scientific data



OPEN

DATA DESCRIPTOR

# VIB5 database with accurate ab initio quantum chemical molecular potential energy surfaces

Lina Zhang[1], Shuang Zhang[1], Alec Owens[2] ✉, Sergei N. Yurchenko[2] & Pavlo O. Dral[1] ✉

**High-level *ab initio* quantum chemical (QC) molecular potential energy surfaces (PESs) are crucial for accurately simulating molecular rotation-vibration spectra. Machine learning (ML) can help alleviate the cost of constructing such PESs, but requires access to the original *ab initio* PES data, namely potential energies computed on high-density grids of nuclear geometries. In this work, we present a new structured PES database called VIB5, which contains high-quality *ab initio* data on 5 small polyatomic molecules of astrophysical significance ($CH_3Cl$, $CH_4$, $SiH_4$, $CH_3F$, and NaOH). The VIB5 database is based on previously used PESs, which, however, are either publicly unavailable or lacking key information to make them suitable for ML applications. The VIB5 database provides tens of thousands of grid points for each molecule with theoretical best estimates of potential energies along with their constituent energy correction terms and a data-extraction script. In addition, new complementary QC calculations of energies and energy gradients have been performed to provide a consistent database, which, e.g., can be used for gradient-based ML methods.**

## Background & Summary

Many physical and chemical processes of molecular systems are governed by potential energy surfaces (PESs) that are functions of potential energy with respect to the molecular geometry defined by the nuclei[1]. Accurate *ab initio* quantum chemical (QC) molecular PESs are essential to predict and understand a multitude of physicochemical properties of interest such as reaction thermodynamics, kinetics[2], and simulation of rovibrational spectra[3–5]. As for the latter, PESs of a number of different molecules have been constructed and used in variational nuclear motion calculations to provide accurate rotation-vibration-electronic line lists to aid the characterization of exoplanet atmospheres, amongst other applications[6–16].

It is necessary to have a global PES covering all relevant regions of nuclear configurations allowing to simulate rotation-vibration (rovibrational) spectra approaching the coveted spectroscopic accuracy of 1 $cm^{-1}$ in a broad range of temperatures. This can be achieved by defining the PES on a high-density grid of nuclear geometries with no holes and having the theoretical best estimate (TBE) of energies computed at a very high QC level of theory. The construction of an optimal grid usually involves many steps and human intervention, and often requires a staggeringly large number of grid points, e.g., ca. 100 thousand points even for a five-atom molecule such as methane[10]. The choice of QC level for TBE calculations is determined by the trade-off between accuracy and computational cost, but typically requires going well beyond the gold-standard[17–19] CCSD(T)[17]/CBS (coupled cluster with single and double excitations and a perturbative treatment of triple excitations/complete basis set) limit and needs many QC corrections on top of it. Just to give a perspective, ca. 24 single processing unit (CPU)-hours are required for calculating TBE energy of each grid point of ~45 thousand methyl chloride ($CH_3Cl$) geometries amounting to over 100 CPU-years when constructing its highly accurate ab initio PES[20].

To reduce the high computational cost, machine learning (ML) has emerged as a powerful approach for constructing full-dimensional PESs[21–27] and the resulting ML PESs can be used[22,24,28–35] for performing vibrational calculations. In particular, substantial cost reduction can be achieved by calculating TBE energies only for a small number of existing grid points and then interpolating between them with ML[36]; such ML grids can be subsequently used for simulating rovibrational spectra with a relatively small loss of accuracy. Importantly,

[1]State Key Laboratory of Physical Chemistry of Solid Surfaces, Fujian Provincial Key Laboratory of Theoretical and Computational Chemistry, Department of Chemistry, and College of Chemistry and Chemical Engineering, Xiamen University, Xiamen, 361005, China. [2]Department of Physics and Astronomy, University College London, Gower Street, WC1E 6BT, London, United Kingdom. ✉e-mail: alec.owens.13@ucl.ac.uk; dral@xmu.edu.cn

| Molecule | Grid size | Reference |
|---|---|---|
| $CH_3Cl$ | 44819 | [7,9,20] |
| $CH_4$ | 97217[a] | [10] |
| $SiH_4$ | 84002 | [8] |
| $CH_3F$ | 82653 | [12] |
| NaOH | 15901 | [14] |
| Total: 5 molecules | 324592[a] | |

**Table 1.** The number of grid points (grid size) for each molecule with references to original studies generating these grid points, theoretical best estimates (TBE), and TBE constituent terms. [a]The number of grid points is slightly smaller than that reported in the original publications as we found very few duplicates in the original data set. See section *Technical Validation*.

much larger savings in computational cost can be achieved[20], when ML is applied to learn various QC corrections using a hierarchical ML (hML) scheme based on Δ-learning[37] rather than to learn the TBE energy directly.

Despite all the above efforts in constructing highly accurate PESs, there is still room for improvement, e.g., via creating denser grids, using higher QC levels, and further development of ML approaches, all of which requires access to data. Unfortunately, the raw data containing geometries, TBEs and TBE constituent terms for many published studies is either missing or scattered. Thus, our data descriptor aims to organize these scattered data generated in the previous studies by some of us into a consolidated, structured PES database that we call VIB5. The VIB5 database contains five molecules $CH_3Cl$[7,9,20], $CH_4$[10], $SiH_4$[8], $CH_3F$[12], and NaOH[14]. The number of grid points ranges from 15 thousand to 100 thousand; altogether more than 300 thousand points (Table 1). In addition, it is also known that inclusion of the energy gradient information can significantly reduce the number of training points for ML, which is efficiently exploited in the gradient-based ML models[38,39]. Thus, for this database, we additionally calculate energies and energy gradients at two levels of theory, MP2/cc-pVTZ (second order Møller-Plesset perturbation theory/correlation-consistent triple-zeta basis set) and CCSD(T)/cc-pVQZ (correlation consistent quadruple-zeta basis set), and provide the HF (Hartree–Fock) energies calculated with the corresponding basis sets cc-pVTZ and cc-pVQZ.

Our database is complementary to existing databases used for developing ML PES models. Some existing databases contain only energies for equilibrium geometries of various compounds calculated at different levels (from density functional theory [DFT] up to coupled-cluster approaches): QM7[40], QM7b[41], QM9[42], revised QM9[43], and ANI-1ccx[44]. Another database (ANI-1[45]) also contains energies at DFT for off-equilibrium geometries. Energies and energy gradients at DFT are available for equilibrium and off-equilibrium geometries of different molecules in the ANI-1x[44] and QM7-X[46] databases. The MD-17 dataset[38,39] is a popular database with energies and energy gradients for geometries taken from MD trajectories of several small- to medium-sized molecules at DFT and for subset of points at CCSD(T) with different basis sets. PESs generated from MD are, however, likely to have limited coverage of high-energy geometries and many holes, making them inapplicable to some kinds of accurate simulations such as diffusion Monte Carlo calculations as was pointed out recently[47]. In contrast to these databases, our database provides reliable, global PESs with QC energies and energy gradients at different levels including very accurate TBEs of energies going beyond CCSD(T)/CBS, which can be used for ML models trained on data from several levels of theory, such as hML, Δ-learning, etc. Finally, our database comes with a convenient data-extraction script that can be used to pull the required information in a suitable format for, e.g., ML.

## Methods

**Grid points generation.** For each molecule, we take grid points directly from the previous studies by some of the authors. Here we only describe in short how these grid points were generated for the sake of completeness. We refer the reader to the original publications cited for each molecule for further details (see Table 1).

*$CH_3Cl$.* 44819 grid points for $CH_3Cl$ were taken from Refs. [7,9,20]. A Monte Carlo random energy-weighted sampling algorithm was applied to nine internal coordinates of $CH_3Cl$: the C–Cl bond length $r_0$; three C–H bond lengths $r_1$, $r_2$, and $r_3$; three ∠($H_i$CCl) interbond angles $\beta_1$, $\beta_2$, and $\beta_3$; and two dihedral angles $\tau_{12}$ and $\tau_{13}$ between adjacent planes containing $H_i$CCl and $H_j$CCl (Fig. 1a). This procedure led to geometries in the range $1.3 \le r_0 \le 2.95$ Å, $0.7 \le r_i \le 2.45$ Å, $65 \le \beta_i \le 165°$ for $i = 1, 2, 3$ and $55 \le \tau_{jk} \le 185°$ with $jk = 12, 13$. The grid also includes 1000 carefully chosen low-energy points to ensure an adequate description of the equilibrium region.

*$CH_4$.* 97271 grid points for $CH_4$ were taken from ref. [10]. The global grid was built in the same fashion as the grid was constructed for $CH_3Cl$. Nine internal coordinates of $CH_4$ are defined as follows: four C–H bond lengths $r_1$, $r_2$, $r_3$ and $r_4$; five∠($H_j$-C-$H_k$) interbond angles $\alpha_{12}$, $\alpha_{13}$, $\alpha_{14}$, $\alpha_{23}$, and $\alpha_{24}$, where $j$ and $k$ label the respective hydrogen atoms (Fig. 1b). Then grid points are in the range $0.71 \le r_i \le 2.60$ Å for $i = 1, 2, 3, 4$ and $40 \le \alpha_{jk} \le 140°$ with $jk = 12, 13, 14, 23, 24$.

*$SiH_4$.* 84002 grid points for $SiH_4$ were taken from ref. [8]. Nine internal coordinates of $SiH_4$ are defined in the same way as $CH_4$: four Si–H bond lengths $r_1$, $r_2$, $r_3$ and $r_4$; five∠($H_j$-Si-$H_k$) interbond angles $\alpha_{12}$, $\alpha_{13}$, $\alpha_{14}$,

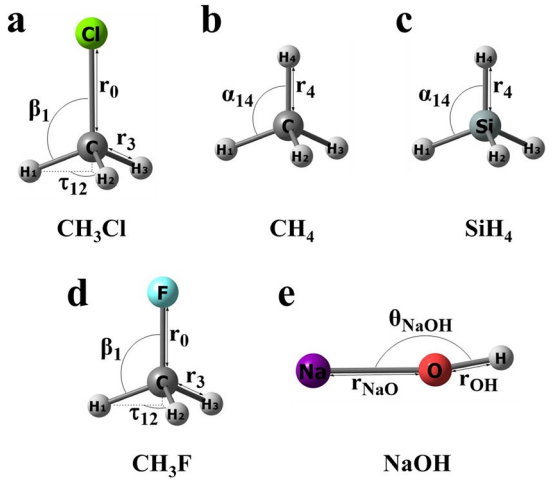


**Fig. 1** Definition of internal coordinates in each molecule. Internal coordinates of (**a**) $CH_3Cl$; $r_0$ is C–Cl bond length, $r_i$ and $\beta_i$ are C–$H_i$ bond lengths and $\angle(H_iCCl)$ angles ($I = 1, 2, 3$), $\tau_{jk}$ are $H_jCClH_k$ dihedral angles ($jk = 12, 13$); only $r_0$, $r_3$, $\beta_1$ and $\tau_{12}$ are shown; (**b**) $CH_4$; $r_i$ and $\alpha_{jk}$ are C–$H_i$ bond lengths and $\angle(H_jCH_k)$ angles ($i = 1, 2, 3, 4$; $jk = 12, 13, 14, 23, 24$); only $r_4$ and $\alpha_{14}$ are shown; (**c**) $SiH_4$; $r_i$ and $\alpha_{jk}$ are Si–$H_i$ bond lengths and $\angle(H_jSiH_k)$ angles ($i = 1, 2, 3, 4$; $jk = 12, 13, 14, 23, 24$); only $r_4$ and $\alpha_{14}$ are shown; (**d**) $CH_3F$; $r_0$ is C–F bond length, $r_i$ and $\beta_i$ are C–$H_i$ bond lengths and $\angle(H_iCF)$ angles ($i = 1, 2, 3$), $\tau_{jk}$ are $H_jCFH_k$ dihedral angles ($jk = 12, 13$); only $r_0$, $r_3$, $\beta_1$ and $\tau_{12}$ are shown; (**e**) NaOH; $r_{NaO}$ and $r_{OH}$ are Na–O and O–H bond lengths, $\theta_{NaOH}$ is $\angle$(NaOH) bond angle.

$\alpha_{23}$, and $\alpha_{24}$, where $j$ and $k$ label the respective hydrogen atoms (Fig. 1c). Then geometries are in the range $0.98 \le r_i \le 2.95$ Å for $i = 1, 2, 3, 4$ and $40 \le \alpha_{jk} \le 140°$ with $jk = 12, 13, 14, 23, 24$.

*$CH_3F$.* 82653 grid points for $CH_3F$ were taken from ref. [12]. Nine internal coordinates of $CH_3F$ are defined in the same way as $CH_3Cl$: the C–F bond length $r_0$; three C–H bond lengths $r_1$, $r_2$, and $r_3$; three $\angle(H_iCF)$ interbond angles $\beta_1$, $\beta_2$, and $\beta_3$; and two dihedral angles $\tau_{12}$ and $\tau_{13}$ between adjacent planes containing $H_iCF$ and $H_jCF$ (Fig. 1d). This procedure led to geometries in the range $1.005 \le r_0 \le 2.555$ Å, $0.705 \le r_i \le 2.695$ Å, $45.5 \le \beta_i \le 169.5°$ for $i = 1, 2, 3$ and $40.5 \le \tau_{jk} \le 189.5°$ with $jk = 12, 13$.

*NaOH.* 15901 grid points for NaOH were taken from ref. [14]. Grid points were generated randomly with a dense distribution around the equilibrium region. Three internal coordinates of NaOH are defined as follows: the Na–O bond length $r_{NaO}$, the O–H bond length $r_{OH}$, and the interbond angle $\angle$(NaOH) (Fig. 1e). This procedure led to geometries in the range $1.435 \le r_{NaO} \le 4.400$ Å, $0.690 \le r_{OH} \le 1.680$ Å, and $40 \le \angle(NaOH) \le 180°$.

**Theoretical best estimates and constituent terms.** For each molecule, we take the TBEs and energy corrections directly from the previous studies by some of us. Here we only briefly introduce how these calculations were performed. We refer the reader to the original publications cited for each molecule for details (see Table 1). TBE is obtained through the sum of many constituent terms: $E_{CBS}$, $\Delta E_{CV}$, $\Delta E_{HO}$, $\Delta E_{SR}$, and, for most molecules, $\Delta E_{DBOC}$. $E_{CBS}$ means the energy at the complete basis set (CBS) limit. $\Delta E_{CV}$ refers to the core-valence (CV) electron correlation energy correction. $\Delta E_{HO}$ refers to the energy correction accounted for by the higher-order (HO) coupled cluster terms and $\Delta E_{SR}$ shows scalar relativistic (SR) effects. $\Delta E_{DBOC}$ means the diagonal Born–Oppenheimer correction and was calculated for $CH_3Cl$, $CH_4$, $CH_3F$, and NaOH, but not for $SiH_4$ due to the little effect of $\Delta E_{DBOC}$ on the vibrational energy levels of this molecule.

The constituent terms were not calculated at the same level of theory across all molecules in the data set. The computational details of five TBE constituent terms ($E_{CBS}$, $\Delta E_{CV}$, $\Delta E_{HO}$, $\Delta E_{SR}$, and $\Delta E_{DBOC}$) for 5 molecules are shown below and summarized in the Table 2.

*$E_{CBS}$.* To extrapolate the energy to the CBS limit, the parameterized, two-point formula[48] ($E^C_{CBS} = (E_{n+1} - E_n)F^C_{n+1} + E_n$) was used. In this process, the method CCSD(T)-F12b[49] and two basis sets cc-pVTZ-F12 and cc-pVQZ-F12[50] were chosen. When performing calculations, the frozen core approximation was adopted and the diagonal fixed amplitude ansatz 3C(FIX)[51] with a Slater geminal exponent value[48] of $\beta = 1.0$ $a_0^{-1}$ were employed. As for the auxiliary basis sets (ABS), the resolution of the identity OptRI[52] basis and cc-pV5Z/JKFIT[53] and aug-cc-pwCV5Z/MP2FIT[54] basis sets for density fitting were used for all 5 molecules. These calculations were carried out with either MOLPRO2012[55] ($CH_3Cl$, $CH_4$, $SiH_4$, $CH_3F$) or MOLPRO2015[55,56] (NaOH). As for the coefficients $F^C_{n+1}$ in this two-point formula, $F^{CCSD\text{-}F12b} = 1.363388$ and $F^{(T)} = 1.769474$[48] were used for all molecules. The extrapolation was not applied to the Hartree–Fock (HF) energy and the HF + CABS (complementary auxiliary basis set) singles correction[49] calculated with the cc-pVQZ-F12 basis set was used.

| Molecule | $E_{CBS}$ | $\Delta E_{CV}$ | $\Delta E_{HO}$ | $\Delta E_{SR}$ | $\Delta E_{DBOC}$ |
|---|---|---|---|---|---|
| $CH_3Cl$ | Software: MOLPRO2012 | The basis set: cc-pCVQZ-F12; Slater geminal exponent value $\beta = 1.5\ a_0^{-1}$; all-electron calculations kept the 1*s* orbital of Cl frozen; Software: MOLPRO2012 | Levels of theory: CCSD(T), CCSDT, and CCSDT(Q); Basis sets for the full triples and the perturbative quadruples calculations are aug-cc-pVTZ(+d for Cl) and aug-cc-pVDZ(+d for Cl), respectively. | Method: one-electron mass velocity and Darwin (MVD1) terms from the Breit–Pauli Hamiltonian in first-order perturbation theory; All electrons correlated (except for the 1*s* of Cl); CCSD(T)/aug-cc-pCVTZ(+d for Cl). Software: CFOUR | The 1*s* orbital of Cl is frozen and all other electrons are correlated; basis set: aug-cc-pCVTZ (+d for Cl) |
| $CH_4$ | Software: MOLPRO2012 | The basis set: cc-pCVTZ-F12; Slater geminal exponent value $\beta = 1.4\ a_0^{-1}$; No frozen orbital; Software: MOLPRO2012 | Levels of theory: CCSD(T), CCSDT, and CCSDT(Q); Basis sets for the full triples and the perturbative quadruples calculations are cc-pVTZ and cc-pVDZ, respectively. | Method: the second-order Douglas–Kroll–Hess approach; frozen core approximation; CCSD(T)/cc-pVQZ-DK. Software: MOLPRO2012 | All electrons are correlated; basis set: aug-cc-pCVDZ |
| $SiH_4$ | Software: MOLPRO2012 | The basis set: cc-pCVTZ-F12; Slater geminal exponent value $\beta = 1.4\ a_0^{-1}$; all-electron calculations kept the 1*s* orbital of Si frozen; Software: MOLPRO2012 | Levels of theory: CCSD(T), CCSDT, and CCSDT(Q); basis sets for the full triples and the perturbative quadruples calculations are cc-pVTZ(+d for Si) and cc-pVDZ(+d for Si), respectively. | Method: the second-order Douglas-Kroll-Hess approach; frozen core approximation; CCSD(T)/cc-pVQZ-DK. Software: MOLPRO2012 | The correction was not included. |
| $CH_3F$ | Software: MOLPRO2012 | The basis set: cc-pCVTZ-F12; Slater geminal exponent value $\beta = 1.4\ a_0^{-1}$; no frozen orbital; Software: MOLPRO2012 | Levels of theory: CCSD(T), CCSDT, and CCSDT(Q); basis sets for the full triples and the perturbative quadruples calculations are cc-pVTZ and cc-pVDZ, respectively. | Method: the second-order Douglas–Kroll–Hess approach; frozen core approximation; CCSD(T)/cc-pVQZ-DK. Software: MOLPRO2012 | All electrons are correlated; basis set: aug-cc-pCVDZ |
| NaOH | Software: MOLPRO2015 | The basis set: cc-pCVTZ-F12; Slater geminal exponent value $\beta = 1.4\ a_0^{-1}$; all-electron calculations kept the 1*s* orbital of sodium frozen; Software: MOLPRO2015 | Levels of theory: CCSD(T) and CCSDT; basis set: cc-pVTZ(+d for Na). | Method: the second-order Douglas–Kroll–Hess approach; frozen core approximation; CCSD(T)/cc-pVQZ-DK. Software: MOLPRO2015 | The 1*s* orbital of Na is frozen and all other electrons are correlated; basis set: aug-cc-pCVDZ(+d for Na) |

**Table 2.** The comparative table of the computational details behind the calculations of the constituent terms of theoretical best estimates for five molecules of the VIB5 database. This table mainly emphasizes differences for each molecule, rather than giving the full description of computational details.

$\Delta E_{CV}$. $\Delta E_{CV}$ was computed at CCSD(T)-F12b/cc-pCVQZ-F12[57] for $CH_3Cl$ and at CCSD(T)-F12b/cc-pCVTZ-F12[57] for the other 4 molecules ($CH_4$, $SiH_4$, $CH_3F$, NaOH). The same ansatz and ABS used for $E_{CBS}$ were employed for calculating $\Delta E_{CV}$ but the Slater geminal exponent value was changed: $\beta = 1.5\ a_0^{-1}$ for $CH_3Cl$ and $\beta = 1.4\ a_0^{-1}$ for the other 4 molecules. For this term, all-electron calculations were adopted, but with the 1*s* orbital of Cl frozen for $CH_3Cl$, the 1*s* orbital of Si frozen for $SiH_4$, and the 1*s* orbital of Na frozen for NaOH. There is no frozen orbital in all-electron calculations for $CH_4$ and $CH_3F$. As for the software used, see the above $E_{CBS}$ part.

$\Delta E_{HO}$. To obtain $\Delta E_{HO}$, the hierarchy of coupled cluster methods was used. $\Delta E_{HO} = E_{CCSDT} - E_{CCSD(T)}$ for NaOH, while $\Delta E_{HO} = \Delta E_T + \Delta E_{(Q)}$ for other 4 molecules ($CH_3Cl$, $CH_4$, $SiH_4$, $CH_3F$) with $\Delta E_T = E_{CCSDT} - E_{CCSD(T)}$ for full triples contribution and $\Delta E_{(Q)} = E_{CCSDT(Q)} - E_{CCSDT}$ for perturbative quadruples contribution. The frozen core approximation was employed in the calculations. Thus, energy calculations at CCSD(T) and CCSDT were performed for NaOH, while energy calculations at CCSD(T), CCSDT, and CCSDT(Q) levels of theory were performed for other 4 molecules. All of these calculations were carried out through the general coupled cluster approach[58,59] implemented in the MRCC code (www.mrcc.hu)[60] interfaced to CFOUR (www.cfour.de)[61]. As for the basis set, aug-cc-pVTZ(+d for Cl)[62–65] & aug-cc-pVDZ(+d for Cl), cc-pVTZ[62] & cc-pVDZ, cc-pVTZ(+d for Si)[62–65] & cc-pVDZ(+d for Si), and cc-pVTZ[62] & cc-pVDZ for full triples and the perturbative quadruples of $CH_3Cl$, $CH_4$, $SiH_4$, and $CH_3F$. For NaOH, cc-pVTZ(+d for Na)[62,66] were used for CCSD(T) and CCSDT calculations.

$\Delta E_{SR}$. $\Delta E_{SR}$ was calculated by using either one-electron mass velocity and Darwin (MVD1) terms from the Breit–Pauli Hamiltonian in first-order perturbation theory[67] or the second-order Douglas–Kroll–Hess approach[68,69]. The former method was used for $CH_3Cl$ and the latter method was used for the other 4 molecules ($CH_4$, $SiH_4$, $CH_3F$, and NaOH). All-electron calculations (except for the 1*s* orbital of Cl) was adopted for $CH_3Cl$ while the frozen core approximation was employed for the other 5 molecules. Calculations were performed at CCSD(T)/aug-cc-pCVTZ(+d for Cl)[70,71] using the MVD1 approach[72] implemented in CFOUR for $CH_3Cl$ and at CCSD(T)/cc-pVQZ-DK[73] using MOLPRO (software versions the same as mentioned in the above $E_{CBS}$ part) for other 4 molecules.

$\Delta E_{DBOC}$. $\Delta E_{DBOC}$ was computed using the CCSD method[74] as implemented in CFOUR. This correction was not included for $SiH_4$. For this term, all-electron calculations were adopted, but with the 1*s* orbital of Cl frozen for

**a MP2/cc-pVTZ**

```
*CFOUR(CALC_LEVEL=MP2,BASIS=cc-pVTZ
SYMMETRY=ON
SCF_PROG=1
SCF_EXPSTART=3
GEO_CONV=10
GEO_MAXCYC=1
MEMORY=800000000)
```

**b CCSD(T)/cc-pVQZ for $CH_3Cl$, $CH_4$, $CH_3F$, NaOH**

```
*CFOUR(CALC_LEVEL=CCSD(T),BASIS=cc-pVQZ
MULTIPLICITY=1
CHARGE=0
FROZEN_CORE=ON
CC_CONV=10,CC_MAXCYC=100,CC_PROG=ECC
SCF_CONV=10,SCF_MAXCYC=100
LINEQ_CONV=8,LINEQ_MAXCYC=100
SYMMETRY=ON
MEMORY=800000000
GEO_CONV=10
GEO_MAXCYC=1
ABCDTYPE=AOBASIS)
```

**c CCSD(T)/cc-pVQZ for $SiH_4$**

```
*CFOUR(CALC_LEVEL=CCSD(T),BASIS=cc-pVQZ
SYMMETRY=ON
FROZEN_CORE=ON
SCF_PROG=1
SCF_EXPSTART=3
SCF_DAMPING=500
GEO_CONV=10
GEO_MAXCYC=1
MEMORY=800000000
ABCDTYPE=AOBASIS)
```

**Fig. 2** Typical CFOUR input options for (**a**) MP2/cc-pVTZ, (**b**) CCSD(T)/cc-pVQZ for $CH_3Cl$, $CH_4$, $CH_3F$, NaOH and (**c**) CCSD(T)/cc-pVQZ for $SiH_4$. The blue options were used for most cases and the light grey options are examples of options used to improve SCF convergence only for some geometries.

$CH_3Cl$, all electrons correlated for $CH_4$ and $CH_3F$, and the 1*s* orbital of Na frozen for NaOH. As for the basis set, calculations were performed at aug-cc-pCVTZ (+d for Cl) for $CH_3Cl$, aug-cc-pCVDZ for $CH_4$, aug-cc-pCVDZ for $CH_3F$, and aug-cc-pCVDZ(+d for Na) for NaOH.

**Complementary energy and gradient calculations.** All complementary *ab initio* QC energy and gradient calculations for a total of 324592 grid points were performed with two levels of theory: MP2[75,76]/cc-pVTZ[62,64,66] and CCSD(T)[17,77,78]/cc-pVQZ[62,64,66] using the CFOUR program package (Versions 1.0 and 2.1[61]; we use CFOUR V2.1 to perform calculations for some grid points in $CH_3Cl$ and NaOH that converge to high energy solutions); see Fig. 2 for the CFOUR input options. In the MP2/cc-pVTZ calculations, we use the default option FROZEN_CORE = OFF so that all electrons and all orbitals are correlated. In the CCSD(T)/cc-pVQZ calculations, the option FROZEN_CORE = ON is used for all molecules to allow valence electrons correlation alone. For $CH_3Cl$, $CH_4$, $CH_3F$ and NaOH, SCF_CONV = 10, CC_CONV = 10 and LINEQ_CONV = 8 are set to specify the convergence criterion for the HF-SCF, CC amplitude and linear equations and CC_PROG = ECC is set to specify that the CC program we used is ECC. For $SiH_4$, we adopted CFOUR default options SCF_CONV = 7, CC_CONV = 7, LINEQ_CONV = 7 and CC_PROG = VCC. We use GEO_MAXCYC = 1 option to set the maximum number of geometry optimization iterations to one to obtain the gradient information of the current nuclear configuration. From these calculations we also extracted HF energies calculated with the corresponding basis sets cc-pVTZ and cc-pVQZ. In addition, for $CH_3Cl$ we include MP2/aug-cc-pVQZ energies calculated using MOLPRO2012[55] as reported in ref. [20].

## Data Records

All data of 5 molecules are stored as a database in JSON format in the file named VIB5.json available for download from https://doi.org/10.6084/m9.figshare.16903288[79]. The first level of the database contains an item corresponding to each molecule in the order of $CH_3Cl$, $CH_4$, $SiH_4$, $CH_3F$, and NaOH. For each molecule, at the next level of the database, chemical formula, chemical name, number of atoms, list of nuclear charges in the same order as they appear in the items with nuclear coordinates are given at first, then the description of properties available for grid points (property type, levels of theory, units) is provided. Finally, the items for each grid point

| No. | Key | Description | Units |
|---|---|---|---|
| 1 | XYZ | Nuclear positions in Cartesian coordinates | Å |
| 2 | INT | Nuclear positions in internal coordinates | Å; degree |
| 3 | HF-TZ | Total energy at HF/cc-pVTZ | Hartree |
| 4 | HF-QZ | Total energy at HF/cc-pVQZ | Hartree |
| 5 | MP2 | Total energy at MP2/cc-pVTZ | Hartree |
| 6 | CCSD-T | Total energy at CCSD(T)/cc-pVQZ | Hartree |
| 7 | TBE | Theoretical best estimate of ab initio deformation energies | $cm^{-1}$ |
| 8 | MP2_grad_xyz | Energy gradient in Cartesian coordinates at MP2/cc-pVTZ | Hartree/Å |
| 9 | MP2_grad_int | Energy gradient in internal coordinates at MP2/cc-pVTZ | Hartree/Å; Hartree/degree |
| 10 | CCSD-T_grad_xyz | Energy gradient in Cartesian coordinates at CCSD(T)/cc-pVQZ | Hartree/Å |
| 11 | CCSD-T_grad_int | Energy gradient in internal coordinates at CCSD(T)/cc-pVQZ | Hartree/Å; Hartree/degree |
| 12 | CBS | Deformation energies at CCSD(T)-F12b/CBS | $cm^{-1}$ |
| 13 | VTZ | Deformation energies at CCSD(T)-F12b/cc-pVTZ-F12 (only for $CH_3Cl$ molecule) | $cm^{-1}$ |
| 14 | VQZ | Deformation energies at CCSD(T)-F12b/cc-pVQZ-F12 (only for $CH_3Cl$ molecule) | $cm^{-1}$ |
| 15 | CV | Deformation energy corrections to account for core-valence electron correlation | $cm^{-1}$ |
| 16 | HO | Deformation higher-order coupled cluster terms beyond perturbative triples | $cm^{-1}$ |
| 17 | SR | Deformation scalar relativistic (SR) effects | $cm^{-1}$ |
| 18 | DBOC | Deformation diagonal Born–Oppenheimer corrections (only for $CH_3Cl$, $CH_4$, $CH_3F$, and NaOH molecules) | $cm^{-1}$ |
| 19 | MP2-aQZ | Deformation energies at MP2/aug-cc-pVQZ (only for $CH_3Cl$ molecule) | $cm^{-1}$ |

**Table 3.** Layout of the VIB5.json file containing the VIB5 database.

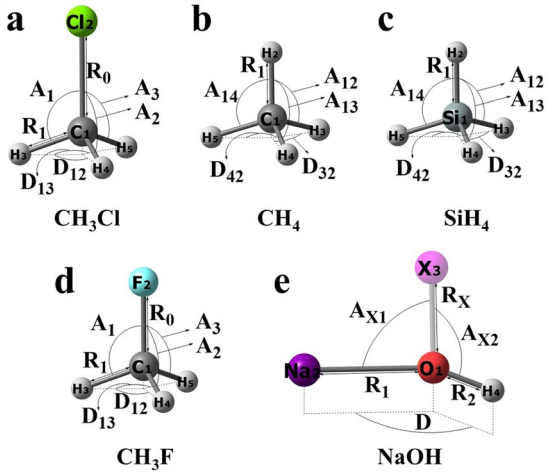


**Fig. 3** Definition of internal coordinates for each molecule used in the database file VIB5.json and in the complimentary calculations. Internal coordinates of (**a**) $CH_3Cl$; $R_0$ is C–Cl bond length, $R_i$ and $A_i$ are C–$H_{i+2}$ bond lengths and ∠($H_{i+2}$CCl) angles ($i = 1, 2, 3$), $D_{jk}$ are $H_{j+2}$CCl$H_{k+2}$ dihedral angles ($jk = 12, 13$); only $R_0$, $R_1$, $A_1$, $A_2$, $A_3$, $D_{12}$, and $D_{13}$ are shown; (**b**) $CH_4$; $R_i$ and $A_{1j}$ are C–$H_{i+1}$ bond lengths and ∠($H_2$C$H_{j+1}$) angles ($i = 1, 2, 3, 4$; $j = 2, 3, 4$), $D_{k2}$ are $H_{k+1}CH_2H_3$ dihedral angles ($k = 3, 4$); only $R_1$, $A_{12}$, $A_{13}$, $A_{14}$, $D_{32}$, and $D_{42}$ are shown; (**c**) $SiH_4$; $R_i$ and $A_{1j}$ are Si–$H_{i+1}$ bond lengths and ∠($H_2$Si$H_{j+1}$) angles ($i = 1, 2, 3, 4$; $j = 2, 3, 4$), $D_{k2}$ are $H_{k+1}SiH_2H_3$ dihedral angles ($k = 3, 4$); only $R_1$, $A_{12}$, $A_{13}$, $A_{14}$, $D_{32}$, and $D_{42}$ are shown; (**d**) $CH_3F$; $R_0$ is C–F bond length, $R_i$ and $A_i$ are C–$H_{i+2}$ bond lengths and ∠($H_{i+2}$CF) angles ($i = 1, 2, 3$), $D_{jk}$ are $H_{j+2}CFH_{k+2}$ dihedral angles ($jk = 12, 13$); only $R_0$, $R_1$, $A_1$, $A_2$, $A_3$, $D_{12}$, and $D_{13}$ are shown; (**e**) NaOH; $R_1$ and $R_2$ are Na–O and H–O bond lengths, $R_X$ is O–X bond length, $A_{X1}$ and $A_{X2}$ are ∠(XONa) and ∠(XOH) angles, and D is NaXOH dihedral angle. X is a dummy atom.

are given containing nuclear positions in both Cartesian and internal coordinates, and the values of properties (energies and energy gradients at different levels of theory, i.e., TBE, TBE constituent terms, complementary data). The JSON keys of items available for each grid point are listed in Table 3 with the brief description and units. The geometry configuration in Cartesian coordinates and in internal coordinates of each grid point for each molecule can be accessed by the "XYZ" key and the "INT" key, respectively. Definition of internal coordinates used in the database is shown in Fig. 3. The "HF-TZ", "HF-QC", "MP2", "CCSD-T", and "TBE" keys can be selected separately to obtain the energy of each grid point at HF/cc-pVTZ, HF/cc-pVQZ, MP2/cc-pVTZ,

CCSD(T)/cc-pVQZ, and TBE, respectively. This database also provides the energy gradients in Cartesian coordinates and internal coordinates at MP2/cc-pVTZ and CCSD(T)/cc-pVQZ theory levels, which can be accessed through "MP2_grad_xyz", "MP2_grad_int", "CCSD-T_grad_xyz", and "CCSD-T_grad_int" keys. See Table 3 for the summary and the keys of other properties.

## Technical Validation

The TBE values and TBE constituent terms were validated by calculating rovibrational spectra and comparing them to experiment in the original peer-reviewed publications cited in the *Methods* section and Table 1. In brief, rovibrational energy levels were computed by fitting analytical expression for PES and performing with it variational calculations using the nuclear motion program TROVE[80]. Then the resulting line list of rovibrational energy levels was compared to experimental values (when available) to validate the accuracy of the underlying PES. The new complementary data we have calculated here was validated by making sure that all calculations fully converged. After the database was constructed, we performed additional checks for repeated geometries, which identified grid points with the same geometrical parameters in the $CH_4$ grid points. We removed such duplicates from the database, which leads to a slightly reduced number of points (97217) compared to the numbers reported in the original publications (97271). This pruned grid is used as our final database.

## Usage Notes

We provide a Python script extraction_data.py that can be used to pull the data of interest from the VIB5.json (Box 1). It is provided together with the database file from https://doi.org/10.6084/m9.figshare.16903288[79].

**Box 1** Using extraction_data.py script to extract required data: an example of extracting CCSD(T)/CBS and CCSD(T)/cc-pVQZ energies and Cartesian geometries for NaOH. The *.dat files contain energies and *.xyz files contain XYZ geometries in the same order as in the database. The user can run `python3 extraction_data.py -h` command to see more options.

```
example$ ls
VIB5.json extraction_data.py
example$ python3 ./extraction_data.py --mols NaOH --energy CBS,CCSD-T -xyz
example$ ls
NaOH_CBS.dat NaOH_CCSD-T.dat NaOH.xyz VIB5.json extraction_data.py
example$ head -n 10 *.dat
==> NaOH_CBS.dat <==
59.280650000000
59.574700000000
59.558345000000
47.465761000000
64.042693000000
59.852814000000
60.391809000000
61.782135000000
33.479406000000
83.271969000000

==> NaOH_CCSD-T.dat <==
-237.644636975222
-237.644635947086
-237.644635792937
-237.644692089151
-237.644614779690
-237.644634762797
-237.644632245341
-237.644626330200
-237.644757449209
-237.644525233060
example$ head -n 10 *.xyz
3

O       0.00000000      0.00000000      1.08916506
Na      0.00000000      0.00000000      -0.84719335
H       0.00000000      0.00000000      2.03971526
3

O       0.00000000      0.00000000      1.08917892
Na      0.00000000      0.00000000      -0.84717949
H       0.00000000      0.00000000      2.03917912
```

## Code availability

All the data generated at the MP2/cc-pVTZ and the CCSD(T)/cc-pVQZ levels of theory were performed with the CFOUR software package. TBE and other data were obtained using various software packages (MOLPRO, CFOUR, MRCC) as described in the Methods section.

Received: 2 November 2021; Accepted: 19 January 2022;
Published online: 11 March 2022

## References

1. Lewars, E. *Computational Chemistry: Introduction to the Theory and Applications of Molecular and Quantum Mechanics* 2nd edn (Springer Science+Business Media B.V., 2011).
2. Upadhyay, S. K. *Chemical Kinetics and Reaction Dynamics* (Anamaya Publishers, 2006).
3. Searles, D. J. & von Nagy-Felsobuki, E. I. In *Ab Initio Variational Calculations of Molecular Vibrational-Rotational Spectra* (Springer-Verlag Berlin Heidelberg, 1993).
4. Császár, A. G., Czakó, G., Furtenbacher, T. & Mátyus, E. In *Annual Reports in Computational Chemistry* **3** (Elsevier, 2007).
5. Bytautas, L., Bowman, J. M., Huang, X. & Varandas, A. J. C. Accurate potential energy surfaces and beyond: chemical reactivity, binding, long-range interactions, and spectroscopy. *Adv. Phys. Chem.* **2012**, 679869 (2012).
6. Tennyson, J. & Yurchenko, S. N. ExoMol: molecular line lists for exoplanet and other atmospheres. *Mon. Not. R. Astron. Soc.* **425**, 21–33 (2012).
7. Owens, A., Yurchenko, S. N., Yachmenev, A., Tennyson, J. & Thiel, W. Accurate ab initio vibrational energies of methyl chloride. *J. Chem. Phys.* **142** (2015).
8. Owens, A., Yurchenko, S. N., Yachmenev, A. & Thiel, W. A global potential energy surface and dipole moment surface for silane. *J. Chem. Phys.* **143** (2015).
9. Owens, A., Yurchenko, S. N., Yachmenev, A., Tennyson, J. & Thiel, W. A global ab initio dipole moment surface for methyl chloride. *J. Quant. Spectrosc. Radiat. Transfer* **184**, 100–110 (2016).
10. Owens, A., Yurchenko, S. N., Yachmenev, A., Tennyson, J. & Thiel, W. A highly accurate ab initio potential energy surface for methane. *J. Chem. Phys.* **145** (2016).
11. Owens, A. & Yurchenko, S. N. Theoretical rotation-vibration spectroscopy of cis- and trans-diphosphene ($P_2H_2$) and the deuterated species $P_2HD$. *J. Chem. Phys.* **150** (2019).
12. Owens, A., Yachmenev, A., Kupper, J., Yurchenko, S. N. & Thiel, W. The rotation-vibration spectrum of methyl fluoride from first principles. *Phys. Chem. Chem. Phys.* **21**, 3496–3505 (2019).
13. Owens, A., Conway, E. K., Tennyson, J. & Yurchenko, S. N. ExoMol line lists – XXXVIII. High-temperature molecular line list of silicon dioxide (SiO2). *Mon. Not. R. Astron. Soc.* **495**, 1927–1933 (2020).
14. Owens, A., Tennyson, J. & Yurchenko, S. N. ExoMol line lists – XLI. High-temperature molecular line lists for the alkali metal hydroxides KOH and NaOH. *Mon. Not. R. Astron. Soc.* **502**, 1128–1135 (2021).
15. Tennyson, J. *et al*. ExoMol molecular line lists XXX: a complete high-accuracy line list for water. *Mon. Not. R. Astron. Soc.* **480**, 2597–2608 (2018).
16. Yurchenko, S. N. & Tennyson, J. ExoMol line lists - IV. The rotation-vibration spectrum of methane up to 1500 K. *Mon. Not. R. Astron. Soc.* **440**, 1649–1661 (2014).
17. Raghavachari, K., Trucks, G. W., Pople, J. A. & Head-Gordon, M. A fifth-order perturbation comparison of electron correlation theories. *Chem. Phys. Lett.* **157**, 479–483 (1989).
18. Helgaker, T., Gauss, J., Jørgensen, P. & Olsen, J. The prediction of molecular equilibrium structures by the standard electronic wave functions. *J. Chem. Phys.* **106**, 6430–6440 (1997).
19. Bak, K. L. *et al*. The accurate determination of molecular equilibrium structures. *J. Chem. Phys.* **114**, 6548–6556 (2001).
20. Dral, P. O., Owens, A., Dral, A. & Csányi, G. Hierarchical machine learning of potential energy surfaces. *J. Chem. Phys.* **152**, 204110 (2020).
21. Behler, J. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Phys. Chem. Chem. Phys.* **13**, 17930–17955 (2011).
22. Manzhos, S., Dawes, R. & Carrington, T. Jr. Neural network-based approaches for building high dimensional and quantum dynamics-friendly potential energy surfaces. *Int. J. Quantum Chem.* **115**, 1012–1020 (2015).
23. Unke, O. T. *et al*. Machine learning force fields. *Chem. Rev.* **121**, 10142–10186 (2021).
24. Manzhos, S. & Carrington, T. Jr. Neural network potential energy surfaces for small molecules and reactions. *Chem. Rev.* **121**, 10187–10217 (2020).
25. Mueller, T., Hernandez, A. & Wang, C. Machine learning for interatomic potential models. *J. Chem. Phys.* **152**, 050902 (2020).
26. Dral, P. O. In *Advances in Quantum Chemistry: Chemical Physics and Quantum Chemistry* **81** (Academic Press, 2020).
27. Dral, P. O. Quantum chemistry in the age of machine learning. *J. Phys. Chem. Lett.* **11**, 2336–2347 (2020).
28. Schmitz, G., Artiukhin, D. G. & Christiansen, O. Approximate high mode coupling potentials using Gaussian process regression and adaptive density guided sampling. *J. Chem. Phys.* **150**, 131102 (2019).
29. Gastegger, M., Behler, J. & Marquetand, P. Machine learning molecular dynamics for the simulation of infrared spectra. *Chem. Sci.* **8**, 6924–6935 (2017).
30. Kamath, A., Vargas-Hernández, R. A., Krems, R. V., Carrington, T. Jr. & Manzhos, S. Neural networks vs Gaussian process regression for representing potential energy surfaces: A comparative study of fit quality and vibrational spectrum accuracy. *J. Chem. Phys.* **148**, 241702 (2018).
31. Manzhos, S. Machine learning for the solution of the Schrödinger equation. *Mach. Learn.: Sci. Technol.* **1**, 013002 (2020).
32. Manzhos, S., Yamashita, K. & Carrington, T. Jr. Using a neural network based method to solve the vibrational Schrodinger equation for $H_2O$. *Chem. Phys. Lett.* **474**, 217–221 (2009).
33. Manzhos, S., Wang, X. G., Dawes, R. & Carrington, T. Jr. A nested molecule-independent neural network approach for high-quality potential fits. *J Phys Chem A* **110**, 5295–5304 (2006).
34. Manzhos, S. & Carrington, T. Jr. A random-sampling high dimensional model representation neural network for building potential energy surfaces. *J. Chem. Phys.* **125**, 084109 (2006).
35. Manzhos, S. & Carrington, T. Jr. Using neural networks, optimized coordinates, and high-dimensional model representations to obtain a vinyl bromide potential surface. *J. Chem. Phys.* **129**, 224104 (2008).
36. Dral, P. O., Owens, A., Yurchenko, S. N. & Thiel, W. Structure-based sampling and self-correcting machine learning for accurate calculations of potential energy surfaces and vibrational levels. *J. Chem. Phys.* **146**, 244108 (2017).
37. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Big data meets quantum chemistry approximations: the Δ-machine learning approach. *J. Chem. Theory Comput.* **11**, 2087–2096 (2015).
38. Chmiela, S. *et al*. Machine learning of accurate energy-conserving molecular force fields. *Sci. Adv.* **3**, e1603015 (2017).
39. Chmiela, S., Sauceda, H. E., Müller, K.-R. & Tkatchenko, A. Towards exact molecular dynamics simulations with machine-learned force fields. *Nat. Commun.* **9**, 3887 (2018).

40. Rupp, M., Tkatchenko, A., Müller, K.-R. & von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Phys. Rev. Lett.* **108**, 058301 (2012).
41. Montavon, G. *et al*. Machine learning of molecular electronic properties in chemical compound space. *New J. Phys.* **15**, 095003 (2013).
42. Ramakrishnan, R., Dral, P. O., Rupp, M. & von Lilienfeld, O. A. Quantum chemistry structures and properties of 134 kilo molecules. *Sci. Data* **1**, 140022 (2014).
43. Kim, H., Park, J. Y. & Choi, S. Energy refinement and analysis of structures in the QM9 database via a highly accurate quantum chemical method. *Sci. Data* **6**, 109 (2019).
44. Smith, J. S. *et al*. The ANI-1ccx and ANI-1x data sets, coupled-cluster and density functional theory properties for molecules. *Sci. Data* **7**, 134 (2020).
45. Smith, J. S., Isayev, O. & Roitberg, A. E. ANI-1, a data set of 20 million calculated off-equilibrium conformations for organic molecules. *Sci. Data* **4**, 170193 (2017).
46. Hoja, J. *et al*. QM7-X, a comprehensive dataset of quantum-mechanical properties spanning the chemical space of small organic molecules. *Sci. Data* **8**, 43 (2021).
47. Qu, C., Houston, P. L., Conte, R., Nandi, A. & Bowman, J. M. MULTIMODE calculations of vibrational spectroscopy and 1d interconformer tunneling dynamics in Glycine using a full-dimensional potential energy surface. *J Phys Chem A* **125**, 5346–5354 (2021).
48. Hill, J. G., Peterson, K. A., Knizia, G. & Werner, H.-J. Extrapolating MP2 and CCSD explicitly correlated correlation energies to the complete basis set limit with first and second row correlation consistent basis sets. *J. Chem. Phys.* **131**, 194105 (2009).
49. Adler, T. B., Knizia, G. & Werner, H.-J. A simple and efficient CCSD(T)-F12 approximation. *J. Chem. Phys.* **127**, 221106 (2007).
50. Peterson, K. A., Adler, T. B. & Werner, H.-J. Systematically convergent basis sets for explicitly correlated wavefunctions: the atoms H, He, B–Ne, and Al–Ar. *J. Chem. Phys.* **128**, 084102 (2008).
51. Ten-no, S. Initiation of explicitly correlated Slater-type geminal theory. *Chem. Phys. Lett.* **398**, 56–61 (2004).
52. Yousaf, K. E. & Peterson, K. A. Optimized auxiliary basis sets for explicitly correlated methods. *J. Chem. Phys.* **129**, 184108 (2008).
53. Weigend, F. A fully direct RI-HF algorithm: implementation, optimised auxiliary basis sets, demonstration of accuracy and efficiency. *Phys. Chem. Chem. Phys.* **4**, 4285–4291 (2002).
54. Hättig, C. Optimization of auxiliary basis sets for RI-MP2 and RI-CC2 calculations: Core-valence and quintuple-ζ basis sets for H to Ar and QZVPP basis sets for Li to Kr. *Phys. Chem. Chem. Phys.* **7**, 59–66 (2005).
55. Werner, H.-J., Knowles, P. J., Knizia, G., Manby, F. R. & Schütz, M. Molpro: a general-purpose quantum chemistry program package. *WIREs Comput. Mol. Sci.* **2**, 242–253 (2012).
56. Werner, H.-J. *et al*. The Molpro quantum chemistry package. *J. Chem. Phys.* **152**, 144107 (2020).
57. Hill, J. G., Mazumder, S. & Peterson, K. A. Correlation consistent basis sets for molecular core-valence effects with explicitly correlated wave functions: the atoms B–Ne and Al–Ar. *J. Chem. Phys.* **132**, 054108 (2010).
58. Kállay, M. & Gauss, J. Approximate treatment of higher excitations in coupled-cluster theory. *J. Chem. Phys.* **123**, 214105 (2005).
59. Kállay, M. & Gauss, J. Approximate treatment of higher excitations in coupled-cluster theory. II. Extension to general single-determinant reference functions and improved approaches for the canonical Hartree–Fock case. *J. Chem. Phys.* **129**, 144101 (2008).
60. MRCC, A string-based quantum chemical program suite written by M. Kállay; see also M. Kállay & P. R. Surján, *J. Chem. Phys.* **115**, 2945 (2001).
61. Stanton, J. F. et al. CFOUR, a quantum chemical program package http://www.cfour.de (2010).
62. Dunning, T. H. Jr. Gaussian basis sets for use in correlated molecular calculations. I. The atoms boron through neon and hydrogen. *J. Chem. Phys.* **90**, 1007–1023 (1989).
63. Kendall, R. A., Dunning, T. H. Jr. & Harrison, R. J. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.* **96**, 6796–6806 (1992).
64. Woon, D. E. & Dunning, T. H. Jr. Gaussian basis sets for use in correlated molecular calculations. III. The atoms aluminum through argon. *J. Chem. Phys.* **98**, 1358–1371 (1993).
65. Dunning, T. H. Jr., Peterson, K. A. & Wilson, A. K. Gaussian basis sets for use in correlated molecular calculations. X. The atoms aluminum through argon revisited. *J. Chem. Phys.* **114**, 9244–9253 (2001).
66. Prascher, B. P., Woon, D. E., Peterson, K. A., Dunning, T. H. Jr. & Wilson, A. K. Gaussian basis sets for use in correlated molecular calculations. VII. Valence, core-valence, and scalar relativistic basis sets for Li, Be, Na, and Mg. *Theor. Chem. Acc.* **128**, 69–82 (2011).
67. Cowan, R. D. & Griffin, D. C. Approximate relativistic corrections to atomic radial wave functions*. *J. Opt. Soc. Am.* **66**, 1010–1014 (1976).
68. Douglas, M. & Kroll, N. M. Quantum electrodynamical corrections to the fine structure of helium. *Ann. Phys.* **82**, 89–155 (1974).
69. Hess, B. A. Relativistic electronic-structure calculations employing a two-component no-pair formalism with external-field projection operators. *Phys. Rev. A* **33**, 3742–3748 (1986).
70. Woon, D. E. & Dunning, T. H. Jr. Gaussian basis sets for use in correlated molecular calculations. V. Core-valence basis sets for boron through neon. *J. Chem. Phys.* **103**, 4572–4585 (1995).
71. Peterson, K. A. & Dunning, T. H. Jr. Accurate correlation consistent basis sets for molecular core-valence correlation effects: The second row atoms Al–Ar, and the first row atoms B–Ne revisited. *J. Chem. Phys.* **117**, 10548–10560 (2002).
72. Klopper, W. Simple recipe for implementing computation of first-order relativistic corrections to electron correlation energies in framework of direct perturbation theory. *J. Comput. Chem.* **18**, 20–27 (1997).
73. Jong, W. A. D., Harrison, R. J. & Dixon, D. A. Parallel Douglas–Kroll energy and gradients in NWChem: estimating scalar relativistic effects using Douglas–Kroll contracted basis sets. *J. Chem. Phys.* **114**, 48–53 (2001).
74. Gauss, J., Tajti, A., Kállay, M., Stanton, J. F. & Szalay, P. G. Analytic calculation of the diagonal Born-Oppenheimer correction within configuration-interaction and coupled-cluster theory. *J. Chem. Phys.* **125**, 144111 (2006).
75. Bartlett, R. J. Many-body perturbation theory and coupled cluster theory for electron correlation in molecules. *Annu. Rev. Phys. Chem.* **32**, 359–401 (1981).
76. Cremer, D. in *Encyclopedia of Computational Chemistry* (John Wiley and Sons, Ltd., 1998).
77. Bartlett, R. J., Watts, J. D., Kucharski, S. A. & Noga, J. Non-iterative fifth-order triple and quadruple excitation energy corrections in correlated methods. *Chem. Phys. Lett.* **165**, 513–522 (1990).
78. Stanton, J. F. Why CCSD(T) works: a different perspective. *Chem. Phys. Lett.* **281**, 130–134 (1997).
79. Zhang, L., Zhang, S., Owens, A., Yurchenko, S. N. & Dral, P. O. VIB5 database with accurate ab initio quantum chemical molecular potential energy surfaces. *figshare* https://doi.org/10.6084/m9.figshare.16903288 (2021).
80. Yurchenko, S. N., Thiel, W. & Jensen, P. Theoretical ROVibrational Energies (TROVE): a robust numerical approach to the calculation of rovibrational energies for polyatomic molecules. *J. Mol. Spectrosc.* **245**, 126–140 (2007).


## Acknowledgements

POD acknowledges funding by the National Natural Science Foundation of China (No. 22003051), the Fundamental Research Funds for the Central Universities (No. 20720210092), and via the Lab project of the State Key Laboratory of Physical Chemistry of Solid Surfaces. SNY and AO thank STFC under grant ST/R000476/1. Their calculations made extensive use of the STFC DiRAC HPC facility supported by BIS National E-infrastructure capital grant ST/J005673/1 and STFC grants ST/H008586/1 and ST/K00333X/1.

## Author contributions

L.Z. has written the original draft of the manuscript. S.Z. performed the complementary calculations, validation, created scripts and database files with assistance of L.Z. and P.O.D. A.O. provided raw data with grids, theoretical best estimates and energy correction terms as well as supporting scripts. A.O., S.N.Y. and P.O.D. supervised the project. S.N.Y. and P.O.D. acquired funding for the project. All authors provided critical feedback and helped shape the database collection, calculations, analysis, and manuscript. P.O.D. conceived the idea of creating a database.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to A.O. or P.O.D.

**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.