QCforever: Quantum Chemistry for Everyone

Masato Sumita,[*,†,‡] Kei Terayama,[¶,†] Ryo Tamura,[§,‡,‖,†] and Koji Tsuda[§,‖,†]

†*RIKEN Center for Advanced Intelligence Project, Tokyo, 103-0027, Japan*

‡*International Center for Materials Nanoarchitectonics(WPI-MANA), National Institute for Materials Science, Tsukuba, 305-0044, Japan*

¶*Graduate School of Medical Life Science, Yokohama City University, Tsurumi-ku, 230-0045, Japan*

§*Graduate School of Frontier Sciences, the University of Tokyo, Kashiwa, 277-8561, Japan*

‖*Research and Services Division of Materials Data and Integrated System, National Institute for Materials Science, Tsukuba, 305-0047, Japan*

E-mail: masato.sumita@riken.jp

Abstract

To obtain observable physical or molecular properties like ionization potential and fluorescent wavelength with quantum chemical (QC) computation, multi-step computation manipulated by a human is required. Hence, automating the multi-step computational process and making it a black box that can be handled by anybody, are important for effective database construction and fast realistic material design through the framework of black-box optimization where machine learning algorithms are introduced as a predictor. Here, we propose a python library, QCforever, to automate the computation of some molecular properties and chemical phenomena induced by molecules. This tool just requires a molecule file for providing its observable properties, automating the computation process of molecular properties (for ionization potential, fluorescence, etc) and output analysis for providing their multi-values for evaluating a molecule. Incorporating the tool in black-box optimization, we can explore molecules that have properties we desired within the limitation of QC.

Introduction

In recent years, black-box optimization using machine learning (ML) algorithms as a predictor has achieved significant results in chemistry and materials science.[1,2] ML itself is not limited to these disciplines and can be applicable in many disciplines by changing the evaluating system (evaluator). Similarly, the evaluator in black-box optimization decides what kind of materials and molecules we desired. If we can install experiments like synthesizing materials and measuring their chemical or physical values as the evaluator, we can obtain the desired materials. Surely, several examples of black-box optimization with the experiments as the evaluators appear in inorganic materials because synthesizing inorganic materials is more efficient than the simulation depending on the target properties.[3–5] However, organic synthesis is not the case.

Organic synthesis is a time-consuming and formidable task including the characterization of synthesized molecules.[6] Hence, several simulation methods are developed as the preliminary methods that are expected to lower the experimental cost to find the expected molecules before the organic synthesis. Quantum chemical computation (QC)[7,8] is also one of them. In contrast to the expectation, QC has been mainly used as a tool to clarify chemical phenomena[9] through QC software packages.[10–12] Although QC is still based on an incomplete theory, tons of chemical phenomena are explained through theoretical chemistry, which can give some speculations to phenomena where the experimental information is not available. To make black-box optimization efficient by incorporating QC instead of chemical experiments, we should develop an automated QC system whose input is a molecule and output is its properties.

Although the QC is a powerful tool to obtain the electronic structures of molecules or materials, multi-step computation is required to obtain the practically meaningful physical or chemical values because most theories of QC are developed based on the orthogonal one-electron states,[13] which are not experimentally observable. Hence, to incorporate QC in black-box optimization, it is necessary to perform QC in a black box by automating the multi-step calculations and the analysis of the obtained results (usually text files). There are several tools for constructing inputs to perform complex computations and parsing output files like cclib,[14] ASE,[15] and QChASM.[16]

However, these tools are still far from the black box that is ready to incorporate QC in black-box optimization because their target is operating structure, distilling the total energy of the system, and one-electron-state based values. Furthermore, multi-objective optimization (optimizing multi-molecular properties) is necessary to obtain the practical materials through black-box optimization. Hence, the black box of QC should be a system that produces physically meaningful multi-properties.

In this paper, we propose a black box of QC that is ready to be incorporated in black-box optimization, QCforever whose input is a well-known sdf file and output is a physically meaning multi-properties, like ionization potential, electronic affinity, absorption wavelength, fluorescent wavelength, etc (surely, one-electron-orbital based properties like HOMO/LUMO gap are also available) because evaluating materials with multi-properties is important for their practical use. In addition, QCforever is useful to exclude the arbitrariness due to the different process in the computation of the physical values with QC. Excluding the arbitrariness, QCforever is also useful for building a database with standardized computational processes. Our implementation is available on GitHub at https://github.com/molecule-generator-collection/QCforever.

Method

Although there are several theories in QC, we employed density functional theory (DFT) [17] implemented in Gaussian16 [10] because of its easiness to use and versatility. Suitable processes for computing molecular properties are important for computational efficiency and reproducing chemical phenomena. Excluding the arbitrariness of computation process is also important for building a reliable database.

Because Gaussian16 supports multi-step jobs, we can summarize multi-step jobs to one input file and facilitate the computational process by reading previous electronic structures (orbitals). Figure 1 shows the computational flow to compute the several molecular properties and phenomena at one time. Different structures are saved as the different formatted checkpoint files. Currently

supported input is a common sdf file of one molecule, which is widely used in chemoinformatics, Gaussian chk, and Gaussian fchk files. When an sdf file is used as input, the number of radical electrons and charge are counted by the tool of RDkit .[18]

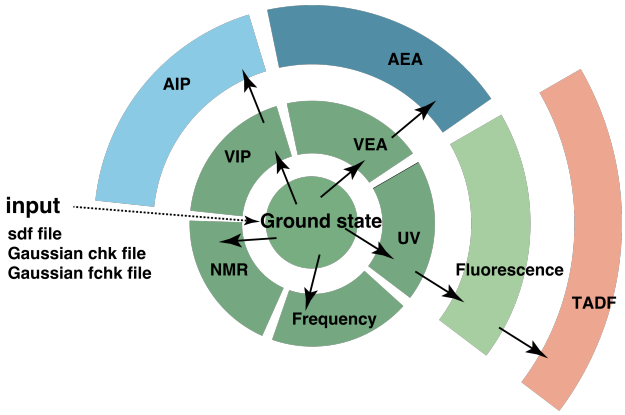

Figure 1: Computational flow of available properties of a molecule in QCforever. An sdf file of one molecule or Gaussian chk/fchk files are accepted as input. Solid arrows indicate reading atomic and electronic structures from the origin of an arrow. The broken arrow indicates that only atomic arrangement is obtained from the state at the origin of the arrow. The base for all computing is the ground state. Same geometries are represented by the same color.

QCforever computes the ground state at the first step. For conformation search, QCforever should rely on the other software.[19] It is possible to perform geometry optimization by option. At the present, the force constant estimation method (Gaussian default) is employed. If geometry optimization is performed, one maximum bond length is printed for checking geometry. After computation of the ground state, several molecular properties based on the orthogonal orbitals are obtained. HOMO/LUMO gap, and their relative energies to some references, and atomization energy are of importance to give speculation to the stability of a molecule in the ground state and its application to several materials. As the reference to compare the HOMO/LUMO level, their relative energies to the SOMO/LUMO energy of an oxygen molecule are computed as the following definition.

$$Ox = E_{\mathrm{O}_2}(\mathrm{LUMO}) - E_t(\mathrm{HOMO}) \tag{1}$$

$$Rd = E_t(\mathrm{LUMO}) - E_{\mathrm{O}_2}(\mathrm{SOMO}) \tag{2}$$

where, $E_{O_2}(\mathrm{LUMO})$ and $E_{O_2}(\mathrm{SOMO})$ are the LUMO and SOMO energies of O_2 respectively. $E_t(\mathrm{HOMO})$ and $E_t(\mathrm{LUMO})$ are the HOMO and LUMO energies of the target molecule respectively. Hence, *Ox* represents the proximity between HOMO of the target molecule and LUMO of O_2, resulting in the oxidation of the target molecule by O_2. On the other hand, because *Rd* represents the energetic proximity between LUMO of the target molecule and SOMO of O_2, *Rd* indicates the possibility of the reduction of target molecules by O_2. QCforever has the data of the SOMO and LUMO energies of O_2 that are computed with each combination of basis sets and functionals in advance. Hence, QCforever gives the values to the stability to O_2, which would be useful to compare the orbital levels to the band level of semiconductors.[20–22] Similarly, because QCforever has the energy of each atom which is computed with several basis sets and functionals, the atomization energy of the target molecule is computed.

Normal vibration modes of a molecule are computed by the vibrational analysis including intensities of frequency infrared (IR) and Raman spectra. Based on the normal mode, Gaussian calculates several thermochemical properties like Gibbs free energy, heat capacity, entropy, etc. QCforever dilutes these values from the log file. Peak positions in nuclear magnetic resonance (NMR) spectrum to the tetramethylsilane (TMS) of the target molecule are also computed using the GIAO method.

QCforever automatically computes the values that are relevant to photochemical properties/phenomena as shown in Figure 2, using the time dependent density functional theory (TD-DFT). Vertical excitation energies to other electronic structures from the ground state, which are observable as ultra-violet visible (UV) absorption measurement, can be computed at single point calculation. By using the TD-DFT, expected fluorescence is computed by optimizing the geometry in the target excited state as shown in Fig. 2.[23] The value (the Delta(S-T); energetic delta between singlet and triplet excited states in Fig. 2) for estimating the probability of thermally activated delayed fluorescence (TADF) [24] is computed through geometry optimization in the triplet state.

Computation for estimating vertical/adiabatic ionization potential (IP) and electronic affinity (EA)[25] is also automated in QCforever through the method called as ΔSCF. Vertical IP (*VIP*)

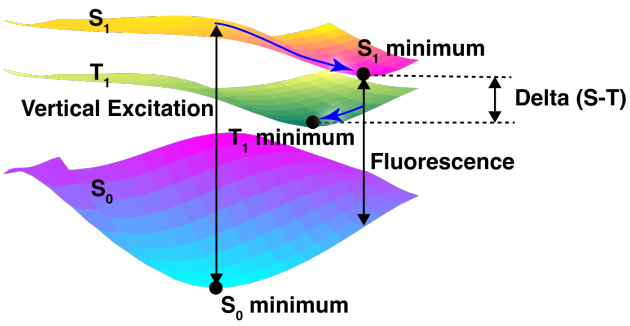

Figure 2: Schematics of potential energy surfaces of the singlet ground state (S_0), singlet excited state (S_1), and first triplet state (T_1) of a molecule. Blue arrows indicate the optimization process starting from the atomic and electronic structures at the origin of the arrows.

and EA (VEA) are the energetic difference between the ground state and the positively/negatively charged state (assuming the ground state is a neutral and singlet state) at the same structure as shown in Figure 3 by using the following equations.

$$VIP = E_{\mathrm{D_0 vertical}} - E_{\mathrm{S_0 minimum}} \tag{3}$$

$$AIP = E_{\mathrm{D_0 minimum}} - E_{\mathrm{S_0 minimum}} \tag{4}$$

$$VEA = E_{\mathrm{S_0 minimum}} - E_{\mathrm{D_0 vertical}} \tag{5}$$

$$AEA = E_{\mathrm{S_0 minimum}} - E_{\mathrm{D_0 minimum}} \tag{6}$$

Here, $E_{\mathrm{S_0 minimum}}$ is the ground state energy of the target molecule in Figure 3 (assuming that the ground state optimization is performed). $E_{\mathrm{D_0 vertical}}$ is the total energy of an electron donated or removed molecule in Figure 3. The values of adiabatic IP (AIP) and adiabatic EA (AEA) are calculated as eq. 4 and 6, where $\mathrm{D_0 minimum}$ is the energy obtained by preforming geometry optimization from $\mathrm{D_0 vertical}$ (Figure 3).

Currently available values are summarized in Table 1 and the keys of the dictionary are also as

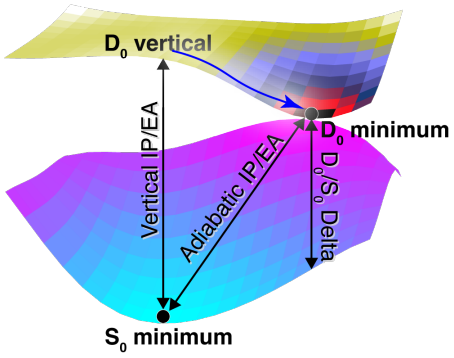

Figure 3: Schematics of potential energy surfaces of a neutral molecule (S_0) and its positively/negatively charged one (D_0), assuming a neutral molecule is in singlet state. A blue arrow indicates the optimization process starting from the structure and electronic structure of origin of the arrow.

the computed values are outputted as the dictionary format of python.

Table 1: Available values of QCforever and keys of the output dictionary.

Option names	Values obtained	key
opt	Geometry optimization in the ground state is performed	GS_MaxBoldLength (Å)
energy	Ground state energy	energy (in E_h)
homolumo	HOMO/LUMO gap	homolumo (in eV)
stable2o2	Stability to O_2	stable2o2 (in E_h)
deen	Atomization energy	deen (in E_h)
dipole	Dipole moment	dipole
cden	Mulliken charge and spin density	cden
symm	Molecular symmetry	symm
nmr	NMR chemical shift of each atom to TMS	nmr (ppm to TMS)
uv	Transition energies to excited state	uv (in nm with oscillator strength)
		state_index
freq	Vibrational analysis (298.15 K, 1.0 atm)	freq (in cm^{-1})
		IR_int (IR intensity)
		Ramman_int (Raman intensity)
		Ezp (Zero point energy)
		Et (Thermal energy)
		E_enth (Enthalpy)
		E_free (Free energy)
		Ei (Thermal energy in Kcal/mol)
		Cv (Heat capacity in mol K)
		Si (Entropy in mol K)
vip	Vertical ionization potential	vip (in eV)
vea	Vertical electronic affinity	vea (in eV)
aip	Adiabatic ionization potential	aip (in eV)
		relaxedIP_MaxBondLength (in Å)
aea	Adiabatic ionization potential	aea (in eV)
		relaxedEA_MaxBondLength (in Å)
fluor	Fluorescent from a specified state	MinEtarget (E_h)
		Min_MaxBondLength (in Å)
		fluor (in nm with oscillator strength)
tadf	Energetic difference between the singlet and the triplet excited state	T_Min (E_h)
		T_Min_MaxBondLength (in Å)
		T_Phos (in nm with oscillator strength)
		Delta(S-T) (in E_h)

Dependences

QCforever needs external quantum chemical computation package but mainly written in Python 3. Currently, only Gaussian16 is supported. Although Gaussian16 users may separate the computational scratch folder and data folder, current QCforever requires that data folder is the same as the scratch. Because Gaussian tools, formchk and unfchk, are used for making fchk or chk files, the path to Gaussian should be suitably set before using QCforever. To count the number of radical electrons and the value of total charge from a sdf file, RDKit[18] are required. Other required python libraries are Numpy. To make the data for computing atomization energy, chemical shift from TMS, and oxygen orbital level, bash scripts are used.

Example usage

It is necessary to make an instance because the main of QCforever is written as a class of python. QCforever needs the kind of functional and basis set, number of cores for Gaussian, and options, and input file names at least as the arguments. If one wants to compute molecular properties in solvent, one can specify the kind of solvents listed in Gaussian.[10] The memory and computational time can be specified by giving the values as the instance variables. The example of code (main.py) for QCforever is shown in List 1.

```
import os, sys
import GaussianRunPack

usage ='Usage; %s infile' % sys.argv[0]

try:
    infilename = sys.argv[1]
except:
    print (usage); sys.exit()
```

```
option = "opt homolumo energy dipole deen stable2o2 fluor=3" # option
   separated by more than one space

test = GaussianRunPack.GaussianDFTRun('B3LYP', 'STO-3G', 8, option,
   infilename)

test.mem = '5GB' # specify the value of memory
test.timexe = 60*60 # specify the maximum time of running Gaussian
outdic = test.run_gaussian()

print (outdic)
```

Listing 1: Example code for QCforever (main.py)

In the example of List 1, QCforever tries to compute the HOMO/LUMO gap, the ground state energy, dipole moment, atomization energy, the stability to O_2 based on the optimized structure of the target molecule in the ground state, and the fluorescence from the third excited state at the B3LYP/STO-3G level.

This code can be executed as the command as shown in List 2

```
$ python main.py ch2o.sdf
```

Listing 2: Example for ch2O.sdf (a sdf file of formaldehyde)

The result can be obtained as shown in List 3, which is the dictionary style of python code with the keys in Table 1. In the "uv" key, four lists are included. The first list indicates the excitation energy to each excited state in nm, the second is the intensity (oscillator strength) to them, the third indicates the length of circular dichroism (CD), and the fourth is the intensity of CD spectrum. Because we use unrestricted DFT calculation, spin allowed and forbidden excited state are mixed. Hence, the indices of spin allowed states are enclosed in the first list of "state_index" key, and those of spin forbidden states are in the second list. The excitation energies to spin allowed states are printed in the "uv" key. In the similar to "uv" key, "fluor" key includes the information of CD emission.

```
$ {'GS_MaxBondLength': 1.2503700019074353, 'homolumo': [6.018801089999999,
    6.018801089999999], 'dipole': [1.3513, -0.0, -0.0001, 1.3513], 'Energy
   ': -112.957313479, 'deen': -0.6349892466540155, 'stable2o2': [0.18606,
   0.23448], 'uv': [[334.14, 137.25, 112.51, 110.82, 92.45, 77.16, 75.95,
   70.19, 69.06], [0.0, 0.0085, 0.1006, 0.0, 0.0956, 0.0347, 0.131,
   0.0231, 0.293], [-0.0, 0.0, -0.0, 0.0, 0.0, -0.0, 0.0, -0.0, 0.0],
   [0.0, 0.0232, 0.0486, 0.0, 0.0765, 0.0488, 0.0348, -0.0039, 0.1926]], '
   state_index': [[2, 5, 7, 8, 10, 14, 16, 18, 20], [1, 3, 4, 6, 9, 11,
   12, 13, 15, 17, 19]], 'MinEtarget': -112.63357, 'Min_MaxBondLength':
   1.6029200006740822, 'fluor': [[864.49, 211.87, 203.81, 171.15, 137.36,
   135.12, 94.51], [0.0, 0.0039, 0.0539, 0.0002, 0.0078, 0.0, 0.0219],
   [-0.0, 0.0, -0.0, 0.0, -0.0, -0.0, 0.0], [0.0, 0.0151, 0.0055, 0.0019,
   0.002, 0.0, 0.0082]], 'log': 'normal'}
```

Listing 3: Example of obtained results

Applications

Using QCforever combined the black box optimization algorithms for discovering and designing materials, we have already reported the several results. Combining a deep learning based de novo molecule generator (DNMG)[26] with machine learning and QCforever, we have successfully demonstrated that molecules designed in silico for optical absorption/emission can be realized experimentally.[27–29] In addition, the DNMG proposed to use an material that had never received attention as an electret material.[30] The DNMG becomes a molecular identifier by setting the computed property by QCforever NMR spectrum.[31] In addition to the collaboration with DNMG, QCforever is useful for screening database. We have also employed QCforever with bound-Less Objective-free eXploration (BLOX) for searching out-of-trend materials from the database.[32] Here, we demonstrate the database screening as an example of the use of QCforever. Recent development of material informatics increases the importance of experimental[33–35] and computational databases[36,37] of molecules. Although PubChemQC[36] provides the observable molecular prop-

erties like absorption wavelength, computational databases basically provides total energies and properties based on one electron states.[38–40] They are might be important features but not practical properties. QCforever might be useful to translate another database to computational one with practical properties.

From the ZINC database,[35] we picked up 100 molecules available from vendors. For these molecules, we have computed the molecular properties, using QCforever with the following options listed in Table 1:

```
opt,freq,nmr,energy,homolumo,dipole,deen,stable2o2,cden,uv,fluor,tadf,vip,
    vea,aip,aea,symm
```

The success ratios for optimization in ground state (GS), fluorescence (Fluor), TADF, and AIP computations are 91, 97, 69, and 90% respectively as tabulated in Table 2. The average computational time per one molecule is about 9 hours for 20 cores. This computation is not definitely light. However, we can build the database for several molecular properties based on the electronic structure theory automatically. Because the multi properties can be simultaneously obtained, the correlation heat map among the computed molecular properties as shown in Fig. 4 is also easily obtained.

This correlation heat map shows the importance of the static analysis based on the database in spite of data of 100 molecules. The HOMO/LUMO gap shows the negative correlation with the absorption wavelength (Abs_wl), VEA, and AEA strongly. Furthermore, the gap has the positive correlation with Stable2o2 (oxidation by O_2), VIP, and AIP. Hence, the HOMO/LUMO gap is a molecular property that dominates not only photochemical properties but also electronic properties. On the other hand, Energy and E_free have no difference (this means that the contribution of the free energy is small in the small molecular size) and other properties (Delta(S-T), Fluor_wl, Freq, IR, Abs_it, Fluor_it) are not interrelated with the HOMO/LUMO gap. This results indicate the difficulty to make prediction model of these properties.

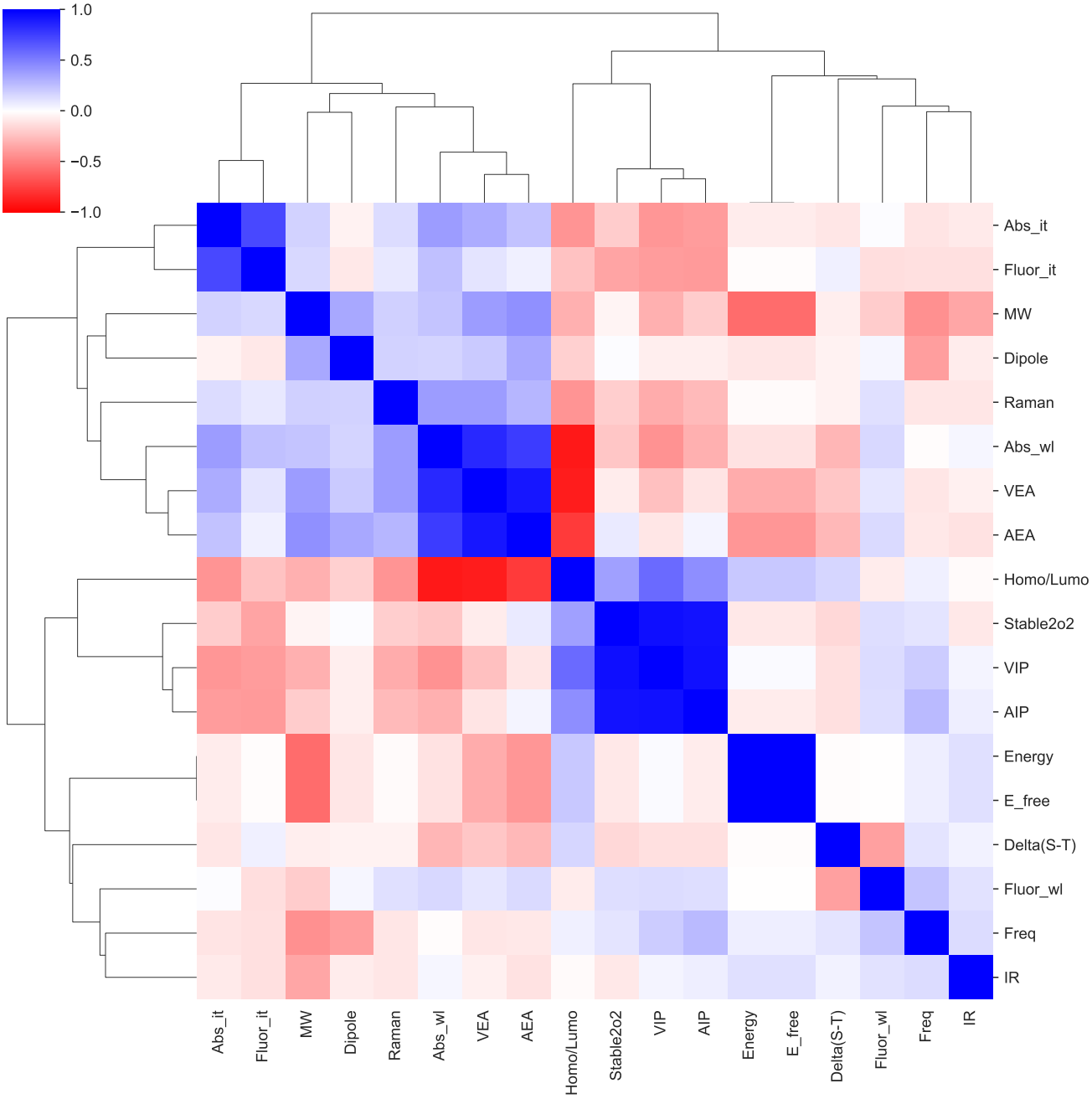

Figure 4: Clusterized correlation heat map among molecular properties of the 100 ZINC molecule computed by QCforever. Abs_it/Fluor_it; oscillator strength of absorption/fluorescence to/from the first excited state. Abs_wl/Fluor_wl; Absorption/Fluorescence wavelength to/from the first excited state. MW; Molecular weight. Dipole; absolute value of the dipole moment. Raman/IR; intensity of the lowest vibration modes of Raman/IR spectra. Freq; the lowest vibration mode in wave number. VEA/AEA; Vertical/Adiabatic electronic affinity. Homo/Lumo; Energetic gap between HOMO and LUMO. Stable2o2; oxdizability by O_2. VIP/AIP; vertical/adiabatic ionization potential. Energy; total energy of the ground state. E_free; Gibbs free energy at 297 K. Delta(S-T); the gap between minimums of the first excited state and the first triplet state

Table 2: Success ratio (%) for 100 molecules with QCforever at the B3LYP/6-31G* level

GS[a]	Fluore[b]	TADF[c]	AIP[d]
91	97	69	90

[a] Ground state optimization without any negative vibrational mode; [b] Geometry optimization in the first excited state valuable for evaluating fluorescence emitting; [c] Computation for evaluating thermally activated delayed fluorescence (TADF); [d]Geometry optimization ionized state to obtain adiabatic ionization potential.

Conclusion

In this paper, we demonstrated a tool automating the process to compute several observable molecular properties through QC; QCforever, which is ready to be equipped with black-box optimization. When QC calculations are used to calculate various physical and chemical properties or phenomena, arbitrary values might be obtained even for the same molecule due to the different computation processes. To avoid this, a standard computation process should be provided. Especially, standardized computation process as is in QCforever would be important for building a database based on QC calculation. As the demonstration of QCforever, we computed 100 molecules picked up from ZINC database.[35] Although the current QCforever could not exclude the several failures including the molecules that have the negative vibrational modes, the computation of 90% of molecules succeeded. In the near future, we will develop QCforever to deal with the negative vibrational mode and several failures like AiiDA.[41]

Simulation tools are expected to reduce the difficulty to develop new materials. QC computation was also one of them. In practice, however, QC is mainly used as a tool giving speculation to chemical phenomena. The history of QC proves that it is a powerful tool to get plausible answers to the forward problems where input is molecules. On the other hand, QC is also used for finding the expected molecules for chemical synthesis in experimental chemistry laboratories. This process corresponds to an inverse problem[42,43] where we should deal with the diversity of the chemical compounds. Surely, the search space is restricted within professional knowledge and favor. Combining QCforever with the black-box optimization algorithm, we can remove this restriction and bias and expand the search space.[27–32]

Acknowledgement

This research was conducted in "Development of a Next-generation Drug Discovery AI through Industry-academia Collaboration (DAIIA)" supported by Japan Agency for Medical Research and Development (AMED) under Grant Number JP22nk0101111. This work was also supported by MEXT as a "Program for Promoting Researches on the Supercomputer Fugaku (Application of Molecular Dynamics Simulation to Precision Medicine Using Big Data Integration System for Drug Discovery)". This research used the computational resources of the supercomputer center of RAIDEN of AIP (RIKEN).

Supporting Information Available

The SMILES list that supports the findings of this study are available in the supplementary material.

References

(1) Terayama, K.; Sumita, M.; Tamura, R.; Tsuda, K. Black-Box Optimization for Automated Discovery. *Acc. Chem. Res.* **2021**, *54*, 1334.

(2) Pollice, R.; Dos Passos Gomes, G.; Aldeghi, M.; Hickman, R. J.; Krenn, M.; Lavigne, C.; Lindner-D'Addario, M.; Nigam, A.; Ser, C. T.; Yao, Z.; Aspuru-Guzik, A. Data-Driven Strategies for Accelerated Materials Design. *Acc. Chem. Res.* **2021**, *54*, 849–860.

(3) Ren, F.; Ward, L.; Williams, T.; Laws, K. J.; Wolverton, C.; Hattrick-Simpers, J.; Mehta, A. Accelerated discovery of metallic glasses through iteration of machine learning and high-throughput experiments. *Sci. Adv.* **2018**, *4*, eaaq1566.

(4) Homma, K.; Liu, Y.; Sumita, M.; Tamura, R.; Fushimi, N.; Iwata, J.; Tsuda, K.; Kaneta, C.

Optimization of a Heterogeneous Ternary Li_3 PO_4 –Li_3 BO_3 –Li_2 SO_4 Mixture for Li-Ion Conductivity by Machine Learning. *J. Phys. Chem. C* **2020**, *124*, 12865–12870.

(5) Sumita, M.; Tamura, R.; Homma, K.; Tsuda, K. Li-Ion Conductive Li_3 PO_4 -Li_3BO_3 -Li_2SO_4 Mixture :Prevision through Density Functional Molecular Dynamics and Machine Learning. *Bull. Chem. Soc. Jpn.* **2019**, *92*, 1100–1106.

(6) Nicolaou, K. C. *Proc. Math. Phys. Eng. Sci.* **2014**, *470*, 20130690.

(7) Szabo, A.; Ostlund, N. S. *Modern Quantum Chemistry*; Dover Publications, Inc., 1989.

(8) Lowe, J. P. *Quantum chemistry*; Academic press, 1993.

(9) Friesner, R. A. *Porc. Natl. Acad. Sci. USA* **2005**, *102*, 6648–6653.

(10) Frisch, M. J. et al. Gaussian~16 Revision C.01. 2016; Gaussian Inc. Wallingford CT.

(11) Barca, G. M. J. et al. Recent developments in the general atomic and molecular electronic structure system. *J. Chem. Phys.* **2020**, *152*, 154102.

(12) Aprá, E. et al. NWChem: Past, present, and future. *J. Chem. Phys.* **2020**, *152*, 184102.

(13) Sumita, M.; Yoshikawa, N. Augmented Lagrangian method for spin-coupled wave function. *Int. J. Quantum Chem.* **2021**, *121*, 026746.

(14) O'Boyle, N. M.; Tenderholt, A. L.; Langner, K. M. cclib: A Library for Package-Independent Computational Chemistry Algorithms. *J. Comput. Chem.* **2007**, *29*, 839–845.

(15) Larsen, A. H. et al. The atomic simulation environment—a Python library for working with atoms. *J. Phys. Condens. Matter* **2017**, *29*, 273002.

(16) Ingman, V. M.; Shaefer, A. J.; Andreola, L. R.; E., W. S. QChASM: Quantum chemistry automation and structure manipulation. *WIREs Comput Mol. Sci. 11*, e1510.

(17) Cohen, A. J.; Mori-Sánchez, P.; Yang, W. Challenges for density functional theory. *Chem. Rev.* **2012**, *112*, 289–320.

(18) Landrum, G. RDKit: Open-Source Cheminformatics Software. 2016; `https://github.com/rdkit/rdkit/releases/tag/Release_2016_09_4`.

(19) Terayama, K.; Sumita, M.; Katouda, M.; Tsuda, K.; Okuno, Y. Efficient Search for Energetically Favorable Molecular Conformations against Metastable States via Gray-Box Optimization. *J. Chem. Theory Comput.* **2021**, *17*, 5419–5427.

(20) Hagfeldt, A.; Boschloo, G.; Sun, L.; Kloo, L.; Pettersson, H. Dye-sensitized solar cells. *Chem. Rev.* **2010**, *110*, 6595–663.

(21) Lu, M.; Liang, M.; Han, H.-Y.; Sun, Z.; Xue, S. Organic Dyes Incorporating Bis-hexapropyltruxeneamin Moiety for Efficient Dye-Sensitized Solar Cells. *J. Phys. Chem. C* **2011**, *115*, 274–281.

(22) Kranthiraja, K.; Saeki, A. Experiment-Oriented Machine Learning of Polymer:Non-Fullerene Organic Solar Cells. *Adv. Funct. Mater.* **2021**, *31*, 1–11.

(23) Atkins, P. *Atkins' physical chemistry*; Oxford University Press., 2017.

(24) Uoyama, H.; Goushi, K.; Shizu, K.; Nomura, H.; Adachi, C. Highly efficient organic light-emitting diodes from delayed fluorescence. *Nature* **2012**, *492*, 234–238.

(25) Boldyrev, A. I.; Simons, J.; Zakrzewski, V. G.; von Niessen, W. *J. Phys. Chem.* **1994**, *98*.

(26) Yang, X.; Zhang, J.; Yoshizoe, K.; Terayama, K.; Tsuda, K. ChemTS: an efficient python library for de novo molecular generation. *Sci. Technol. Adv. Mater.* **2017**, *18*, 972–976.

(27) Sumita, M.; Yang, X.; Ishihara, S.; Tamura, R.; Tsuda, K. *ACS Cent. Sci.* **2018**, *4*, 1126.

(28) Fujita, T.; Terayama, K.; Sumita, M.; Tamura, R.; Nakamura, Y.; Naito, M.; Tsuda, K. Understanding the evolution of a de novo molecule generator via characteristic functional group monitoring. *Sci. Technol. Adv. Mater.* **2022**, *23*, 352–360.

(29) Sumita, M.; Terayama, K.; Suzuki, N.; Ishihara, S.; Chahal, M. K.; Payne, D. T.; Yoshizoe, K.; Tsuda, K. *Sci. Adv.* **2022**, *8*, eabj3906.

(30) Zhang, Y.; Zhang, J.; Suzuki, K.; Sumita, M.; Terayama, K.; Li, J.; Mao, Z.; Tsuda, K.; Suzuki, Y. Discovery of polymer electret material via de novo molecule generation and functional group enrichment analysis. *Appl. Phys. Lett.* **2021**, *118*, 223904.

(31) Zhang, J.; Terayama, K.; Sumita, M.; Yoshizoe, K.; Ito, K.; Kikuchi, J.; Tsuda, K. NMR-TS: de novo molecule identification from NMR spectra. *Sci. Technol. Adv. Mater.* **2020**, *21*, 552–561.

(32) Terayama, K.; Sumita, M.; Tamura, R.; Payne, D. T.; Chahal, M. K.; Ishihara, S.; Tsuda, K. Pushing property limits in materials discovery via boundless objective-free exploration. *Chem. Sci.* **2020**, *11*, 5959–5968.

(33) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **2021**, *49*, D1388–D1395.

(34) Joung, J. F.; Han, M.; Jeong, M.; Park, S. Experimental database of optical properties of organic compounds. *Sci. Data* **2020**, *7*, 1–6.

(35) Irwin, J. J.; Sterling, T.; Mysinger, M. M.; Bolstad, E. S.; Coleman, R. G. ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* **2012**, *52*, 1757.

(36) Nakata, M.; Shimazaki, T. PubChemQC Project: A Large-Scale First-Principles Electronic Structure Database for Data-Driven Chemistry Maho. *J. Chem. Inf. Model.* **2017**, *57*, 1300.

(37) Ramakrishnan, R.; Dral, P. O.; Rupp, M.; Von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **2014**, *1*, 1.

(38) 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.* **2012**, *52*, 2864–2875.

(39) von Lilienfeld, O. A.; Müller, K. R.; Tkatchenko, A. Exploring Chemical Compound Space with Quantum-Based Machine learning. *Nat. Rev. Chem.* **2020**, *4*, 347–358.

(40) Cai, J.; Chu, X.; Xu, K.; Li, H.; Wei, J. Machine learning-driven new material discovery. *Nanoscale Adv.* **2020**, *2*, 3115–3130.

(41) Huber, S. P. Automated reproducible workflows and data provenance with AiiDA. *Nat. Rev. Phys.* **2022**,

(42) Sanchez-lengeling, B.; Aspuru-Guzik, A. Inverse molecular design using machine learning: Generative models for matter engineering. *Science* **2018**, *365*, 360–365.

(43) Kim, K. et al. Deep-learning-based inverse design model for intelligent discovery of organic molecules. *npj Comput. Mater.* **2018**, 1–7.