


OPEN

DATA DESCRIPTOR

QM-sym, a symmetrized quantum chemistry database of 135 kilo molecules

Jiechun Liang¹, Yanheng Xu², Rulin Liu² & Xi Zhu^{1,2*}

Applying deep learning methods in materials science research is an important way of solving the time-consuming problems of typical *ab initio* quantum chemistry methodology, but due to the size of large molecules, large and uncharted fields still exist. Implementing symmetry information can significantly reduce the calculation complexity of structures, as they can be simplified to the minimum symmetric units. Because there are few quantum chemistry databases that include symmetry information, we constructed a new one, named QM-sym, by designing an algorithm to generate 135k organic molecules with the C_n symmetry composite. Those generated molecules were optimized to a stable state using Gaussian 09. The geometric, electronic, energetic, and thermodynamic properties of the molecules were calculated, including their orbital degeneracy states and orbital symmetry around the HOMO-LUMO. The basic symmetric units were also included. This database provides consistent and comprehensive quantum chemical properties for structures with C_n symmetries. QM-sym can be used as a benchmark for machine learning models in quantum chemistry or as a dataset for training new symmetry-based models.

Background & Summary

Designing novel molecules and structures with specific physicochemical properties is attractive for researchers, and many methodologies are focused on this field. Among these methodologies, high-throughput screening has been proposed as the most straightforward approach¹, but this method came with a presumption that all the approximations and assumptions made for the adopted modelling techniques are applicable for all stable structures in the entire chemical space². Many quantum chemistry databases have been constructed and reported, including QM7^{3,4}, QM7b⁵, and QM9⁶. Among these databases, the QM9 database, which includes the first 134k molecules of the chemical universe GDB-17 database, is the most widely used one in chemistry deep learning applications. All molecules in QM9 are reported along with 15 properties obtained at the B3LYP/6-31G(2df,p) level of theory to reach a higher accuracy compared to experimental values⁷, including the primary energies, enthalpy, and bandgap. The inclusion of these properties makes the database suitable for the small and basic *de novo* design of new molecules, but the molecules in QM9 consist of at most 9 heavy atoms, which makes the molecule size too small for predicting large molecules, including proteins and polymers. Moreover, there is no symmetry information inside the database. Some essential properties, such as the point group, orbital degeneracy, and selection rules for excitation, are also missing, which makes it impossible to derive excitation events from the QM9 database. In addition, many unstable computer-generated structures containing long N-N chains are included. Due to their low stability and high endothermic properties, long N-N chains tend to decompose and eliminate N_2 ⁸. To identify all of these molecules, we selected all the molecules in the QM9 database with nitrogen chains of more than two nitrogen atoms. All these suspiciously unstable molecules are available in the Supplementary Information, along with their xyz files.

In this work, we provide a dataset of larger symmetrical structures⁹. The benefits of the implemented symmetric properties include considerably reducing the *ab initio* complexity, and the new database offers the possibility of symmetry recognition by the deep learning architecture to construct a connection between a basic symmetric unit and a conventional structure. Our symmetrical database (QM-sym) includes 135 k organic structures with

¹Shenzhen Institute of Artificial Intelligence and Robotics for Society (AIRS), 2001 Longxiang Road, Longgang District, Shenzhen, Guangdong, 518172, China. ²School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen, 2001 Longxiang Road, Longgang District, Shenzhen, Guangdong, 518172, China. *email: zhuxi@cuhk.edu.cn

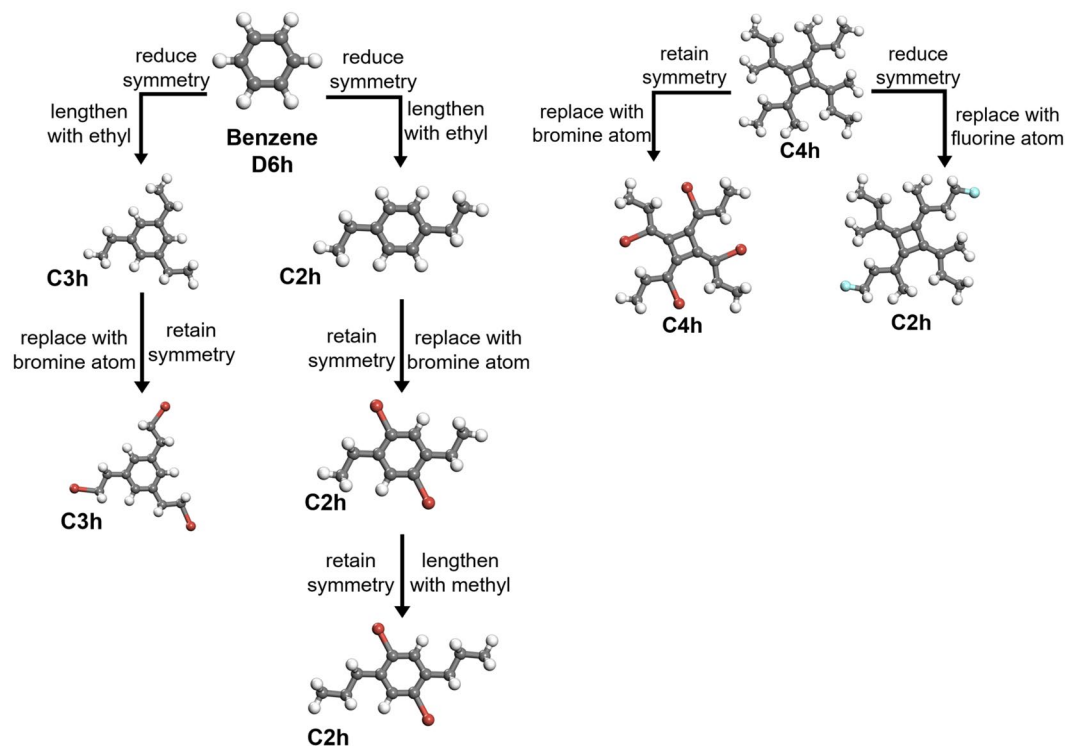


Fig. 1 Generation map of some molecules in the QM-sym database by reducing or retaining symmetry through replacements. The left part is the generation map of some C2h and C3h molecules. Starting from benzene (D6h), the symmetry group can be reduced to C2h or C3h or be retained correspondingly depending on whether replacing or lengthening is carried out during generation. The right part is a generation map of some C2h and C4h molecules. Starting from a C4h molecule, the symmetry group can be reduced to C2h by replacing two atoms or be retained by replacing four atoms. Grey, white, red, and blue balls denote carbon, hydrogen, and fluorine atoms, respectively.

H, B, C, N, O, F, Cl, and Br atoms. All of them have symmetries other than C1, including C2h, C3h, and C4h. Information about the basic symmetric unit and symmetry centre is also recorded in this database. The structure of this work is organized as follows. We first introduce the methods by which the database is generated and the major difference from the previous QM9 database and discuss some general results. Then, we randomly sample 100 molecules from the QM-sym database and discuss the benchmark with other numerical methods, such as G4MP2⁷, G4¹⁰, and CBS-QB3¹¹, with the same validation as the QM9 database. The QM-sym database, which can serve a function similar to that of the previous ones, provides an efficient training and evaluation database for the data-driven-based machine learning (ML) models in quantum chemistry¹². Because of the enclosed symmetrical information, the QM-sym database can provide more applications than precious databases in the orbital symmetry-dependent properties, such as excitation degeneracy and the selection rules of transitions. This symmetry-enclosed database can benefit more from the understanding and discovery of structure properties from the ML point of view¹³.

Methods

Generation of atomic coordinates. The generation of the QM-sym database includes two steps. First, we construct the raw molecular structures based on typical molecular information, such as bond angle and bond length, and then grow them with a genetic algorithm to find a relatively stable structure with given symmetric point groups. Examples of the generation map are shown in Fig. 1. For simplification, we initiate 3 point groups, C2h, C4h, and D6h, which correspond to the ethane, cyclobutene, and benzene, respectively, and then extend the molecular information by adding aliphatic hydrocarbon chains to branches. As shown in the left part of Fig. 1, CH₃CH₂ radicals extend the original benzene molecule from the hydrogen sites in two different ways, maintaining the C3h and C2h point groups throughout this step.

Furthermore, to increase the number of structures and structure diversity, we can also randomly sample the halogen elements (F, Cl, and Br) to replace the hydrogen atoms in the corresponding carbon chains and ring motifs. In each sampling, the point group can be retained or reduced by a corresponding partial replacement, such as from C4h to C2h, as shown in the right part of Fig. 1. Additionally, the implemented symmetry can simplify the input of the molecular structure by the primary structure information. For example, benzene, C₆H₆, can be represented as CH under the D6h point group.

The QM-sym database records the primary structure information for future symmetrical applications as well. After the raw molecular structures are established, similar to the QM9 database, each of the structures was further precisely optimized at the B3LYP/6-31G(2df,p) level of theory by Gaussian 09¹⁴ and was set to strictly follow its

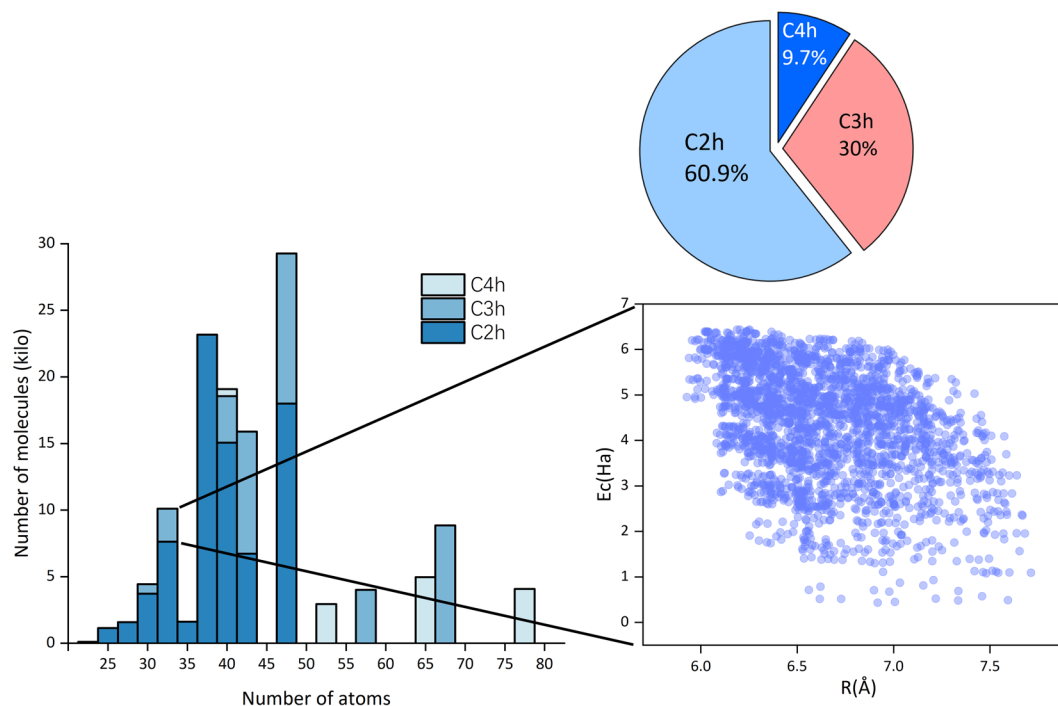


Fig. 2 An overview of the QM-sym database. The proportion of each space group in the QM-sym database is shown on the top. The left inset indicates the distribution of molecules with respect to their size according to the number of atoms and space groups. The right inset corresponds to some of the C_3h molecules, with their radius of rotation plotted versus their cohesive energy, and shows a distinct tendency of the cohesive energy to decrease when the rotation radius increases.

Reference	MAE	RMSE	maxAE
G4MP2	6.1 (5.0)	7.3 (6.1)	18.2 (16.0)
G4	5.4 (4.9)	6.3 (5.9)	15.4 (14.4)
CBS-QB3	5.6 (4.5)	6.7 (5.5)	16.7 (13.4)

Table 1. The benchmark of atomization enthalpies at B3LYP/6-31G(2df, p) level compared. The data in the parenthesis are from QM9 database.

designed symmetric group, but some of the raw structures generated by the first step were rational in only geometry without any guarantee in chemistry, which would cause SCF convergence to take a long time or even fail. In addition, more structures would be locally trapped in saddle points during the structure optimization due to the complexity and the initial symmetry settings. To solve this problem, we applied a similar strategy to that applied in the QM9 dataset⁶. We chose 200 maximal SCF cycles for all the molecule structures, and for those structures that failed the SCF convergence after 200 steps in 1 cycle, we applied very tight convergence criteria for the Gaussian 09 input and restarted. If the SCF convergence still failed, we ignored those structures. The frequency calculations were included, and for the molecules with imaginary frequencies after the procedure discussed above, we further applied additional iterations by using keywords, `opt (calcfc, maxstep = 5, maxcycles = 1000)`. We retained only the molecular structures with good SCF convergence in the ground state and non-negative frequencies.

The molecular structure with the D_{6h} point group derived from benzene is excellent in both SCF convergence and frequency distribution. It is observed that the DFT calculation can be well accelerated by the symmetrized molecular structures. For the molecular orbitals, we calculate and record at least 5 orbitals upward and downward from the LUMO and HOMO, i.e., from HOMO - 5 to LUMO + 5. The number of contained orbitals depends on the degeneracy; for example, if there is degeneracy between HOMO-6 and HOMO - 5, we count in HOMO-6 as well. A similar operation is performed from the LUMO to the LUMO + 5.

Data Records

The QM-sym database is publicly available at GitHub and Figshare⁹ (see the Code Availability part below). It now includes 13.5 k molecular structures (QM_sym_xyz_number.tar) and all the properties, including the point group, information on the basic symmetric unit, enthalpies, atomization, zero-point energy, energy and symmetry labels from at least HOMO - 5 to LUMO + 5. Detailed information on the available properties is documented in the README file. The proportion of each space group is shown in Fig. 2. Additionally, the 100 randomly chosen structures, benchmarked with other numerical methods (G4MP2⁷, G4¹⁰, and CBS-QB3¹¹), are also included in the database as benchmarked.tar.gz. The data are shown in Table 1, and the QM9 benchmark is also included

Line	Content
1	Number of atoms $N \cdot n_a$ (+1)
2	Properties of molecule
3, ..., $2 + n_a$	Coordinates of atoms in the first subgroup
$3 + n_a$, ..., $2 + 2 \cdot n_a$	Coordinates of atoms in the second subgroup
...	...
$3 + (N - 1)n_a$, ..., $2 + N \cdot n_a$	Coordinates of atoms in the last subgroup
$(3 + N \cdot n_a$ to end)	(Coordinates of atom on rotation axes)

Table 2. xyz file format for molecular structure and properties. The coordinate lines are shown in this format: atom, x position, y position, z position, charge.

No.	Property	Unit	Description
1	S	/	Symmetry group
2	E_g	eV	Bandgap
3	E_c	eV	LUMO
4	E_v	eV	HOMO
5, 6, 7	B_v	GHz	Rotational constant
8, 9, 10, 11	μ	D	Dipole moment
12	α	a_0^3	Isotropic polarizability
13	R^2	au	Electronic spatial extent
14, 15	ε_v	J/mol; Kcal/mol	Zero-point vibrational energy
16	$\varepsilon_0 + \varepsilon_{ZPE}$	Ha	Sum of electronic and zero-point energies
17	$\varepsilon_0 + E_{tot}$	Ha	Sum of electronic and thermal energies
18	$\varepsilon_0 + H_{corr}$	Ha	Sum of electronic and thermal enthalpies
19	$\varepsilon_0 + G_{corr}$	Ha	Sum of electronic and thermal free energies
20	C_v	Cal · Mol · Kelvin ⁻¹	Heat capacity
21–32	D	/	Degeneracy of orbitals
33–44	S_o	/	Symmetry of orbitals
45– ($N \cdot n_a + 44$)	S_g	/	Indication of subgroups and atoms

Table 3. Calculated properties. Properties are stored in the order given by the first column. The orbital degeneracy and symmetry are from ‘HOMO – 5’ to ‘LUMO + 5’, see further explanation in the following.

with the results derived from the G4MP2, G4, and CBS-QB3 methods. We randomly choose 100 molecules from 135k molecules and calculate the errors relative to the individual method, identified by the mean absolute error (MAE), root-mean-square error (RMSE), and maximal absolute error (maxAE). The units are kcal/mol. The values from the QM9 database are in parentheses and derived directly from the literature⁶.

File format. The QM_sym.xyz file contains the atomic coordinates, together with predicted properties information from the Gaussian09 calculation; each structure is indexed by QM_sym_i.xyz, where i is the index of the structure ordered in the database. The xyz file format is one of the most common file formats for molecular chemistry; it is ASCII coded and can be viewed by many free software like VESTA¹⁵ and Jmol¹⁶. The basic outline of the xyz format is shown in Table 2.

The original xyz file format includes only the structure information. Here, in the QM-sym database, we add more property and symmetry information to the comment lines. The user can directly read out all the information from the modified xyz file, as indicated in Tables 2 and 3. N in the tables is the number of subgroups, depending on the symmetry C_N , and n_a is the number of atoms in each subgroup. ‘+1’ and ‘ $4 + N \cdot n_a$ ’ in brackets will be present in the xyz files when there are atoms on rotation axes. An indication of the coordinates is shown at the end of the property line. Take ‘11 12 13 14 15 16 17 18 19 110 21 22 23 24 25 26 27 28 210 29 011’, which is a list of ‘signs’, from a C_{2h} structure as an example. The order of signs is the same as the order of the subsequent atom coordinates. Each sign is composed of two parts: a subgroup ID and a position ID. Taking ‘110’ and ‘210’ as an example, the first numbers, ‘1’ and ‘2’, denote that these atoms belong to subgroups 1 and 2, respectively, which have the same primitive structure as the two ‘CH’ subgroups in benzene (C₆H₆). The ‘10’ is the position ID for these atoms, which means that ‘110’ and ‘210’ are atoms of the same element at the same position in different subgroups. To differ atoms on rotation axes that belong to all subgroups, we use ‘0’ to denote the subgroup ID and a number that does not belong to any other atom to show its position. The atom coordinates are below the comment line, and the coordinates of the atoms are followed by the Mulliken charges.

Properties. The properties are numerically derived from the DFT calculation after the full relaxation of the molecular geometry with the initial symmetry settings. The technical details of the DFT calculation can be expressed as B3LYP/6-31G(2df,p), with a 10^{-5} eV criterion for energy convergence. The initially given symmetry

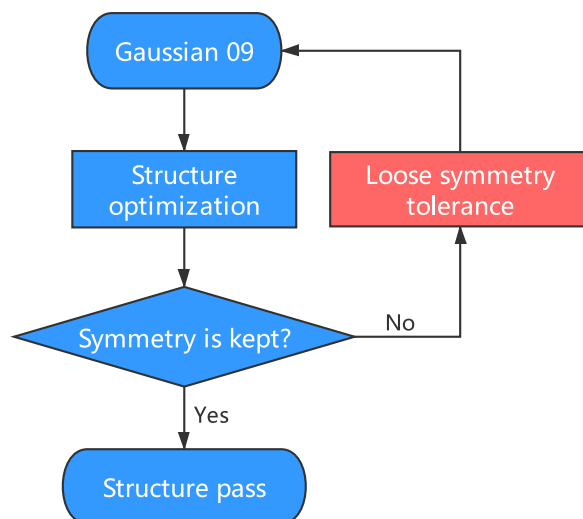


Fig. 3 Flow chart of the geometry check.

information and all the other properties of the 135 k molecular structures are listed in Table 3. Compared with the QM9 database, the QM-sym database is supplemented with the orbital symmetry and eigenvalue information from ‘HOMO – 5’ to ‘LUMO + 5’ and the symmetric unit cell segments by the point group. According to the symmetry information above, the user can calculate the excitation transition probability from ‘HOMO – 5’, ..., HOMO to LUMO, ..., ‘LUMO + 5’. To simplify the expressions of orbitals, in the xyz file, we regard the degenerated orbitals as one sign. For example, ‘1|2|1|1|BU|BG|AU|AG’ contains 5 orbitals, and there are two orbitals with ‘BG’ in the symmetry that are degenerated. Since all the molecules and the properties are symmetrized, a symmetrized neural network can be used to further optimize the efficiency from the symmetrical input¹⁷.

Technical Validation

Validation of geometry and symmetry consistency. In the QM9 database⁶, the authors applied InChI (IUPAC International Chemical Identifier) strings¹⁸ and SMILES (simplified molecular-input line-entry system) strings¹⁹ for the forward-feedback double-check with semi-empirical methods (SEMs) such as PM7 in MOPAC²⁰. Since both the InChI and SMILES descriptions lose geometric information regarding the bond lengths, bond angles, and dihedral angles, there are approximately 3000 structures that fail the consistency check in QM9. Here, in the QM-sym database, the arrangements of the elements are all constrained by the given point group; i.e., the symmetry can be the only measurement of the geometry. Due to the initial setting of the symmetry, both the DFT and SEM calculations are initially fully symmetrized in the given point group. For some structures, the symmetry information may be lost during the structure optimization, causing geometric inconsistency. The flow chart of the symmetry check is shown in Fig. 3. This issue can be solved by setting a ‘loose’ criterion for the symmetry identification and redoing the Gaussian 09 calculation. We find that there are approximately 2000 structures that fail the symmetry invariant test out of the 135 k molecules in the database. When we look closely into the 2000 structures by distributing the atomization energy and element distribution (except for hydrogen and carbon), we find that most of the structures that fail the symmetry check are of low stability with unphysical chemical structures; thus, we do not add them to QM-sym.

Validation of the quantum chemistry results. In the QM-sym database, as discussed above, all 135 k molecules are first generated by the symmetry operation; then, the structures are optimized at the B3LYP/6-31G(2df, p) level of DFT, with the same theory quality as that in the previous QM7^{3,4}, QM7b⁵, and QM9⁶ databases. The additional benchmarks with the G4MP2, G423, and CBS-QB3 functions are summarized in Table 1 as well. The QM-sym database arrives with an accuracy comparable to that of QM9, as the atomization enthalpies in the benchmark belong to the scalar properties and there is no significant benefit gained from accuracy in this domain. For all three additional functions, within the 100 randomly selected molecules, the MSE is approximately 6.1 kJ/mol, and the RMSE is 6 kJ/mol. In addition to the low numerical errors, the orbital degeneracy and symmetry-dependent calculation, similar to the transition selection rules, obtains the exact results.

Spectral transition probability. With the database present, according to methods provided by F. Albert Cotton²¹, the spectral transition probability could also be calculated based on the symmetry of the orbitals. By defining symmetry operations of the initial and targeting orbitals ψ_i , ψ_j , and the transition moment operator μ , the intensity of the transition is given by the equation:

$$I \propto \int \psi_i \otimes \mu \otimes \psi_j d\tau \quad (1)$$

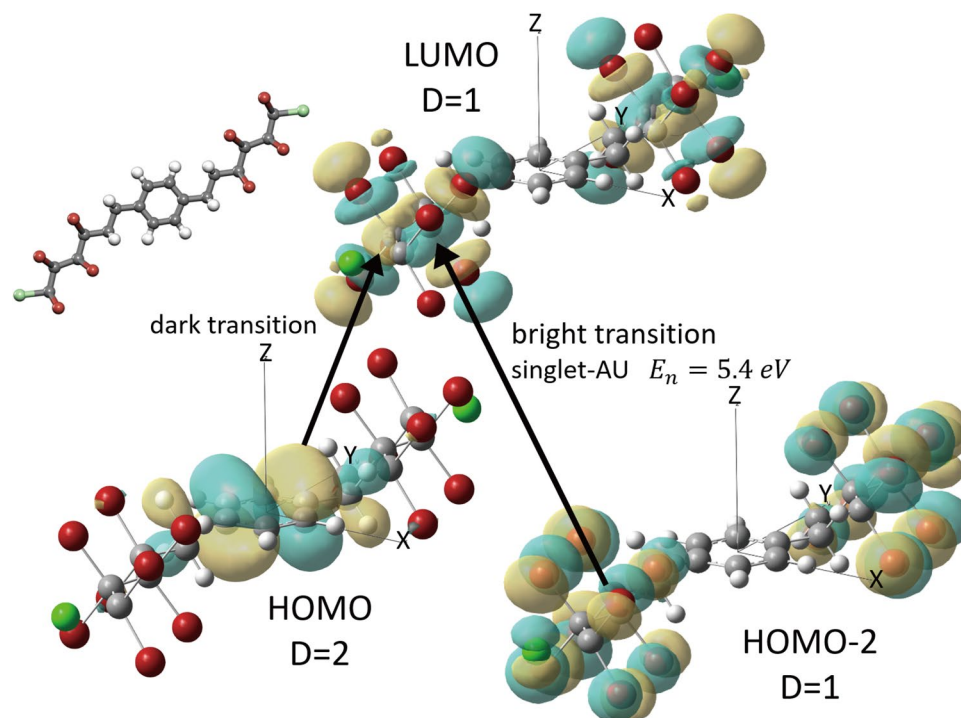


Fig. 4 Sketch of the excitations between orbitals with different energy levels. The degeneracy levels of the HOMO, HOMO – 2 and LUMO are 2, 1 and 1, respectively. From the results of both group theory and Gaussian 09, the transition from the HOMO to the LUMO is dark, while that from HOMO – 2 to the LUMO is bright, with a singlet with AU in terms of symmetry and an energy of 5.4 eV. An example molecule used for the spectral transition probability calculation is shown at the top left.

C_{2h}	E	C_2	i	σ_h	linear
A_g	1	1	1	1	R_z
B_g	1	-1	1	-1	R_x, R_y
A_u	1	1	-1	-1	z
B_u	1	-1	-1	1	x, y

Table 4. Character table of the symmetry group C_{2h} . It could be observed that x and y polarized light would be yielded by the symmetry operation B_u and that z polarized light would be yielded by A_u .

$T_{i \rightarrow f}$	Polarization of light	Direct products	Resulting characters				P_T
			E	C_2	i	σ_h	
$T_{HOMO \rightarrow LUMO}$	x, y	$B_g \times B_u \times A_g$	1	-1	-1	1	\times
	z	$B_g \times A_u \times A_g$	1	1	-1	-1	\times
$T_{HOMO-2 \rightarrow LUMO}$	x, y	$A_u \times B_u \times A_g$	1	-1	1	-1	\times
	z	$A_u \times A_u \times A_g$	1	1	1	1	$\sqrt{\quad}$

Table 5. Probabilities for the transition from T_i to T_f calculated using direct products. Only the last one is bright since only the direct product $A_u \times A_u \times A_g$ contains the total symmetry A_g . P_T is the probability of each transition, the transition is bright with $\sqrt{\quad}$ and dark with \times .

where ' \otimes ' refers to the direct product of symmetry operations. The characters of the representation of a direct product are equal to the products of the characters of the representations based on the individual sets of functions. Only when the total symmetry operation is present in the result of the direct product $\psi_i \otimes \mu \otimes \psi_j$ will this integral be nonzero; i.e., the transition of electrons from the ψ_i orbital to the ψ_j orbital via operator μ is possible. According to the results of Gaussian 09, part of the energy level diagram of the C_{2h} molecule in Fig. 4 is also shown in Fig. 4 as an example. To check the spectral transition, the operator μ must contain the Cartesian coordinates x , y or z (full character table for C_{2h} is available in Table 4). In this case, for x or y polarized light, $\mu = B_u$; for the z polarized light, $\mu = A_u$. Details regarding the characters calculated via Eq. 1 are shown in Table 5.

From the decomposition formula $a_i = \sum_R \chi_r(R) \chi_i(R) / h$ (R refers to all the symmetry operators of the symmetry group, χ_r refers to the character of the reducible representation Γ_r , χ_i refers to the characters of the i th irreducible representation, h refers to the dimensions of the symmetry group, and a_i refers to the number of times the i th irreducible representation occurs in Γ_r)²¹, it could be concluded that the total symmetry operation A_g occurs only in the transition $T_{HOMO-2 \rightarrow LUMO}$; thus, this is the spectral transition of this molecule with the lowest energy. The energy level $HOMO - 1$ is not listed here because it is degenerate with the $HOMO$; therefore, it has the same symmetry property as that of the $HOMO$. Thus, it would have the same transition property as that of $HOMO$ as well. To verify the group theory results, we randomly choose 350 molecules and perform TD-DFT transition calculations in Gaussian 09. The result for the above molecule shows that excitations from the $HOMO$ to the $LUMO$ yield dark transitions, while excitations from $HOMO - 2$ to the $LUMO$ yield bright transitions in only the z-direction, which agrees with our group theory calculation. The symmetry information in the QM-sym database provides the exact selection rules for the transition states of the molecules.

Code availability

The newest version of database is available on Figshare and GitHub. Figshare <https://doi.org/10.6084/m9.figshare.9638093>. GitHub <https://github.com/XI-Lab/QM-sym-database>.

Received: 25 June 2019; Accepted: 10 September 2019;

Published online: 18 October 2019

References

1. Curtarolo, S. *et al.* The high-throughput highway to computational materials design. *Nature Materials* **12**, 191 (2013).
2. Kirkpatrick, P. & Ellis, C. Chemical space. *Nature* **432**, 823 (2004).
3. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and Accurate Modeling of Molecular Atomization Energies with Machine Learning. *Physical Review Letters* **108**, 058301 (2012).
4. Blum, L. C. & Raymond, J.-L. 970 Million Druglike Small Molecules for Virtual Screening in the Chemical Universe Database GDB-13. *Journal of the American Chemical Society* **131**, 8732–8733 (2009).
5. Montavon, G. *et al.* Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics* **15**, 095003 (2013).
6. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **1**, 140022 (2014).
7. Curtiss, L. A., Redfern, P. C. & Raghavachari, K. Gaussian-4 theory using reduced order perturbation theory. *The Journal of Chemical Physics* **127**, 124105 (2007).
8. Zhang, Q. & Shreeve, Jn. M. Growing catenated nitrogen atom chains. *Angewandte Chemie International Edition* **52**, 8792–8794 (2013).
9. Liang, J., Xu, Y., Liu, R. & Zhu, X. Qm-sym-database. [figshare, https://doi.org/10.6084/m9.figshare.9638093](https://doi.org/10.6084/m9.figshare.9638093) (2019).
10. Curtiss, L. A., Redfern, P. C. & Raghavachari, K. Gaussian-4 theory. *The Journal of Chemical Physics* **126**, 084108 (2007).
11. Montgomery, J. A., Frisch, M. J., Ochterski, J. W. & Petersson, G. A. A complete basis set model chemistry. VII. Use of the minimum population localization method. *The Journal of Chemical Physics* **112**, 6532–6542 (2000).
12. Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O. & Walsh, A. Machine learning for molecular and materials science. *Nature* **559**, 547–555 (2018).
13. Le, T., Epa, V. C., Burden, F. R. & Winkler, D. A. Quantitative Structure–Property Relationship Modeling of Diverse Materials Properties. *Chemical Reviews* **112**, 2889–2919 (2012).
14. Gaussian 16 Rev. B.01 (Wallingford, CT, 2016).
15. Momma, K. & Izumi, F. VESTA: a three-dimensional visualization system for electronic and structural analysis. *Journal of Applied Crystallography* **41**, 653–658 (2008).
16. Jmol: an open-source Java viewer for chemical structures in 3D, <http://www.jmol.org>.
17. Ge, R., Kudritipudi, R., Li, Z. & Wang, X. Learning Two-layer Neural Networks with Symmetric Inputs. *arXiv preprint arXiv:1810.06793* (2018).
18. Heller, S. R. & McNaught, A. D. The IUPAC international chemical identifier (InChI). *Chemistry International* **31**, 7 (2009).
19. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of chemical information and computer sciences* **28**, 31–36 (1988).
20. Stewart, J. J. P. MOPAC2012. *Stewart Computational Chemistry, Colorado Springs, CO, USA*, <http://OpenMOPAC.net> (2012).
21. Cotton, F. A. *Chemical applications of group theory*. (John Wiley & Sons, 2003).

Acknowledgements

This work is supported by the Robotic Discipline Development Fund (2016-1418) from the Shenzhen Government, the Shenzhen Fundamental Research Foundation (JCYJ20170818103918295, JCYJ20180508162801893), and the National Natural Science Foundation of China (Grant No. 21805234).

Author contributions

All authors designed and performed research and wrote the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information is available for this paper at <https://doi.org/10.1038/s41597-019-0237-9>.

Correspondence and requests for materials should be addressed to X.Z.

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.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

The Creative Commons Public Domain Dedication waiver <http://creativecommons.org/publicdomain/zero/1.0/> applies to the metadata files associated with this article.

© The Author(s) 2019