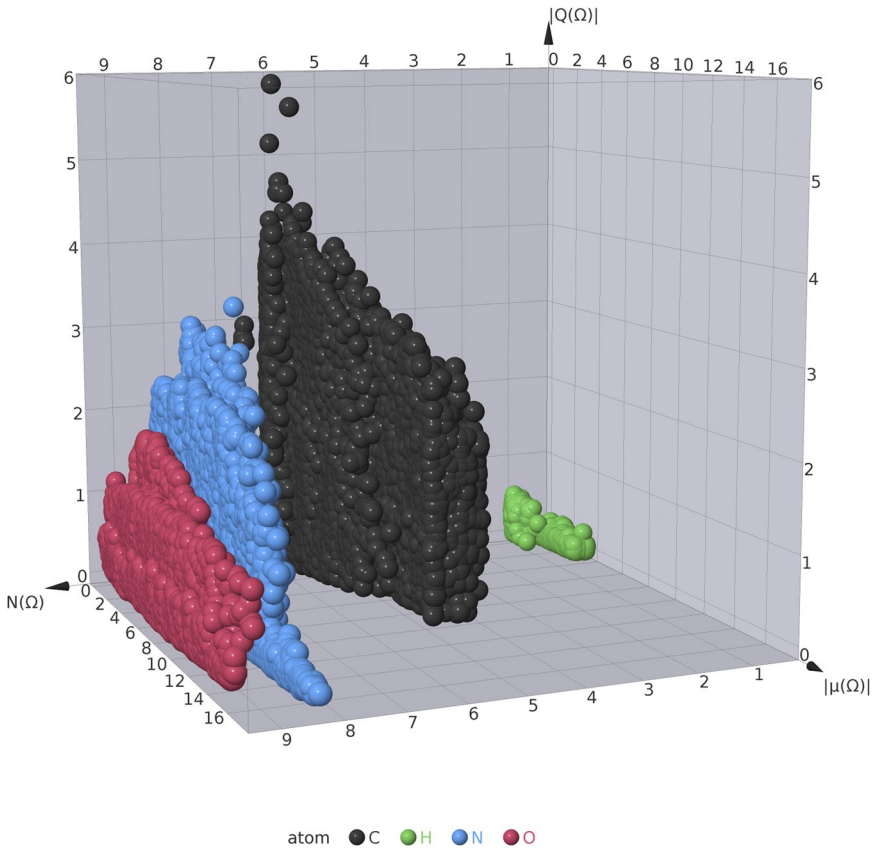

Fig. 7 3D distribution of the calculated atomic properties, grouped by atom kind. All properties are presented in atomic units.

calculated atomic properties. In Fig. 4 the error distributions are presented. Figure 4A shows a direct comparison between E(mol) and $\sum_{\Omega}^{N_{\Omega}} E(\Omega)$. The results reveal that molecules with larger errors (>|0.80| kcal/mol) are observed within the range of −3000.0 to −2000.0 kcal/mol. The complete collection of molecules contained in the AIMEl data set presents error values below 1.0 kcal/mol. These errors follow a normal distribution, as shown in Fig. 4B.

S. Senthil *et al.*[28] have studied the lack of chemical sense in molecules within the QM9 dataset. They found that the use of ωB97XD/6-31G(2df, p) leads to geometrically stable structures. In this work, we have found that using a $\overline{x} \pm 4\sigma$ approximation refines the chemical structures, filtering out molecules with geometric instabilities. Besides, to compare our results with this level of theory, we carried out single-point calculations on a validation subset of 4,397 molecules and obtained their QTAIM descriptors. The results are shown in Table 3. There are no meaningful discrepancies in this comparison. Although more significant differences are observed in the magnitude of the dipole and quadrupole moments, the metrics for atomic population and energy are close. The higher differences in the dipole and quadrupole magnitudes can be attributed to the nature of these properties, as a small perturbation in electronic density can lead to a significant change in those properties.

Database characterization. Figure 5 presents the correlation matrix between the atomic properties. This matrix shows a strong correlation between atomic energy $E(\Omega)$ and the atomic population $N(\Omega)$. This relationship is interesting because we can obtain the energy of an atomic basin considering only its atomic population; that is, the stabilization of the atom correlates directly with its electronic population. The weakest correlations occur with the quadrupole moment, $|Q(\Omega)|$ followed by $|\mu(\Omega)|$, which implies that these properties can be used to characterize the distribution of the data set.

We present the normalized histograms of the properties studied for atoms *H*, *C*, *O*, and *N* in Fig. 6. The values are presented by filtering the entire database, where each property of each atom is selected using a $\overline{x} \pm 4\sigma$ approximation. $\overline{x}$ and σ represent the mean and standard deviation of the sample. In the case of the atomic population, *H* and *C* present many instances centered around 1 and 6 a.u. corresponding to their respective atomic numbers. In contrast, for *O* and *N*, the distributions are broader and skewed toward larger atomic populations, which could be related to their electronegative nature. Similarly, lower atomic dipoles are observed for *H* and *C*, for which most cases range from 0 to 5 a.u. For the case of *N* and *O*, the atomic dipoles span a wider range of values (0 to 15 a.u.). This observation can be attributed to the generally higher reactivity of the *N* and *O* atoms[29]. Concerning the atomic quadrupole moment, the histograms span different range values depending on the atom, and no clear structural distributions are observed. However, in the cases of *N* and *O*, the values exhibit distinct regions in the broader range, reflecting the possible diversity of structures within the database. Finally, the atomic energies briefly resemble the population distributions. Hydrogen and carbon are centered around

approximately −0.62 Ha and −38.0 Ha, respectively, accounting for the small spread in population values for these two atoms. Nitrogen and oxygen exhibit a more diverse distribution with at least two observed peaks, ranging from approximately −55.5 to −54.5 Ha for *N* and approximately −76.2 to −75.6 Ha for *O*.

Finally, a three-dimensional visualization is presented in Fig. 7, which shows three electronic properties: $N(\Omega)$, $|Q(\Omega)|$, and $|\mu(\Omega)|$. The colors represent atom types. As expected, the atomic population groups atoms by kind. In this regard, $N(\Omega)$ can characterize atoms within a molecule, since there is no wide data distribution. In contrast, the dipole and quadrupole magnitudes show a wider distribution, which captures the diversity of atoms in each group. Therefore, $|Q(\Omega)|$ and $|\mu(\Omega)|$ can be used to characterize the diversity of the database and illustrate the broad reactivity of the analyzed molecules.

This study introduces a novel data set of atomic properties for approximately 44K organic molecules. The data provide fundamental information on the atomic properties based on the Quantum Theory of Atoms in Molecules (QTAIM). In particular, the data set includes atomic basin energies, populations, dipole moments, and quadrupole moments. The data set can enable powerful new machine-learning models for predicting atomic properties and chemical reactivity directly from molecular structure. The public availability of this large database could facilitate new studies in chemical informatics, machine learning applied to chemistry, and computational molecular design.

Code availability

The processing of molecular structures and input generation was performed using OpenBabel[30]. All analyzes were performed using the Python programming language, version 3.9.13. Datawarrior was used for the calculation of molecular properties and chemical space visualization[31]. All quantum mechanical calculations were performed with Gaussian 16[24]. The QTAIM analysis was performed using the AIMAll package[25]. No custom code was generated for this work.

Received: 13 February 2024; Accepted: 29 July 2024;
Published online: 29 August 2024

References

1. Karelson, M., Lobanov, V. S. & Katritzky, A. R. Quantum-chemical descriptors in qsar/qspr studies. *Chemical reviews* **96**, 1027–1044 (1996).
2. Huang, B. & Von Lilienfeld, O. A. Ab initio machine learning in chemical compound space. *Chemical reviews* **121**, 10001–10036 (2021).
3. Bader, R. & Bader, R.Atoms in Molecules: A Quantum Theory. International series of monographs on chemistry, https://books.google.com.mx/books?id=up1pQgAACAAJ (Clarendon Press, 1990).
4. Bader, R. Quantum topology of molecular charge distributions. iii. the mechanics of an atom in a molecule. *The Journal of Chemical Physics* **73**, 2871–2883 (1980).
5. Hernández-Trujillo, J. & Bader, R. F. Properties of atoms in molecules: atoms forming molecules. *The Journal of Physical Chemistry A* **104**, 1779–1794 (2000).
6. Gallegos, M., Vassilev-Galindo, V., Poltavsky, I., Martín Pendás, Á. & Tkatchenko, A. Explainable chemical artificial intelligence from accurate machine learning of real-space chemical descriptors. *Nature Communications* **15**, 4345 (2024).
7. Handley, C. M., Hawe, G. I., Kell, D. B. & Popelier, P. L. Optimal construction of a fast and accurate polarisable water potential based on multipole moments trained by machine learning. *Physical Chemistry Chemical Physics* **11**, 6365–6376 (2009).
8. Fletcher, T. L., Davie, S. J. & Popelier, P. L. Prediction of intramolecular polarization of aromatic amino acids using kriging machine learning. *Journal of chemical theory and computation* **10**, 3708–3719 (2014).
9. Zubatyuk, R., Smith, J. S., Leszczynski, J. & Isayev, O. Accurate and transferable multitask prediction of chemical properties with an atoms-in-molecules neural network. *Science advances* **5**, eaav6490 (2019).
10. Duarte, L. J. & Bruns, R. E. Qtaim atomic charge and polarization parameters and their machine-learning transference among boron-halide molecules. *The Journal of Physical Chemistry A* **124**, 3407–3416 (2020).
11. Gallegos, M., Guevara-Vela, J. M. & Pendás, Á. M. Nnaimq: A neural network model for predicting qtaim charges. *The Journal of Chemical Physics* **156** (2022).
12. Knight, E. T. & Allen, L. C. Rotational barriers originate from energy changes in individual atoms. *Journal of the American Chemical Society* **117**, 4401–4402 (1995).
13. Symons, B. C., Bane, M. K. & Popelier, P. L. Dl_fflux: a parallel, quantum chemical topology force field. *Journal of Chemical Theory and Computation* **17**, 7043–7055 (2021).
14. Ramírez-Palma, D. I., García-Jacas, C. R., Carpio-Martínez, P. & Cortés-Guzmán, F. Predicting reactive sites with quantum chemical topology: carbonyl additions in multicomponent reactions. *Physical Chemistry Chemical Physics* **22**, 9283–9289 (2020).
15. Coley, C. W., Barzilay, R., Green, W. H., Jaakkola, T. S. & Jensen, K. F. Convolutional embedding of attributed molecular graphs for physical property prediction. *Journal of chemical information and modeling* **57**, 1757–1772 (2017).
16. Coley, C. W. *et al.* A graph-convolutional neural network model for the prediction of chemical reactivity. *Chemical science* **10**, 370–377 (2019).
17. Coley, C. W., Barzilay, R., Jaakkola, T. S., Green, W. H. & Jensen, K. F. Prediction of organic reaction outcomes using machine learning. *ACS central science* **3**, 434–443 (2017).
18. Ramakrishnan, R., Dral, P., Dral, P. O., Rupp, M. & Anatole von Lilienfeld, O. Quantum chemistry structures and properties of 134 kilo molecules https://springernature.figshare.com/collections/Quantum_chemistry_structures_and_properties_of_134_kilo_molecules/978904/4 (2014).
19. Nakata, M. & Maeda, T. Pubchemqc b3lyp/6-31g*//pm6 data set: The electronic structures of 86 million molecules using b3lyp/6-31g* calculations. *Journal of Chemical Information and Modeling* **63**, 5734–5754, https://doi.org/10.1021/acs.jcim.3c00899 (2023).
20. Smith, J. S., Isayev, O. & Roitberg, A. E. Ani-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules. *Scientific Data* **4**, 170193, https://doi.org/10.1038/sdata.2017.193 (2017).
21. Next Move Software. Pistachio, reaction data, querying and analytics.
22. Kearnes, S. M. *et al.* The open reaction database. *Journal of the American Chemical Society* **143**, 18820–18826, https://doi.org/10.1021/jacs.1c09820 (2021).
23. Landrum, G. *et al.* rdkit/rdkit: 2024_03_3 (q1 2024) *Zenodo* https://doi.org/10.5281/zenodo.11396708 (2024).
24. Frisch, M. J. *et al.* Gaussian ~ 16 Revision C.01 (2016). Gaussian Inc. Wallingford CT.
25. Keith, T. A. Aimall (version 19.10.12) (2019). TK Gristmill Software, Overland Park KS, USA, aim.tkgristmill.com.

26. Cortés-Guzmán, F. & Bader, R. Transferability of group energies and satisfaction of the virial theorem. *Chemical physics letters* **379**, 183–192 (2003).
27. Meza-González, B. *et al*. AIMEl-DB: Atomic Properties for 44K small organic molecules *Zenodo* https://doi.org/10.5281/zenodo.11406726 (2024).
28. Senthil, S., Chakraborty, S. & Ramakrishnan, R. Troubleshooting unstable molecules in chemical space. *Chem. Sci.* **12**, 5566–5573, https://doi.org/10.1039/D0SC05591C (2021).
29. Robert, J. & Ouellette, J. D. R. Organic Chemistry: Structure, Mechanism, and Synthesis, http://gen.lib.rus.ec/book/index.php?md5=CF08DB61306D6CCEFE111EAF5CDA32EA (Elsevier, 1 edn. 2014).
30. O'Boyle, N. M. *et al*. Open babel: An open chemical toolbox. *Journal of Cheminformatics* **3**, 33, https://doi.org/10.1186/1758-2946-3-33 (2011).
31. Sander, T., Freyss, J., von Korff, M. & Rufener, C. Datawarrior: An open-source program for chemistry aware data visualization and analysis. *Journal of Chemical Information and Modeling* **55**, 460–473, https://doi.org/10.1021/ci500588j (2015).
32. Ramakrishnan, R., Dral, P. O., Rupp, M. & Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific data* **1**, 1–7 (2014).
33. Isert, C., Atz, K., Jiménez-Luna, J. & Schneider, G. Qmugs, quantum mechanical properties of drug-like molecules. *Scientific Data* **9**, 273, https://doi.org/10.1038/s41597-022-01390-7 (2022).
34. Schneider, N., Stiefl, N. & Landrum, G. A. What's what: The (nearly) definitive guide to reaction role assignment. *Journal of Chemical Information and Modeling* **56**, 2336–2346, https://doi.org/10.1021/acs.jcim.6b00564 (2016).

Acknowledgements

The authors thank DGTIC-UNAM (LANCAD-UNAM-DGTIC-194) for the computer time and CONAHCyT (CF2019-1561802/2020) and DGAPA-UNAM (IN207822) for financial support. BMG also thanks CONAHCyT for the financial support (Grant 660455). DIRP and PCM thank DGAPA-UNAM for the postdoctoral fellowship.

Author contributions

F.C.G. conceived the idea. B.M.G., D.I.R.P., and P.C.M. performed the quantum-chemical calculations. B.M.G., P.C.M., .D.I.R.P., K.M.M., and F.C.G. analyzed the results. B.M.G. and D.V.C. coded the scripts. All authors wrote and reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to F.C.-G.

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.