Supplementary Information (SI) for Physical Chemistry Chemical Physics.
This journal is © the Owner Societies 2024

Supporting information to: Machine-learning to predict anharmonic frequencies: a study of models and transferability

Basis set information of the ML dataset

Table S1: Clusters used for ML dataset and corresponding basis set information (-r = ring; -c = chain; -l = linear; -rb = boat ring).

Cluster	basis set (def2-)	Cluster	basis set (def2-)
HF	TZVP	$(HBr)_2$	TZVP
$(HF)_2$	TZVP	$(HBr)_3$-r	SVPD
$(HF)_3$-r	TZVP	$(HBr)_3$-c	SVPD
$(HF)_3$-c[a]	SVPD	$(HBr)_4$-r	SVP
$(HF)_3$-cl[b]	TZVP	CH_3F	TZVP
$(HF)_4$-r	TZVP	$(CH_3F)_2$	SVPD
$(HF)_4$-rb	TZVP	$(CH_3F)_2$	SVP
$(HF)_5$-r	TZVP	CH_3Cl	TZVP
HCl	TZVP	$(CH_3Cl)_2$	SVP
$(HCl)_2$	TZVP	CH_3Br	TZVP
$(HCl)_3$-r	TZVP	$(CH_3Br)_2$	TZVP
$(HCl)_3$-c	SVPD	C_2H_5F	TZVP
$(HCl)_4$-r	SVPD	C_2H_5Cl	TZVP
$(HCl)_4$-rb	SVPD	C_2H_5Br	TZVP
HBr	TZVP		

S1 Quantitative analysis of the ML model

Convergence of the GBR model

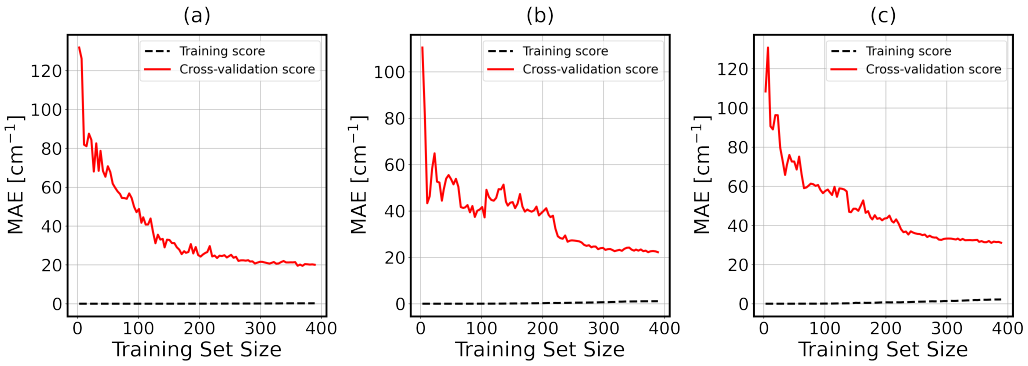

Figure S1: Learning curves of the gradient boosting regression leave-one-out cross validation for (a) Diagonal, (b) VSCF and (c) VSCF-PT2 frequencies

Convergence of the MLR model

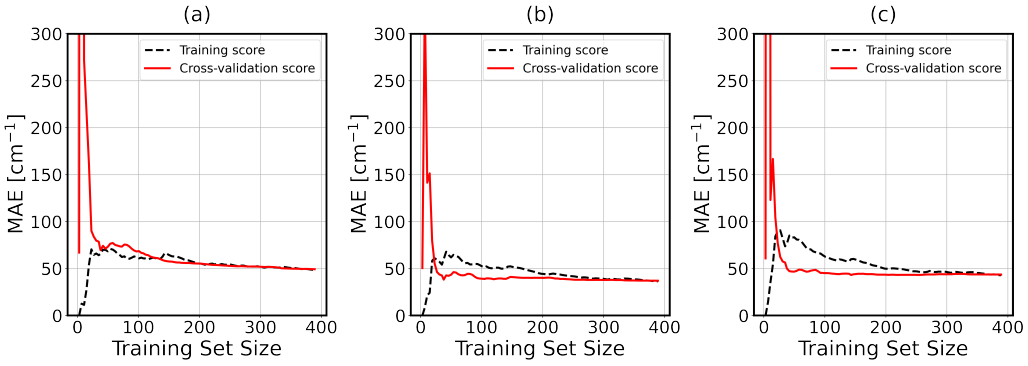

Figure S2: Learning curves of the linear regression leave-one-out cross validation for (a) Diagonal, (b) VSCF and (c) VSCF-PT2 frequencies

MLR predictions

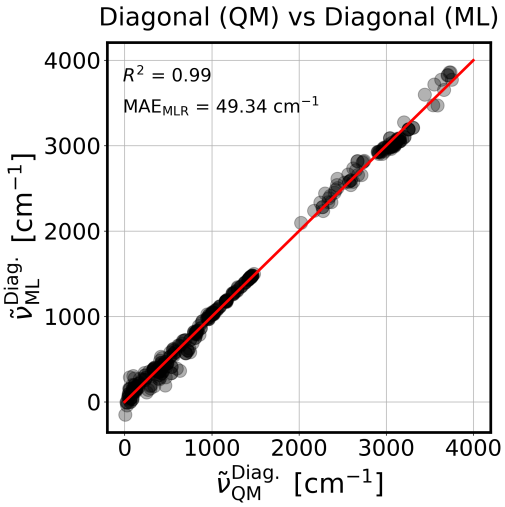

Figure S3: QM calculated Diagonal frequencies ($\tilde{\nu}^{\text{Diag}}_{\text{QM}}$) vs. ML predicted Diagonal frequencies ($\tilde{\nu}^{\text{Diag}}_{\text{ML}}$). $\tilde{\nu}^{\text{Diag}}_{\text{ML}}$ are LOOCV linear regression predictions. Red line is ideal diagonal

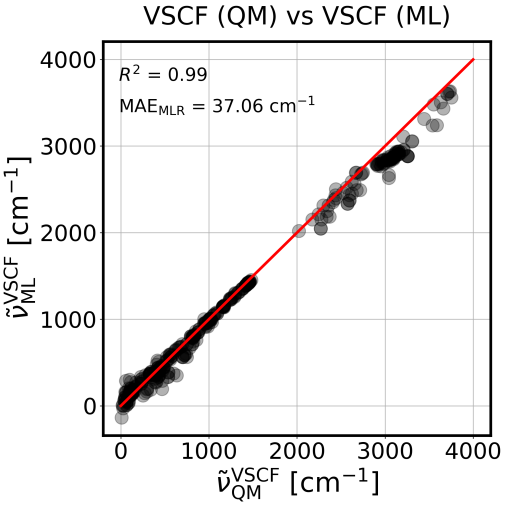

Figure S4: Same as Fig. S3 for VSCF frequencies. $\tilde{\nu}^{\text{VSCF}}_{\text{ML}}$ are LOOCV linear regression predictions. Red line is ideal diagonal

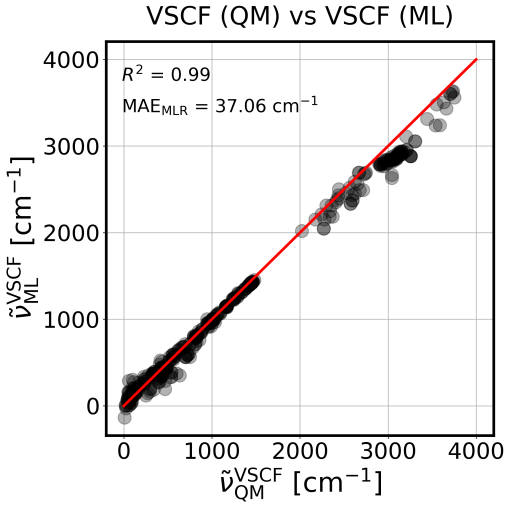

Figure S5: Same as Fig S3 for VSCF-PT2 frequencies. $\tilde{\nu}_{\mathrm{ML}}^{\mathrm{VSCF-PT2}}$ are LOOCV linear regression predictions. Red line is ideal diagonal

S2 Comparison of VSCF-calculated, AIMD and experimental[a] spectra of CH_3F

AIMD computational details

Velocity-Verlet algorithm as implemented in NWChem was used for AIMD simulations. Electronic potential was calculated using density functional theory with def2-TZVP basis set and BLYP exchange correlation functional. Grimme's dispersion correction with Becke-Johnson damping D3(BJ) were employed to account for Van der Waals interactions. The nuclear time step was set to 0.2419 fs and in total 5000 steps were simulated. NVE ensemble, i.e, no thermostat with the temperature of 298.15 K, to generate initial velocities from Maxwell-Boltzmann distribution, were utilized. Corresponding IR spectrum was generated using NWChem postprocessing tool `qmd_analysis`.

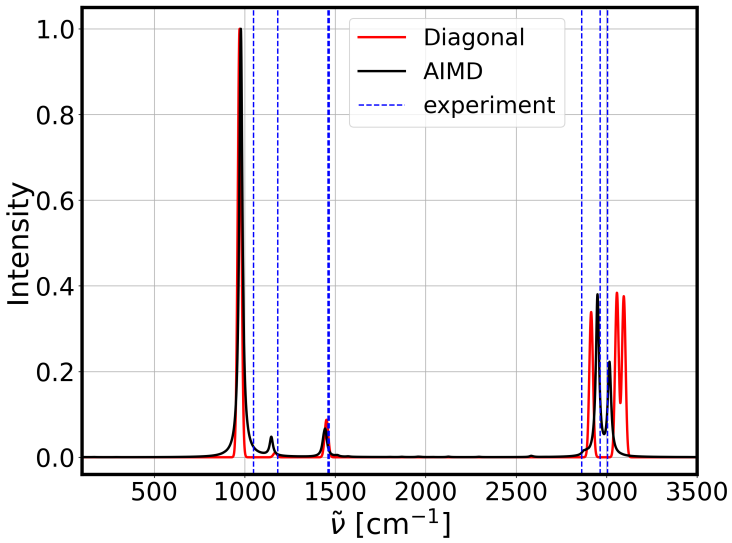

Figure S6: Diagonal and AIMD spectra of CH_3F; blue dashed vertical lines are experimental intensity peak locations

[a]J. Chem. Phys. 76, 809–816 (1982); J. Chem. Phys. 66, 970–975 (1977)

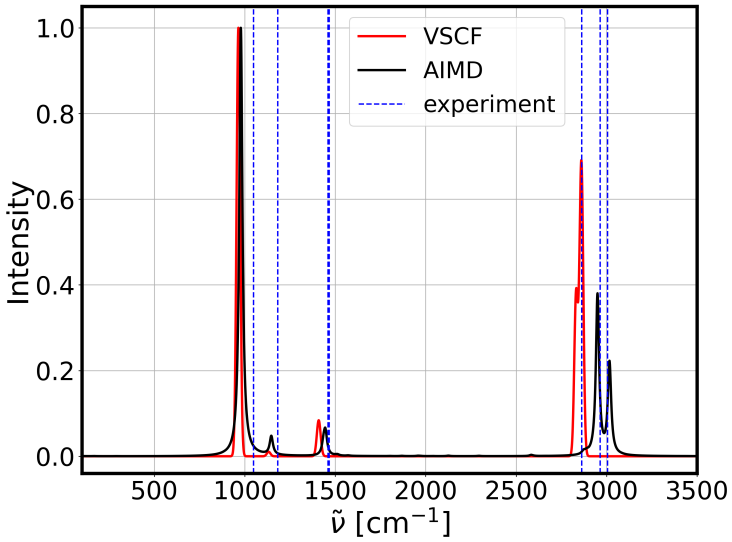

Figure S7: VSCF and AIMD spectra of CH_3F; blue dashed vertical lines are experimental intensity peak locations

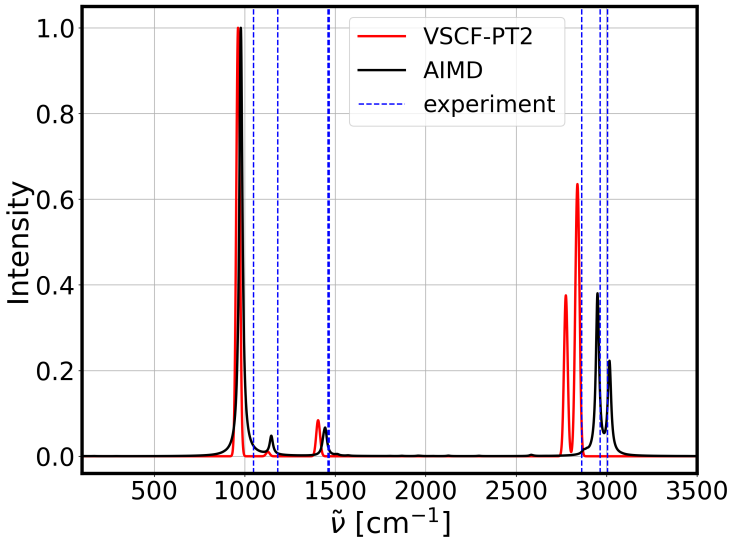

Figure S8: VSCF-PT2 and AIMD spectra of CH_3F; blue dashed vertical lines are experimental intensity peak locations

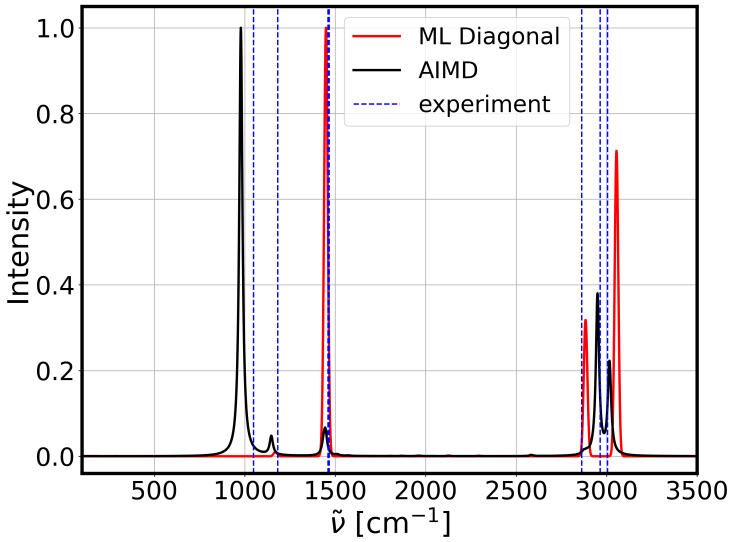

Figure S9: ML predicted Diagonal and AIMD spectra of CH_3F; blue dashed vertical lines are experimental intensity peak locations

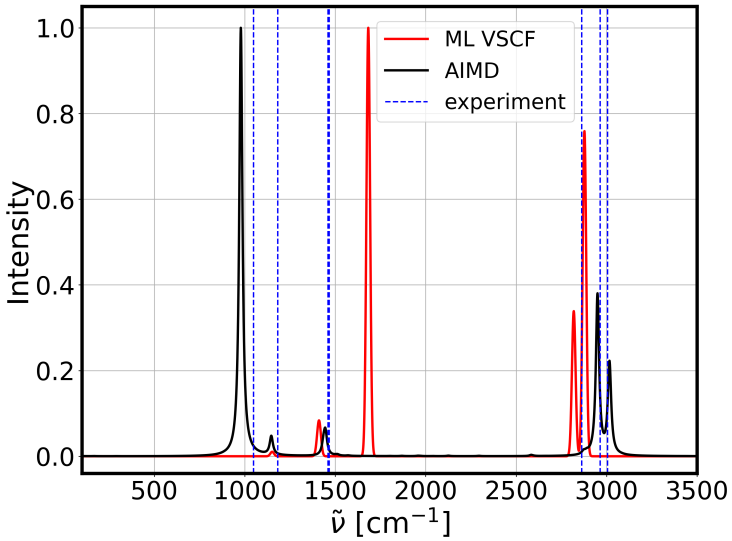

Figure S10: ML predicted VSCF and AIMD spectra of CH_3F; blue dashed vertical lines are experimental intensity peak locations

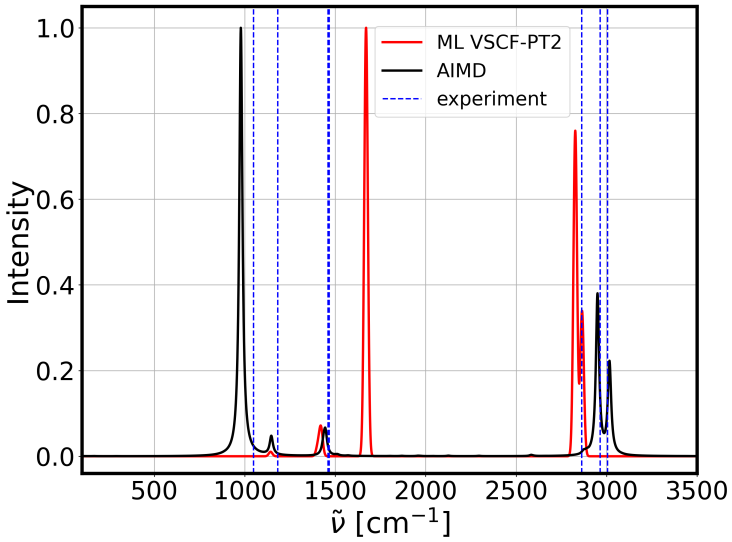

Figure S11: ML predicted VSCF-PT2 and AIMD spectra of CH_3F; blue dashed vertical lines are experimental intensity peak locations

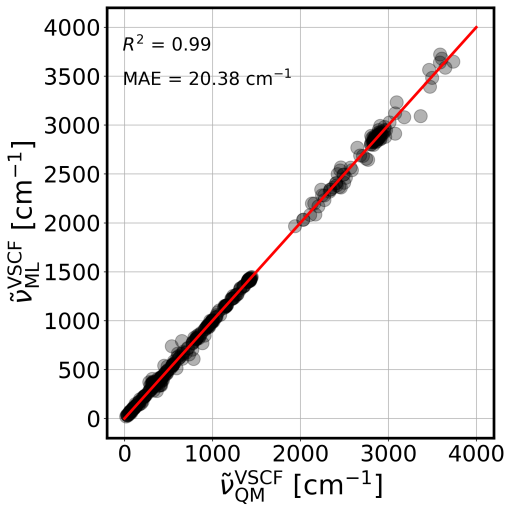

Figure S12: QM calculated vs. ML predicted VSCF frequencies. Here "Diagonal - VSCF" shift is used as target and Diagonal frequencies are included to the original harmonic based descriptor set.

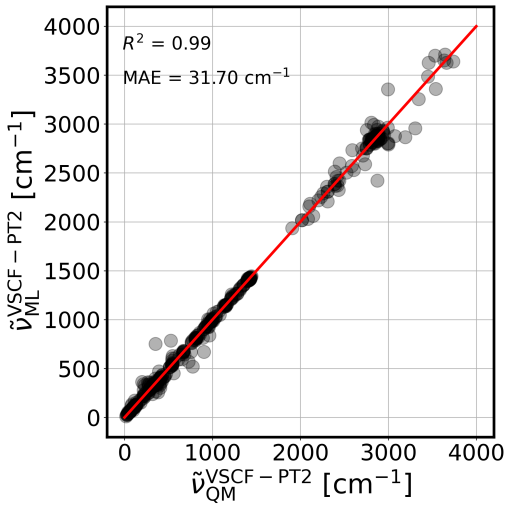

Figure S13: QM calculated vs. ML predicted VSCF-PT2 frequencies. Here "Diagonal - VSCF-PT2" shift is used as target and Diagonal frequencies are included to the original harmonic based descriptor set.

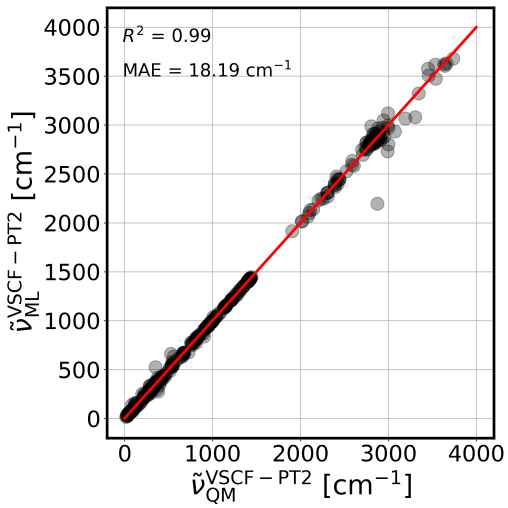

Figure S14: QM calculated vs. ML predicted VSCF-PT2 frequencies. Here "VSCF - VSCF-PT2" shift is used as target and VSCF frequencies are included to the original harmonic based descriptor set.

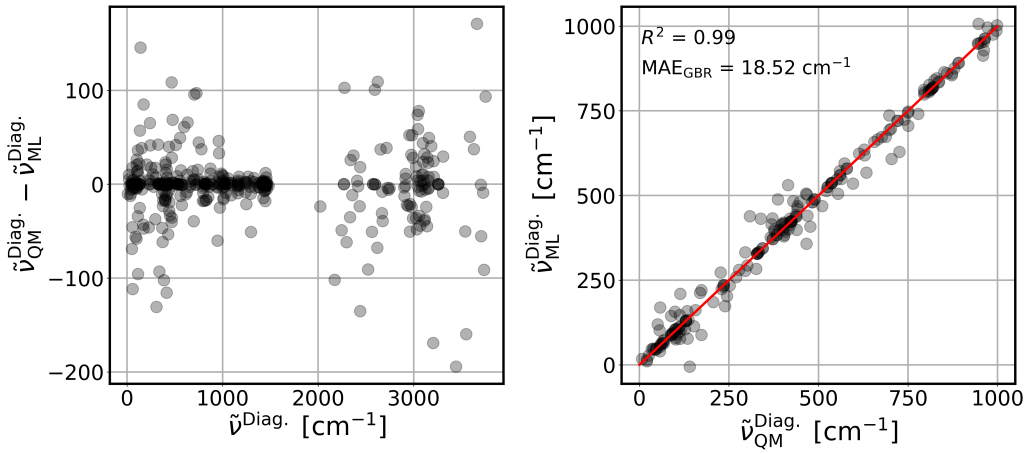

Figure S15: Left panel: difference between ML predicted and QM calculated Diagonal frequencies vs. Diagonal frequencies; Right panel: predictions of the Diagonal frequencies below 1000 cm^{-1}

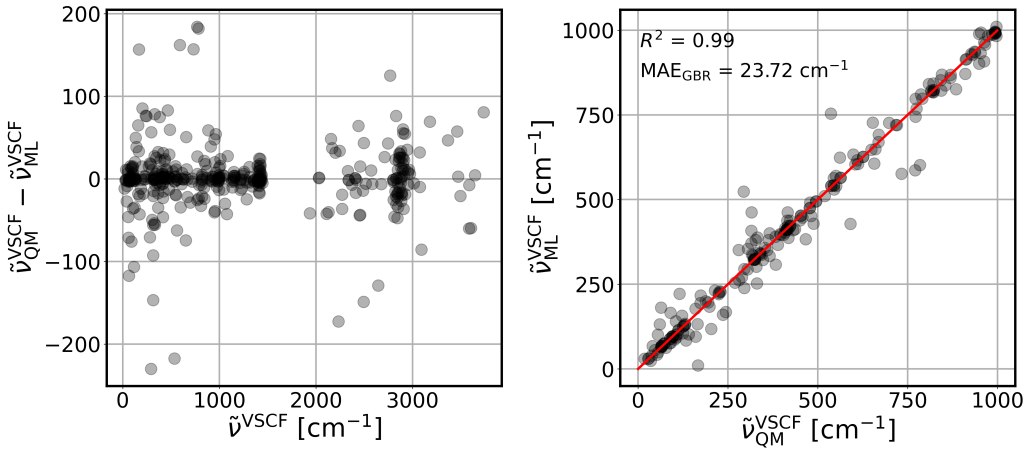

Figure S16: Left panel: difference between ML predicted and QM calculated VSCF frequencies vs. VSCF frequencies; Right panel: predictions of the VSCF frequencies below 1000 cm^{-1}

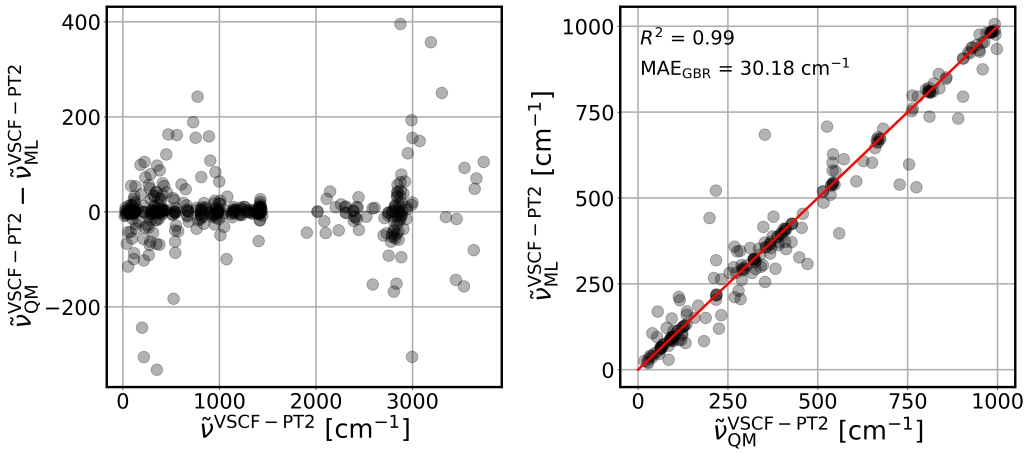

Figure S17: Left panel: difference between ML predicted and QM calculated VSCF-PT2 frequencies vs. VSCF-PT2 frequencies; Right panel: predictions of the VSCF-PT2 frequencies below 1000 cm^{-1}

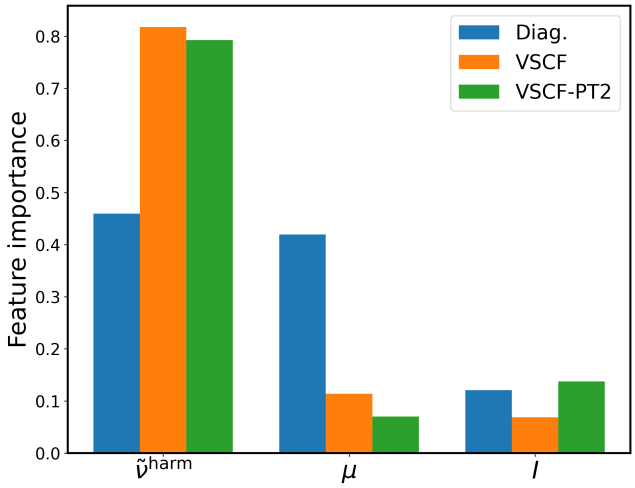

Figure S18: Feature importance of the descriptor set for 3 anharmonic frequency types. Here $\tilde{\nu}^{\mathrm{harm}}$ - harmonic frequency, μ - reduced mass and I - intensity