scientific data

Check for updates

OPEN

DATA DESCRIPTOR

Simulated carbon K edge spectral database of organic molecules

Kiyou Shibata[1]✉, Kakeru Kikumasa[1], Shin Kiyohara[1,2] & Teruyasu Mizoguchi[1]✉

Here we provide a database of simulated carbon K (C-K) edge core loss spectra of 117,340 symmetrically unique sites in 22,155 molecules with no more than eight non-hydrogen atoms (C, O, N, and F). Our database contains C-K edge spectra of each carbon site and those of molecules along with their excitation energies. Our database is useful for analyzing experimental spectrum and conducting spectrum informatics on organic materials.

Background & Summary

Carbon-based organic molecules form an immense configuration space and still have a lot of potentials for unknown functionalities in various fields. In the research and development of organic materials and their functionalities, accurate characterization of configuration and prediction of functionalities are the keys for success.

Among analytical clues for materials characterization, core-loss spectra, electron-energy loss near edge structure (ELNES) and X-ray absorption near edge structure (XANES), have been widely used as one of the most effective fingerprints for determination of local atomic structures and electronic states. These core-loss spectroscopy measure energy loss due to excitation of an electron from a core orbital and the core-loss spectra contain a partial density of states of the unoccupied states and possess useful information on atomic structure and electronic structure. Analysis of the core-loss spectrum have been performed by a comparison of measured spectra to reference spectra. Recent developments in experimental equipment and facilities have enabled core loss spectroscopy at high resolution regarding time, space, and energy. On the other hand, because of the extremely large amount of spectral data obtained, it is becoming increasingly important to establish methods for efficient and automatic analysis. Organic molecules have a variety of molecular structures, and the correlation between their structures and chemical bonds and core-loss spectral shapes is complex and has been remaining elusive.

These situations have been stimulating the development of reference spectral databases including huge variety of spectra obtained by experiments[1] and calculation[2]. Although a database of the core-loss spectra for inorganic materials has been recently developed[3], databases of the organic molecules are highly limited. Furthermore, in the field of materials science, application of informatics on predicting and designing various materials properties from spectrum data in data-driven manner has been attracting much interest. These facts represent increasing need for spectrum database of organic materials.

Here, we calculated core-loss spectra of carbon K (C-K) edge of 117,340 symmetrically unique sites in 22,155 molecules in a structure and property database of organic molecules[4,5], based on density functional theory (DFT). Our dataset provides theoretical fingerprints for analyzing experimental core-loss spectra, and also offers an opportunity for trying data-driven spectrum informatics on organic molecules.

Methods

Density functional theory calculations. The calculations of C-K edge spectra and excitation energies were carried out based on DFT by the first-principles plane-wave basis pseudopotential method using CASTEP code[6–11]. The generalized gradient approximation in a Perdew-Burke-Ernzerhof (GGA-PBE)[12] was adopted for the exchange-correlation functional. Spin polarization was not considered. For each carbon site in each molecule, an excited electronic structure was separately calculated using an on-the-fly potential of carbon. In the pseudopotential calculation, the core-hole effect can be taken into account by employing a special pseudopotential designed for the excited atom with a core-hole[11]. We consider a neutral excited state including a full core hole, *i.e.* a state with a 1 s core hole and an additional electron at an orbital corresponding to the original lowest unoccupied molecular orbital (LUMO) state, to calculate both the excitation energy and the spectral feature. The

[1]Institute of Industrial Science, the University of Tokyo, 4-6-1, Komaba, Meguro-ku, Tokyo, 153-8505, Japan. [2]Laboratory for Materials and Structures, Institute of Innovative Research, Tokyo Institute of Technology, Yokohama, 226-8503, Japan. ✉e-mail: kiyou@iis.u-tokyo.ac.jp; teru@iis.u-tokyo.ac.jp

absolute value of the excitation energy was calculated as the total energy difference between the ground state and the excited state, so called Δ self-consistent field (Δ SCF) method[13]. The calculating the excitation energy in the Δ SCF scheme has been applied not only for all electron calculations but also for the pseudopotential calculations with PAW method by correcting contribution of the core orbitals to energy[14,15]. The spectral feature whose energy is relative to the energy of LUMO level was calculated as the pair of the dipole transition matrix elements between the core states and the unoccupied eigenstates. The peak positions relative to the transition to LUMO state are obtained as the difference in eigenvalues between the LUMO state and the unoccupied eigenstates. Using this method, not only the spectral features but also the chemical shift of the spectrum can be calculated. The cut-off energy of the plane wave 500 eV was used for all calculations. For the calculation of spectrum, 1,000 unoccupied levels were considered for all molecules.

We have selected the calculation method for the present study because of the following reasons: (1) The pseudopotential method is more efficient than the all-electron method because core-orbitals are not explicitly calculated[9]. (2) The CASTEP code adopts a pseudopotential method which provides more accurate theoretical spectra than those obtained by a muffin-tin potential method especially for systems with anisotropic covalent bondings[16]. (3) The calculation method has been applied for the core-loss spectra of many kinds of materials, including crystalline, doped graphene, liquid, and gaseous materials[14,17–23].

Limitations and inaccuracies of this method can be summarized as the following: By setting the limited number of unoccupied levels, the spectra corresponding to the transitions to 1,000 unoccupied eigenstates are obtained. In the case of the carbon K-edge of the molecule calculated here, this corresponds to an energy range of up to 25 eV relative to the LUMO level, although it varies depending on the molecule and site. Using the pseudopotential method can save calculation costs, but it needs correction of energy contribution from the core orbitals that are not explicitly calculated. It has been reported that the excitation energy obtained by Δ SCF may vary up to about 20 eV depending on target elements, the calculation basis function set, potential description, and their practical implementations in calculation code[18]. The treatment of the exchange-correlation functional and DFT formalism also affects the evaluation of excitation energy[24]. From the view point of accuracy in the excitation energy, there are some other DFT formalisms such as the *GW* method[25] and the time-dependent DFT[26], but these methods require much higher calculation cost to construct a comprehensive spectral database. There are also other methods in dealing with the core-hole effect depending on the DFT formalism such as a method of Slater transition with half core hole in the Hartree-Fock-Slater method (or X α) method[27] and approximation using an atom of the atomic number of $Z+1$ including a full core hole[28]. The calculation condition adopted in this study is only one of them, and there is room for further discussion on how to incorporate this effect because the extent of the core-hole effect is affected by the life time of the core hole and its shielding effects which are largely dependent on the materials.

We carried out the calculations on 22,288 molecules which contain at least one carbon atom and no more than eight non-hydrogen atoms (C, O, N, and F) in QM9 dataset version 2[4,5]. Atomic configuration of each molecule was extracted from the dataset and was converted to the format of CASTEP input, namely "*.cell" files with a 15×15×15 Å periodic calculation cell. The Γ-point sampling was used for all calculations by explicitly setting KPOINT_MP_GRID, SPECTRAL_KPOINT_MP_GRID, and ELNES_KPOINT_MP_GRID as (1, 1, 1) in all the "*.cell" files. The atomic configurations in the dataset were used as is for all calculations, and only the electronic structure was optimized. For each molecule, symmetrically unique carbon sites were identified by a python program using the "PointGroupAnalyzer" class in the "pymatgen.symmetry" package, which is originally based on the spglib library[29], and we adopted the parameter "tolerance" = 0.285. Calculations of spectrum were carried out considering the excitation at the symmetrically unique carbon sites. For molecules with multiple equivalent carbon sites, we only carried out calculation on only one of the equivalent sites to reduce calculation cost. For each symmetrically unique carbon site, three calculations for C-K edge spectrum and excitation energy were carried out as follows: First, excited electronic state with a full core hole in the 1 s orbital was calculated by the self-consistent field method. Secondly, theoretical C-K edge was calculated for the same system with consideration of a large number of unoccupied bands using the obtained electronic states in the first step. Finally, an electronic structure of the ground state was calculated for evaluating the theoretical excitation energy. The CASTEP code is based on the pseudopotential method and supports generation of a pseudopotential during runtime ("on-the-fly pseudopotential"). For all the three calculation steps, an on-the-fly ultrasoft pseudopotential was used for the symmetrically unique carbon site of interest in order to include the core hole effect and to evaluate excitation energies[16,18]. For the excited state, a full core hole in the 1 s orbital was introduced in an on-the-fly pseudopotential by setting "C:ex 2|1.4|6|11|11|20:21(qc = 7)1s1" in the "*.cell" file, and self-consistent field calculation was carried out to get total energy of the system and electron density. The C-K edge spectrum was then calculated using the calculation results on the excited state. The ground state was also calculated using an on-the-fly pseudopotential at the carbon site of interest but with the filled core state by setting "C:ex 2|1.4|6|11|11|20:21(qc = 7)1s2" in the "*.cell" file. The cutoff in number of unoccupied bands for the spectrum calculation was set to 1,000 by specifying "elnes_nextra_bands: 1000" in the "*.param" file. In each step, atomic energy of carbon at the site of interest was calculated by both pseudopotential method and all electron method and was stored in the "*.castep" file.

The spectral data is calculated through the dynamical structure factor for each eigen value of unoccupied states by calculating the transition matrix element using the projector augmented wave (PAW) approach[30] within the dipole approximation[11]. The operator for transition process in ELNES and XANES can be written in the form of $\hat{O}_{\mathrm{E}} = \exp(i\mathbf{q}\cdot\mathbf{r})$ and $\hat{O}_{\mathrm{X}} = \exp(i(\omega/c)\hat{\mathbf{n}}\cdot\mathbf{r})\hat{\varepsilon}\cdot\mathbf{p}$, respectively, where $\mathbf{r}$ is the relative position of the excited electron from the core, $\mathbf{q}$ is the scattering vector, $\hat{\varepsilon}$ is the polarization vector, $\hat{\mathbf{n}}$ is the unit vector of propagation, and $\mathbf{p}$ is the momentum operator. In either case, the dipole approximation can be applied under $\delta \ll 1$: $\exp(i\delta) \sim 1 + i\delta$. Thus, the transition matrix element is proportional to $\phi_c|r_\alpha|\psi_{n,k}$ for a specific posi-

State/Calculation	All electron	Pseudo atomic
Ground state	−1027.632	−148.5146
Excited state	−723.254	−241.7997

Table 1. Atomic energies used for correcting the excitation energies in the unit of eV. These energies were used for calculating the correction term $\Delta E_{\mathrm{core(atom)}} = 397.6631\mathrm{eV}$ for calculating the excitation energy $E_{\mathrm{TE}} = \Delta E_{\mathrm{valence}} + \Delta E_{\mathrm{core(atom)}}$, where $\Delta E_{\mathrm{valence}}$ is given by the difference in the final energies of the ground state and the excited state.

tional operator $r_\alpha = x, y, z$, depending on direction of momentum transfer vector **q** in case of ELNES or that of propagation vector $\hat{\mathbf{n}}$ in case of XANES, where ϕ_c and $\psi_{n,k}$ are the core state and the unoccupied final state. In the CASTEP code, this is calculated through expansion with pseudofunctional basis set[11] using PAW projector function $\tilde{p}_i$, its corresponding pseudofunction $\tilde{\phi}_i$, and a pseudofunction $\tilde{\psi}_{n,k}$ corresponding to $\psi_{n,k}$ as: $\langle\phi_c|r_\alpha|\psi_{n,k}\rangle = \langle\phi_c|r_\alpha|\tilde{\psi}_{n,k}\rangle + \sum_i(\langle\phi_c|r_\alpha|\phi_i\rangle - \langle\phi_c|r_\alpha|\tilde{\phi}_i\rangle)\langle\tilde{p}_i|\tilde{\phi}_{n,k}\rangle$. The dynamical structure factor then can be obtained by the square of the absolute value of the transition matrix element above, and spectrum is calculated for each positional operator r_α by merging the dynamical structure factors for each unoccupied eigen states. From physical point of view, the spectrum for a specific r_α is proportional to the spectrum taken by the incident direction of r_α. The difference in spectra for r_α is originating from the anisotropic environment of the excitation site.

Some of the calculations of 133 molecules failed due to any errors in electronic structure optimization or other calculation procedures, and such molecules are not included in the dataset, resulting in the valid dataset for 22,155 molecules. For confirming the detailed calculation conditions, some of the raw CASTEP files ("*.cell", "*.param", and "*.castep") are available at figshare[31].

Parsing data and creating database. *Excitation energy.* Since we adopted pseudopotential calculations using the CASTEP code, the excitation energy cannot be directly obtained from the difference of total energies. To obtain the excitation energy, we corrected the change in energies of the core orbitals using atomic energies obtained by all-electron and pseudo-atomic calculations using the reported procedure[14]. All the energies used for calculating the excitation energy were extracted from the output "*.castep" files. The atomic energies of the carbon atom obtained by all-electron and pseudo-atomic calculations are the same for all the carbon sites and are summarized in Table 1.

Eigenvalues and dynamical structure factors. The CASTEP code outputs the energies and the dynamical structure factors for $r_\alpha = x, y, z$ to a text-formatted "*.elnes" file, and also outputs raw information about eigenvalue and transition matrix elements to "*.bands" and "*.eels_mat" files, respectively. For creating spectral database, we extracted data from "*.bands" and "*.eels_mat" files and formed spectra. The pairs of the eigenvalues and the transition matrix elements are parsed from "*.bands" and "*.eels_mat" file by a Python script which we published at GitHub[32]. The dynamical structure factors were obtained by the square of the absolute values of the transition matrix elements.

Site-specific C-K edge spectra. The C-K edge spectra for each symmetrically unique carbon site were calculated from the pairs of the eigenvalues and the dynamical structure factors by applying Gaussian smearing of 0.5 eV relative to the LUMO state eigenvalue. The total spectra were calculated by averaging over the three r_α directions.

Molecular C-K edge spectra. For each molecule, molecular spectra were obtained by shifting the site specific C-K edge spectra by the difference in excitation energies followed by averaging considering multiplicity of symmetrically unique sites. It should be emphasized that the relative energy difference between corresponding peaks in spectra from different sites is reported to reproduce that of experimental spectra[14], which assures the summing up the calculated spectra from the individual sites to the calculate whole molecular spectra. Since there are some errors on the calculation of spectrum or excitation energy for any sites, C-K edge spectra of the molecules are missing. As a result, 22,155 molecular spectra were obtained in total from C-K edge spectra of 117,340 symmetrically unique sites. Figure 1 shows some typical spectra after applying smoothing with a Gaussian function of standard deviation of 0.5 eV.

Data Records

The spectral data of the pair of raw eigenvalues and dynamical structure factors of each symmetrically unique carbon site are provided in HDF5 format at figshare[31]. For a typical use, the site and molecular specific C-K edge spectral data that are smoothed with a Gaussian filter with a standard deviation of 0.5 eV and sampled with 0.1 eV steps are provided as HDF5 files and csv files at figshare[31]. The site specific spectra database provides spectra of symmetrically unique carbon sites in terms of symmetrical equivalence along with their multiplicity and excitation energy. The molecular specific spectra database provides spectra of molecules obtained by weighted averaging considering multiplicity and excitation energy of the symmetrically unique sites in each molecule. For generating spectra with a desired smearing parameters, we also provide a Python script which can calculate smeared spectra from the HDF5 file of the site specific eigenvalues and dynamical structure factors. The raw CASTEP input and some output files are provided at figshare[31] for reproduction and transparency of the calculation.

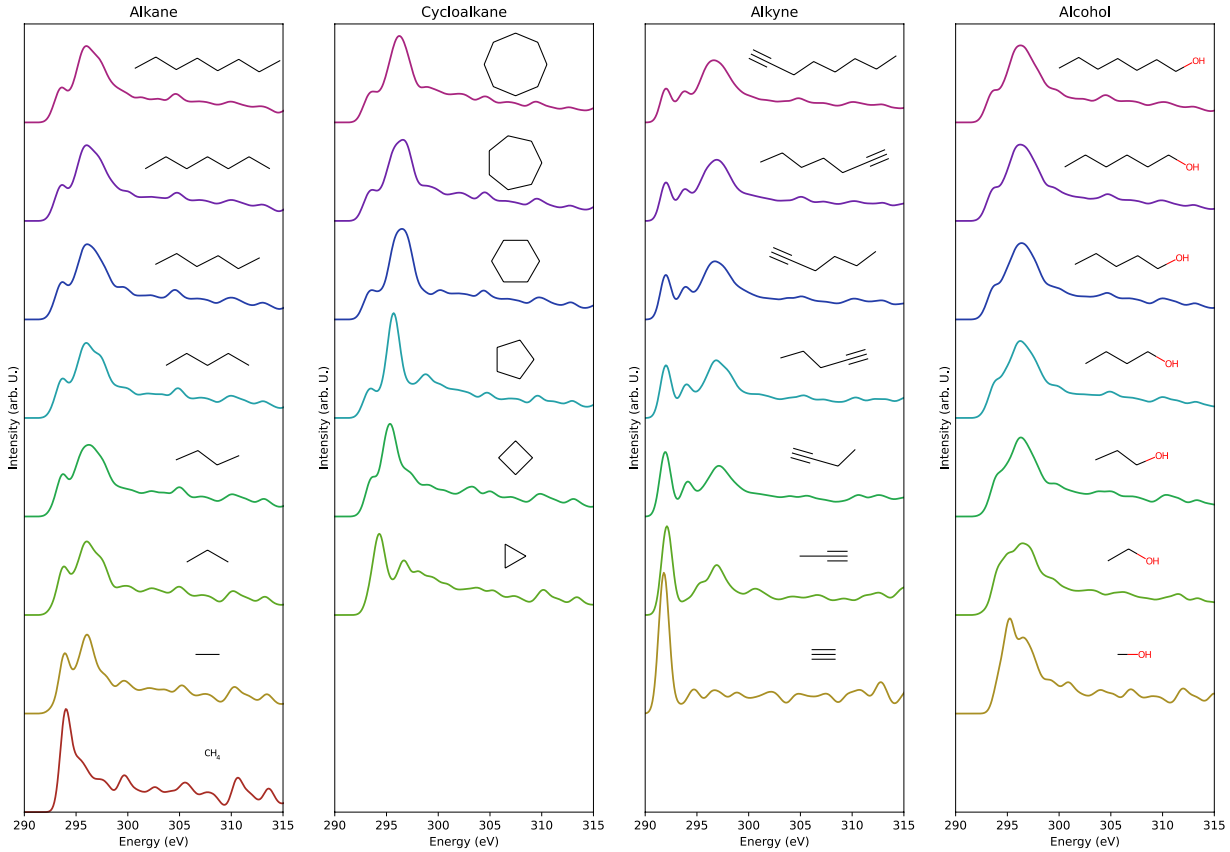

Fig. 1 The calculated spectra smoothed with Gaussian filter with the standard deviation of 0.5 eV of some typical molecules in the dataset.

Column	Header	Content	Unit
1	energy	energy loss	eV
2	tot	total dynamical structure factor	Å^2
3	x	*x* component of dynamical structure factor	Å^2
4	y	*y* component of dynamical structure factor	Å^2
5	z	*z* component of dynamical structure factor	Å^2

Table 2. CSV file format for molecular and site-specific C-K edge core-loss spectra.

Column	Header	Content	Unit
1	site	site number	
2	specie	atomic specie	
3	x	*x* coordinate of the site	Å
4	y	*y* coordinate of the site	Å
5	z	*z* coordinate of the site	Å
6	multiplicity	multiplitiy of the site	
7	representative	representative site used for calculation	

Table 3. CSV file format for a molecular structure.

Spectra database files in HDF5 format. The pairs of the eigenvalues and the dynamical structure factors, site specific spectra, and molecular spectra are separately provided in the format of HDF5 at figshare[31].

Site specific eigenvalues and dynamical structure factors. This database contains the most primitive information related to spectrum: the final energies for the excited state and the ground state, the excitation energy, the site multiplicity, the eigenvalues and the dynamical structure factors, *i.e.* the square of the absolute values of the transition matrix element. The information of the *j*-th site of the *i*-th molecule is stored in the */i/j* group in the

ID in QM9	name of molecule[ref. no.]
#4	Acetylene[40]
#7	Ethane[41]
#10	Acetonitrile[42]
#11	Acetaldehyde[43]
#13	*n*-Propane[44]
#16	Cyclopropane[45]
#21	Isobutane[44]
#29	2-Butyne[46]
#33	Propargyl alcohol[47]
#37	Methyl formate[47]
#39	*n*-Butane[48]
#40	*n*-Propanol[47]
#47	Cyclobutane[45]
#51	Imidazole[49]
#54	2,2-Dimethylpropane[44]
#55	*t*-Butyl alcohol[50]
#83	Isopentane (i.e. 2-Methylbutane)[44]
#133	*n*-Pentane[44]
#136	Diethyl ether[51]
#156	Cyclopentene[45]
#159	Cyclopentane[45]
#160	Tetrahydrofuran[52]
#214	Benzene[53]
#215	Pyridine[54]
#218	*s*-Triazine[49]
#266	Ethyl carbamate[51]
#286	Alanine[55]
#543	*n*-Hexane[44]
#659	Tetrahydropyran[52]
#940	Aniline[56]
#949	Phenol[57]
#1132	Isopropyl ether[51]
#2071	Cyclohexanone[58]
#4208	Monofluorobenzene[53]
#4336	*m*-Xylene[59]
#4563	*p*-Xylene[59]
#4591	Hydroquinone[57]
#4958	*o*-Xylene[59]
#5357	Benzaldehyde[60]
#7723	1,3-Cyclohexanedione[58]
#10793	1,4-Cyclohexanedione[58]
#12304	1,2-Cyclohexanedione[58]
#17954	Cyclooctatetraene[45]
#23866	Hexafluoroethane[46]

Table 4. List of the ID in QM9 dataset, molecule name, and reference to the original article of the 44 molecules extracted from the Gas Phase Core Excitation Database published and maintained by the Hitchcock group[33] for the comparison of experimental spectral data to our calculated spectral data. The superscript numbers to the name of the molecules are the reference numbers of the original literatures of the spectra used for the comparison.

HDF5 file, where *i* is the ID of the molecule in the QM9 dataset and *j* is the integer number corresponding to the count of the sites in the xyz file of the QM9 dataset starting from 0. The eigenvalues and the dynamical structure factors are stored in data groups named */i/j/eigen_values* and */i/j/dsf*, respectively. The eigenvalues are stored as a one-dimensional array with length n_{band} in the unit of eV, where n_{band} is the number of the calculated bands. The dynamical structure factors were calculated as $|\phi_c|r_\alpha|\psi_{n,k}|^2$ in the unit of Å^2 for the three $r_\alpha(=x, y, z)$ and eigenstate index n, and thus stored as a two-dimensional array with shape (n_{band}, 3). The number of electrons adopted in the calculation using pseudopotentials are stored as the attribute value "num_electrons" of each site group */i/j*. The site multiplicity, final energy of the ground state, final energy of the excited state, and excitation

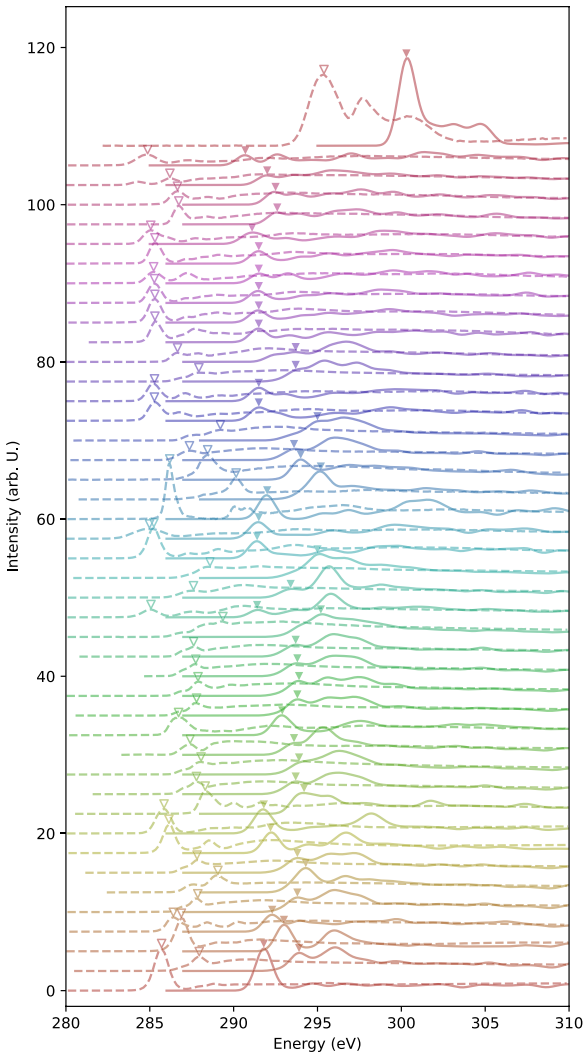

Fig. 2 Comparison of the shape of the spectra between the calculated spectra smoothed with Gaussian filter with the standard deviation of 0.5 eV (solid lines) and the 44 experimental spectra extracted from literatures summarized in Table 4 (dotted lines). The triangle markers denote the position of the first peak. Spectra are shifted for visibility, and the order of the spectra from bottom to top corresponds to the order of molecules listed in Table 4.

energy of the site are stored as the attribute values "multiplicity", "final_energy_gs", "final_energy_ex" and "excitation_energy" of the */i/j* group, respectively.

Note that the pairs of eigenvalues and dynamical structure factors below LUMO is also included. When forming a site-specific spectrum from these data, the dynamical structure factors below LUMO should be ignored, which can be achieved by ignoring the "num_electrons/2" levels from the lowest energy levels. In addition, for each eigenvalue, the dynamical scattering factor must be doubled because our calculation does not consider spin polarization. As for the site or molecular spectra, total non-directional spectra can be calculated as the average of the three $r_\alpha (=x, y, z)$. In order to get a molecular spectrum, the site-specific spectra should be multiplied by the site multiplicity followed by summed up over the symmetrically unique sites.

Site specific spectra. The information of the *j*-th site of the *i*-th molecule is stored in the */i/j* group in the HDF5 file. Note that only one representative site is included among the symmetrically equivalent carbon sites. The spectral data is stored in datasets named */i/j/spectrum* and */i/j/energies*. The first, second, third, and fourth columns of */i/j/spectrum* are the total, and the *x*, *y*, and *z* direction spectra in the unit of Å^2, respectively. */i/j/energies* is a one-dimensional array of the energy which is relative to the eigenvalue of the LUMO state in the unit of eV. The range of energies covers the minimum and maximum of energies relative to LUMO level with a margin of 5 eV. The site multiplicity, final energy of the ground state, final energy of the excited state, and excitation energy of the site are stored as the attribute values "multiplicity", "final_energy_gs", "final_energy_ex" and "excitation_energy" of the */i/j* group, respectively.

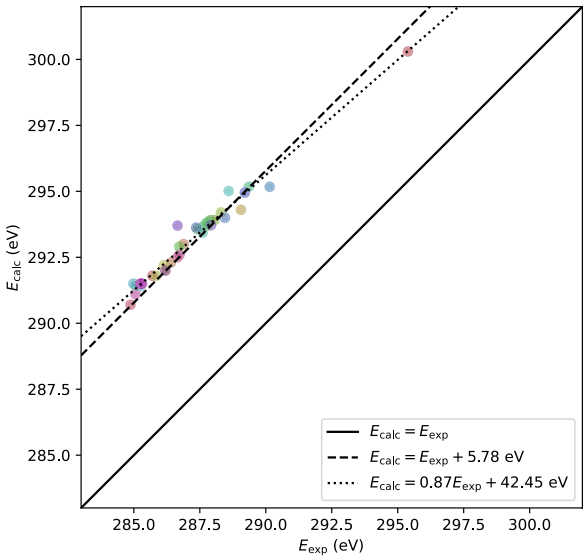

Fig. 3 Comparison in the energy position of first peak between the calculated spectra and experimental spectra in literatures, E_{calc} and E_{exp}, respectively. Solid line shows the line: $E_{calc} = E_{exp}$. Dashed line shows the line only considering energy shift: $E_{\mathrm{calc}} = E_{\mathrm{exp}} + 5.78\mathrm{eV}$. Dotted line shows the line obtained by linear regression: $E_{\mathrm{calc}} = 0.87E_{\mathrm{exp}} + 42.45\mathrm{eV}$. The colors of markers are the same for the line color in Figs. 2 and 4.

Molecular specific spectra. The information of *i*-th molecule is stored in the */i* group in the HDF5 file. The spectral data is stored in a data set called */i/spectrum*, where the first, second, third, and fourth columns are the total, and the *x*, *y*, and *z* direction spectra in the unit of Å^2, respectively. The energy corresponding to the */i/spectrum* in the unit of eV is stored in */i/energies*.

Spectra database files in csv format. This CSV format database provides spectral data that are smoothed with a Gaussian filter with a standard deviation of 0.5 eV and thinned out in 0.1 eV steps. The data for each molecule is stored in a directory with a directory name corresponding to the molecule number *i* in QM9. In each directory, three types of data are stored in CSV file format: spectral data of the molecule, site-specific spectral data, and data of the molecular structure. The file name of the molecular spectrum data is "*i*.csv", the file name of the site spectrum data is "*i_j_m*.csv", and the file name of the molecular structure data is "*i*_structure.csv", where *i* is the molecule number in the QM9 database, *j* is the site number in molecule *i*, and *m* is the multiplicity of site *j* in molecule *i*. The file format of the csv file for spectral data and structures are summarized in Tables 2 and 3.

Raw CASTEP files. The most of raw CASTEP input and output files for carbon sites are available at figshare[31] except for PBE pseudopotential files and spectra data files. This record covers input files which are necessary for recalculation and output files which are sufficient for confirming calculation process.

Technical Validation

Comparison to experimental spectra. Some of the calculated spectra were compared with experimental spectra. We extracted experimental spectral data of 44 molecules from the Gas Phase Core Excitation Database published and maintained by the Hitchcock group[33], and compared their energy position and shape of the spectrum with those of our calculated spectral data. The list of the 44 molecules used for comparison is summarized in Table 4.

Figure 2 shows the calculated spectra after applying smoothing with a Gaussian function of standard deviation of 0.5 eV and experimental spectra. A mismatch in the absolute value of transition energy can be seen between the experimental and calculated spectra. To analyze the difference in the transition energy, we extracted the positions of the first peak. Figure 3 shows a scatter plot of the energy position of the first peak in the experimental data and our calculations. The data points are distributed linearly, which confirms that it is a systematic misalignment. We tried linear regression and obtained a fitting line of $E_{\mathrm{calc}} = 0.87E_{\mathrm{exp}} + 42.45\mathrm{eV}$. It was also found that taking into account an energy shift of 5.78 eV, assuming the slope of 1.0, is a sufficient approximation. Practically, this means that energies of our calculation tends to be higher by 5.78 eV than the experiment in terms of energy. These systematic deviations indicate that the calculated transition energy are of sufficient quality to withstand comparison with that of experimental data by considering the systematic error of 5.78 eV. It should be noted that this level of systematic errors in the transition energies of core-loss spectrum calculations is generally observed in various calculation methods[16,18,34–38].

For comparison of the spectral features, the experimental spectrum and our calculated spectrum with the energy position shifted by 5.78 eV are shown in Fig. 4. There are a few molecules for which the spectral shapes do not match, but the agreement is generally good, taking into account the existence of the variation in the experimental spectral data depending on the measurement conditions. Those spectral features can satisfactorily

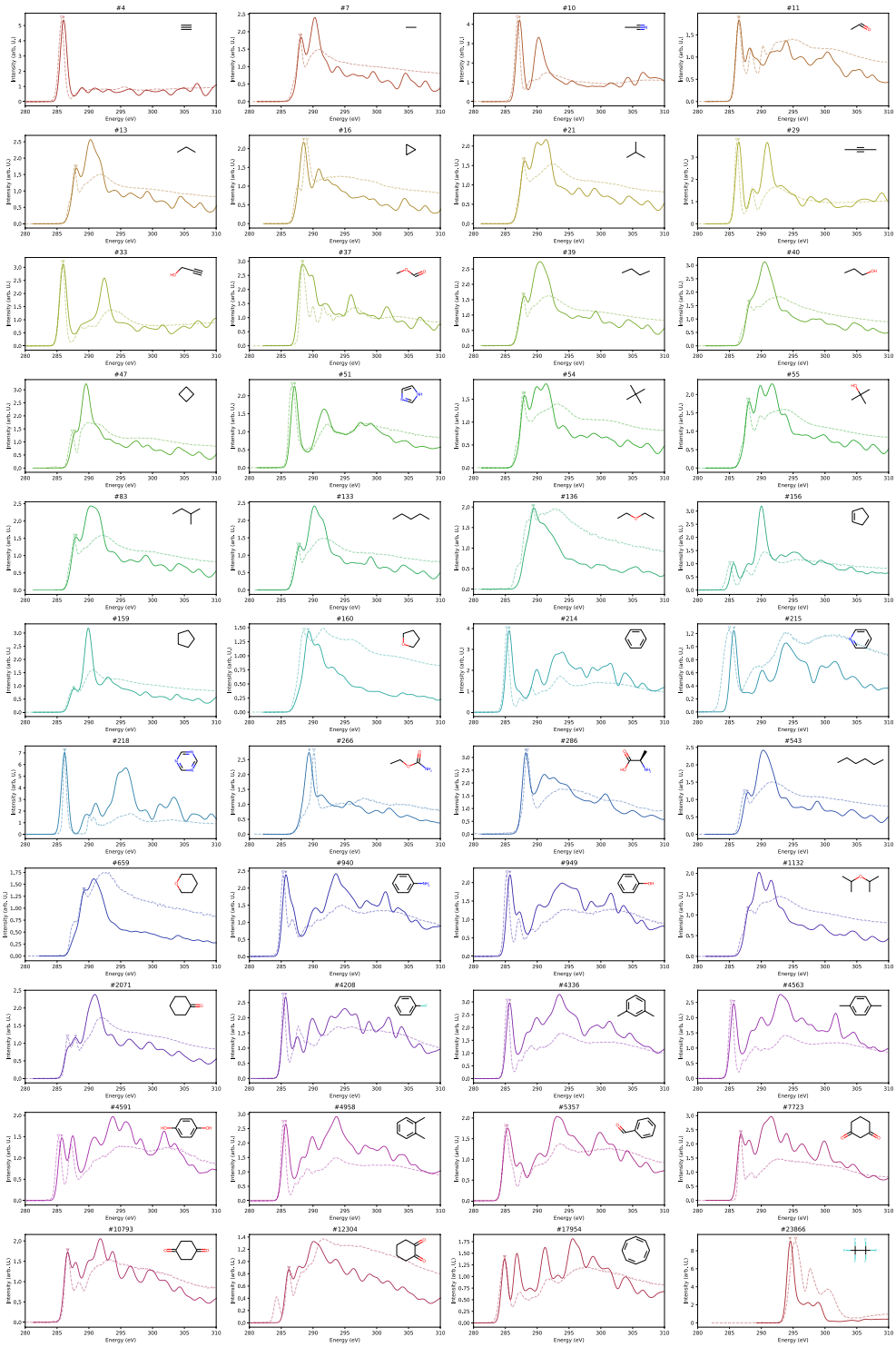

Fig. 4 Comparison of the shape of the spectra between the calculated spectra smoothed with Gaussian filter with the standard deviation of 0.5 eV and shifted by 5.78 eV (solid lines) and experimental spectra in literatures (dotted lines) for each molecule.

reproduce the experimental spectra, and the chemical shift between those peaks are quantitatively reproduced by the present method. It is worthwhile that the chemical shift among different molecules can be quantitatively reproduced by the present calculation.

The simulated spectral intensities tend to become weaker than that of the experiment in high energy regions above about 310 eV, which is attributed to the fact that the number of unoccupied levels is only considered up to 1,000 and only intra-atomic transitions are taken into account.

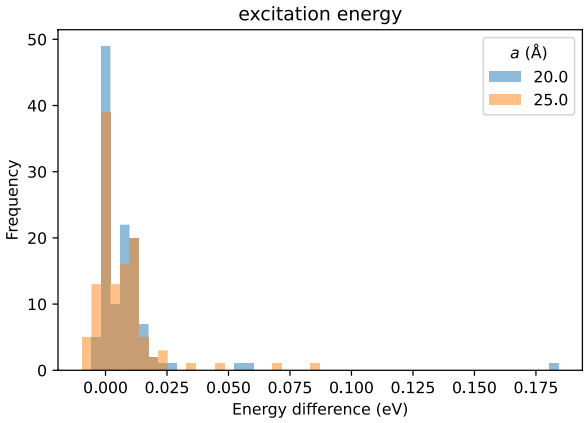

Fig. 5 Histogram of excitation energy difference for different cell sizes compared to that of the 15 Å cubic cell.

Concerning the effect of spin polarization, we also carried out calculations considering spin polarization for the 44 molecules, and found only a small difference in excitation energy less than 0.41 eV compared to that of non-magnetic calculation, which indicates the consideration of spin does not yield significant difference compared to the mismatch between simulated excitation energies and experimental ones.

Calculation cell size dependence. We checked calculation cell size dependence of the excitation energies for the 44 molecules in Table 4 from 15.0 to 25.0 Å cubic cell. Figure 5 shows the histograms of the deviations of excitation energies compared to those of 15.0 Å cubic cell except for the site 4 of #5357 whose calculations did not converged for the larger cell sizes. The largest deviation of the excitation energy was 184 meV, but most of them are located within 25 meV, which is relatively small compared to the realistic smearing width of the spectrum like 0.3 eV, and they do not have large impact when considered as data for comparison with experiments or for machine learning.

Usage Notes

In this work, we present raw files of CASTEP, which are of use for confirming the detail of calculation conditions. The spectra database files contain all the calculated spectra, and can be used for finger-print method and spectrum informatics. Spin polarization and spin-orbit interaction are not considered for all calculations. The spectra was calculated using 1,000 unoccupied states and only inner atom transition was considered, which yields mismatch to experimental results especially in high energy loss regions. The absolute value of the excitation energies tends to be overestimated systematically compared to the experimental values by about 6 eV as can be seen in Fig. 3, which might be due to the used of difference in atomic energies between ground states and excited states for the correction of energies calculated using pseudopotentials.

Code availability

A proprietary code, Academic Release CASTEP version 8.0[6–11] was used to perform DFT and core-loss spectra calculations. The configuration files used in the calculation is provided for reproducing the site C-K edge spectra at figshare[31] along with some of the output files for confirmation of calculation condition. For making input files, parsing output files, creating database and visualization, we have used the following python libraries: Numpy, Pandas, h5py, rdkit and Matplotlib. The Python code used for parsing the CASTEP input and output files is available at GitHub[32] under the MIT license. The Python code used for making the Gaussian smeared spectra dataset from the database HDF5 file of eigenvalues and dynamical structure factors is also available at GitHub[39] under the MIT license, and can be used for making spectra with arbitrary smearing parameters.

Received: 1 September 2021; Accepted: 1 April 2022;
Published: 16 May 2022

References

1. Ewels, P., Sikora, T., Serin, V., Ewels, C. P. & Lajaunie, L. A complete overhaul of the electron energy-loss spectroscopy and X-ray absorption spectroscopy database: eelsdb.eu. *Microscopy and Microanalysis* **22**, 717–724, https://doi.org/10.1017/S1431927616000179 (2016).
2. Mathew, K. *et al.* High-throughput computational X-ray absorption spectroscopy. *Scientific Data* **5**, 180151, https://doi.org/10.1038/sdata.2018.151 (2018).
3. Zheng, C. *et al.* Automated generation and ensemble-learned matching of X-ray absorption spectra. *npj Computational Materials* **4**, 1–9, https://doi.org/10.1038/s41524-018-0067-x (2018).
4. 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. *Journal of Chemical Information and Modeling* **52**, 2864–2875, https://doi.org/10.1021/ci300415d (2012).
5. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data* **1**, 140022, https://doi.org/10.1038/sdata.2014.22 (2014).
6. Hohenberg, P. & Kohn, W. Inhomogeneous electron gas. *Physical Review* **136**, B864–B871, https://doi.org/10.1103/PhysRev.136.B864 (1964).
7. Kohn, W. & Sham, L. J. Self-consistent equations including exchange and correlation effects. *Physical Review* **140**, A1133–A1138, https://doi.org/10.1103/PhysRev.140.A1133 (1965).

8. Monkhorst, H. J. & Pack, J. D. Special points for Brillouin-zone integrations. *Physical Review B* **13**, 5188–5192, https://doi.org/10.1103/PhysRevB.13.5188 (1976).
9. Payne, M. C., Teter, M. P., Allan, D. C., Arias, T. A. & Joannopoulos, J. D. Iterative minimization techniques for ab initio total-energy calculations: molecular dynamics and conjugate gradients. *Reviews of Modern Physics* **64**, 1045–1097, https://doi.org/10.1103/RevModPhys.64.1045 (1992).
10. Clark, S. J. *et al*. First principles methods using CASTEP. *Zeitschrift für Kristallographie - Crystalline Materials* **220**, 567–570, https://doi.org/10.1524/zkri.220.5.567.65075 (2005).
11. Gao, S.-P., Pickard, C. J., Perlov, A. & Milman, V. Core-level spectroscopy calculation and the plane wave pseudopotential method. *Journal of Physics: Condensed Matter* **21**, 104203, https://doi.org/10.1088/0953-8984/21/10/104203 (2009).
12. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Physical Review Letters* **77**, 3865–3868, https://doi.org/10.1103/PhysRevLett.77.3865 (1996).
13. Jones, R. O. & Gunnarsson, O. The density functional formalism, its applications and prospects. *Rev. Mod. Phys.* **61**, 689–746, https://doi.org/10.1103/RevModPhys.61.689 (1989).
14. Mizoguchi, T., Tanaka, I., Gao, S. P. & Pickard, C. J. First-principles calculation of spectral features, chemical shift and absolute threshold of ELNES and XANES using a plane wave pseudopotential method. *Journal of Physics Condensed Matter* **21**, https://doi.org/10.1088/0953-8984/21/10/104204 (2009).
15. Ljungberg, M., Mortensen, J. & Pettersson, L. An implementation of core level spectroscopies in a real space projector augmented wave density functional theory code. *Journal of Electron Spectroscopy and Related Phenomena* **184**, 427–439, https://doi.org/10.1016/j.elspec.2011.05.004 (2011).
16. Ikeno, H. & Mizoguchi, T. Basics and applications of ELNES calculations. *Journal of Electron Microscopy* **66**, 305–327, https://doi.org/10.1093/jmicro/dfx033 (2017).
17. Gao, S. P. Ab initio calculation of ELNES/XANES of BeO polymorphs. *Physica Status Solidi (B) Basic Research* **247**, 2190–2194, https://doi.org/10.1002/pssb.200945574 (2010).
18. Mizoguchi, T., Olovsson, W., Ikeno, H. & Tanaka, I. Theoretical ELNES using one-particle and multi-particle calculations. *Micron* **41**, 695–709, https://doi.org/10.1016/j.micron.2010.05.011 (2010).
19. Matsui, Y. & Mizoguchi, T. First principles calculation of oxygen K edge absorption spectrum of acetic acid: relationship between the spectrum and molecular dynamics. *Chemical Physics Letters* **649**, 92–96, https://doi.org/10.1016/j.cplett.2016.02.043 (2016).
20. Kepaptsoglou, D. *et al*. Electronic structure modification of ion implanted graphene: the spectroscopic signatures of p- and n-type doping. *ACS Nano* **9**, 11398–11407, https://doi.org/10.1021/acsnano.5b05305 (2015).
21. Matsui, Y., Seki, K., Hibara, A. & Mizoguchi, T. An estimation of molecular dynamic behaviour in a liquid using core-loss spectroscopy. *Scientific Reports* **3**, 3503, https://doi.org/10.1038/srep03503 (2013).
22. Katsukura, H., Miyata, T., Shirai, M., Matsumoto, H. & Mizoguchi, T. Estimation of the molecular vibration of gases using electron microscopy. *Scientific Reports* **7**, 16434, https://doi.org/10.1038/s41598-017-16423-0 (2017).
23. Katsukura, H., Miyata, T., Tomita, K. & Mizoguchi, T. Effect of the van der Waals interaction on the electron energy-loss near edge structure theoretical calculation. *Ultramicroscopy* **178**, 88–95, https://doi.org/10.1016/j.ultramic.2016.07.012 (2017).
24. Takahashi, O. & Pettersson, L. G. M. Functional dependence of core-excitation energies. *The Journal of Chemical Physics* **121**, 10339–10345, https://doi.org/10.1063/1.1809610 (2004).
25. Hybertsen, M. S. & Louie, S. G. Electron correlation in semiconductors and insulators: band gaps and quasiparticle energies. *Phys. Rev. B* **34**, 5390–5413, https://doi.org/10.1103/PhysRevB.34.5390 (1986).
26. Petersilka, M., Gossmann, U. J. & Gross, E. K. U. Excitation energies from time-dependent density-functional theory. *Phys. Rev. Lett.* **76**, 1212–1215, https://doi.org/10.1103/PhysRevLett.76.1212 (1996).
27. Slater, J. C. Statistical exchange-correlation in the self-consistent field. *Advances in Quantum Chemistry* **6**, 1–92 (1972).
28. Fujikawa, T. Theory of the X-Ray absorption near edge structure (XANES) at a deep $L_{2,3}$ edge studied by the short-range order multiple scattering theory. *Journal of the Physical Society of Japan* **52**, 4001–4007, https://doi.org/10.1143/JPSJ.52.4001 (1983).
29. Togo, A. & Tanaka, I. Spglib: a software library for crystal symmetry search. *arXiv:1808.01590v1 [cond-mat.mtrl-sci]* https://arxiv.org/abs/1808.01590 (2018).
30. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979, https://doi.org/10.1103/PhysRevB.50.17953 (1994).
31. Shibata, K., Kikumasa, K., Kiyohara, S. & Mizoguchi, T. C-K edge core-loss spectral database of organic molecules. *figshare* https://doi.org/10.6084/m9.figshare.c.5494395.v1 (2021).
32. Shibata, K. castep_elnes_parser. *GitHub* https://doi.org/10.5281/zenodo.6352252 (2021).
33. Hitchcock, A. & Mancini, D. Bibliography and database of inner shell excitation spectra of gas phase atoms and molecules. *Journal of Electron Spectroscopy and Related Phenomena* **67**, 1–12, https://doi.org/10.1016/0368-2048(94)87001-2 (1994).
34. Ching, W.-Y. & Rulis, P. X-ray absorption near edge structure/electron energy loss near edge structure calculation using the supercell orthogonalized linear combination of atomic orbitals method. *Journal of Physics: Condensed Matter* **21**, 104202, https://doi.org/10.1088/0953-8984/21/10/104202 (2009).
35. Hébert, C., Luitz, J. & Schattschneider, P. Improvement of energy loss near edge structure calculation using Wien2k. *Micron* **34**, 219–225, https://doi.org/10.1016/S0968-4328(03)00030-1 (2003).
36. Hébert, C. Practical aspects of running the WIEN2k code for electron spectroscopy. *Micron* **38**, 12–28, https://doi.org/10.1016/j.micron.2006.03.010 (2007).
37. Mo, S.-D. & Ching, W. Y. Ab initio calculation of the core-hole effect in the electron energy-loss near-edge structure. *Physical Review B* **62**, 7901–7907, https://doi.org/10.1103/PhysRevB.62.7901 (2000).
38. Hamann, D. R. & Muller, D. A. Absolute and approximate calculations of electron-energy-loss spectroscopy edge thresholds. *Physical Review Letters* **89**, 126404, https://doi.org/10.1103/PhysRevLett.89.126404 (2002).
39. Shibata, K. ck_edge_maker. *GitHub* https://doi.org/10.5281/zenodo.6352097 (2021).
40. Hitchcock, A. & Brion, C. Carbon K-shell excitation of C_2H_2, C_2H_4, C_2H_6 and C_6H_6 by 2.5 keV electron impact. *Journal of Electron Spectroscopy and Related Phenomena* **10**, 317–330, https://doi.org/10.1016/0368-2048(77)85029-9 (1977).
41. Ishii, I. *et al*. The σ^* molecular orbitals of perfluoroalkanes as studied by inner-shell electron energy loss and electron transmission spectroscopies. *Canadian Journal of Chemistry* **66**, 2104–2121, https://doi.org/10.1139/v88-336 (1988).
42. Hitchcock, A. P., Tronc, M. & Modelli, A. Electron transmission and inner-shell electron energy loss spectroscopy of acetonitrile, isocyanomethane, methyl thiocyanate, and isothiocyanatomethane. *The Journal of Physical Chemistry* **93**, 3068–3077, https://doi.org/10.1021/j100345a039 (1989).
43. Hitchcock, A. & Brion, C. Inner-shell excitation of formaldehyde, acetaldehyde and acetone studied by electron impact. *Journal of Electron Spectroscopy and Related Phenomena* **19**, 231–250, https://doi.org/10.1016/0368-2048(80)87006-X (1980).
44. Hitchcock, A. & Ishii, I. Carbon K-shell excitation spectra of linear and branched alkanes. *Journal of Electron Spectroscopy and Related Phenomena* **42**, 11–26, https://doi.org/10.1016/0368-2048(87)85002-8 (1987).
45. Hitchcock, A. P. *et al*. Carbon K-shell excitation of gaseous and condensed cyclic hydrocarbons: C_3H_6, C_4H_8, C_5H_8, C_5H_{10}, C_6H_{10}, C_6H_{12}, and C_8H_8. *The Journal of Chemical Physics* **85**, 4849–4862, https://doi.org/10.1063/1.451719 (1986).
46. Robin, M., Ishii, I., McLaren, R. & Hitchcock, A. Fluorination effects on the inner-shell spectra of unsaturated molecules. *Journal of Electron Spectroscopy and Related Phenomena* **47**, 53–92, https://doi.org/10.1016/0368-2048(88)85005-9 (1988).

47. Ishii, I. & Hitchcook, A. The oscillator strengths for C1s and O1s excitation of some saturated and unsaturated organic alcohols, acids and esters. *Journal of Electron Spectroscopy and Related Phenomena* **46**, 55–84, https://doi.org/10.1016/0368-2048(88)80005-7 (1988).
48. Hitchcock, A. & Ishii, I. Carbon K-shell excitation spectra of linear and branched alkanes. *Journal of Electron Spectroscopy and Related Phenomena* **42**, 11–26, https://doi.org/10.1016/0368-2048(87)85002-8 (1987).
49. Apen, E., Hitchcock, A. P. & Gland, J. L. Experimental studies of the core excitation of imidazole, 4,5-dicyanoimidazole, and s-triazine. *The Journal of Physical Chemistry* **97**, 6859–6866, https://doi.org/10.1021/j100128a019 (1993).
50. Ishii, I., McLaren, R., Hitchcock, A. P. & Robin, M. B. Inner-shell excitations in weak-bond molecules. *The Journal of Chemical Physics* **87**, 4344–4360, https://doi.org/10.1063/1.452893 (1987).
51. Urquhart, S. G., Hitchcock, A. P., Priester, R. D. & Rightor, E. G. Analysis of polyurethanes using core excitation spectroscopy. Part II: Inner shell spectra of ether, urea and carbamate model compounds. *Journal of Polymer Science Part B: Polymer Physics* **33**, 1603–1620, https://doi.org/10.1002/polb.1995.090331105 (1995).
52. Newbury, D. C., Ishii, I. & Hitchcock, A. P. Inner shell electron-energy loss spectroscopy of some heterocyclic molecules. *Canadian Journal of Chemistry* **64**, 1145–1155, https://doi.org/10.1139/v86-900 (1986).
53. Hitchcock, A. P., Fischer, P., Gedanken, A. & Robin, M. B. Antibonding σ^* valence MOs in the inner-shell and outer-shell spectra of the fluorobenzenes. *The Journal of Physical Chemistry* **91**, 531–540, https://doi.org/10.1021/j100287a009 (1987).
54. Horsley, J. A. *et al*. Resonances in the K shell excitation spectra of benzene and pyridine: Gas phase, solid, and chemisorbed states. *The Journal of Chemical Physics* **83**, 6099–6107, https://doi.org/10.1063/1.449601 (1985).
55. Cooper, G. *et al*. Inner shell excitation of glycine, glycyl-glycine, alanine and phenylalanine. *Journal of Electron Spectroscopy and Related Phenomena* **137–140**, 795–799, https://doi.org/10.1016/j.elspec.2004.02.102 (2004).
56. Turci, C. C., Urquhart, S. G. & Hitchcock, A. P. Inner-shell excitation spectroscopy of aniline, nitrobenzene, and nitroanilines. *Canadian Journal of Chemistry* **74**, 851–869, https://doi.org/10.1139/v96-094 (1996).
57. Francis, J. T. & Hitchcock, A. P. Inner-shell spectroscopy of p-benzoquinone, hydroquinone, and phenol: distinguishing quinoid and benzenoid structures. *The Journal of Physical Chemistry* **96**, 6598–6610, https://doi.org/10.1021/j100195a018 (1992).
58. Francis, J. T. & Hitchcock, A. P. Distinguishing keto and enol structures by inner-shell spectroscopy. *The Journal of Physical Chemistry* **98**, 3650–3657, https://doi.org/10.1021/j100065a018 (1994).
59. Eustatiu, I., Huo, B., Urquhart, S. & Hitchcock, A. Isomeric sensitivity of the C 1s spectra of xylenes. *Journal of Electron Spectroscopy and Related Phenomena* **94**, 243–252, https://doi.org/10.1016/S0368-2048(98)00189-3 (1998).
60. Hitchcock, A. P., Urquhart, S. G. & Rightor, E. G. Inner-shell spectroscopy of benzaldehyde, terephthalaldehyde, ethylbenzoate, terephthaloyl chloride and phosgene: models for core excitation of poly(ethylene terephthalate). *The Journal of Physical Chemistry* **96**, 8736–8750, https://doi.org/10.1021/j100201a015 (1992).

Acknowledgements

This study was supported by the Grant-in-Aid for Scientific Research (Grant Nos. 17H06094, 19H05787, 19H00818, and 20J00773) from the MEXT and CREST (Grant No. JPMJCR1993) from the JST.

Author contributions

T.M. conceived the project, K.K. and S.K. carried out the calculation, K.K., S.K., K.S. and T.M. analysed the results. K.S. wrote the manuscript. All authors reviewed the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to K.S. or T.M.

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 Author(s) 2022