Electronic Supplementary Material (ESI) for Physical Chemistry Chemical Physics.
This journal is © the Owner Societies 2023

Supplementary Information – HyDRA challenge

Silvan Käser [ORCID: 0000-0002-3641-8519, email: silvan.kaeser@unibas.ch][1], Kai Töpfer [ORCID: 0000-0002-4650-9641, email: kai.toepfer@unibas.ch][1], Luis I. Vazquez-Salazar [ORCID: 0000-0001-6347-5108, email: luisitza.vazquezsalazar@unibas.ch][1], Eric D. Boittier [ORCID: 0000-0002-9611-1017, email: ericdavid.boittier@unibas.ch][1], and Markus Meuwly [ORCID: 0000-0001-7930-8806, email: m.meuwly@unibas.ch][1]

[1]Department of Chemistry, University of Basel, Klingelbergstrasse 80 , CH-4056 Basel, Switzerland.

1 Computational Details

Different approaches including DVR3D, kernel, and neural network (NN) based methods were explored for the study of the hydrogen-bonded OH stretching vibration $\omega_{\mathrm{OH_b}}$. This document summarizes the efforts from the Meuwly Research group of the University of Basel.

1.1 Data generation - Eric D. Boittier & Silvan Käser

The structures of the mono-hydrates (i.e. a molecule and hydrogen-bonded water) given in the HyDRA training and test sets were optimized followed by a frequency calculation at the B3LYP[1]/aug-cc-pVTZ[2] + D3[3] level of theory using Gaussian09[4]. Starting from the 20 molecules given in the HyDRA data set, a possible extension to the data set was considered which would include new, chemically similar molecules to the test and training data while adding unseen chemical information to supplement training data for machine learning models. To this end, using the GDB11 database[5, 6] of molecules between 5 and 10 heavy atoms, a filter that only included molecules with the hydrogen–bonding motifs present in the training and test set (i.e. -OH, -NH, etc.) was applied, and a similarity search was performed in the RDKit[7]. A total of approximately 200 molecules were extracted using this procedure. For each of these molecules, a water molecule was positioned at the anticipated hydrogen bonding site, before an optimization and frequency calculation as previously described was performed.

The Tanimoto measure was used as the similarity metric and was calculated as implemented in RDKit[7] and using the canonical SMILES of the host molecule. Identical structures (those with a similarity index equal to one) were ignored. To achieve a suitable guess for the water orientation (before optimization), two consecutive alignments using the Kabsch algorithm[8] are performed. First, between the selected hydrogen-bonding motif, and secondly, using the maximum common substructure[7] for the query and the molecule identified in GDB11. The coordinates for the guest water molecule, taken from the initial optimized complex, were appended to the end of the file.

1.2 Neural Network + Transfer Learning - Silvan Käser

This approach was based on a combination of neural network and transfer learning (NN + TL) strategies. It included obtaining a NN model capable of predicting the harmonic frequency $\omega_{\mathrm{OH_b}}$ followed by TL based on the experimental frequencies given in the HyDRA training set. The training of a base NN model for predicting the harmonic frequency $\omega_{\mathrm{OH_b}}$ was assumed to benefit from an extended data set (see Sec. 1.1) as ~ 20 molecules are likely insufficient to obtain a robust machine learning model. The power of such an approach is that it can be systematically improved (e.g. by adding further *ab initio* and experimental data, optimizing the descriptor used for training and the size of the NN, or choosing a more sophisticated NN architecture).

Neural Network: A base NN able to predict the harmonic frequency $\omega_{\mathrm{OH_b}}$ was trained on the extended data set (i.e. optimized geometries and its *ab initio* $\omega_{\mathrm{OH_b}}$. Note that the frequencies were standardized, i.e. $\omega \rightarrow (\omega - \mu)/\sigma$, before training). The descriptor/feature vector used in this work is based on the FCHL descriptor[9, 10], as implemented in the *QML* Python toolkit[11]. As is common in machine learning, a principal component analysis was performed to obtain a concise descriptor starting from the FCHL representation giving a descriptor of length 193.

A feed-forward neural network was then used, consisting of 5 hidden layers with [193, 193, 193, 193, 24] nodes each. After the final layer a linear transformation was used to obtain the final output. The shifted-softplus function was used as activation function and the parameters of the NN were optimized using the Adam optimizer[12]. The NN was written using Tensorflow's Keras module[13]. It is to note that the molecules from the HyDRA training and test set are all in the training set of the base NN. A generalization to other molecules can be advantageous but is not mandatory as, in principle, new target molecules (for which an estimate for the experimental frequency is needed) can be added.

Transfer learning: Starting from the original model, TL was carried out using the experimental frequencies given in the HyDRA challenge set, to account for the anharmonicities. Therefore, the parameters from the original model were loaded and the parameters from all but the last two layers were frozen. Then, in the actual TL step, the parameters of the last two layers were re-optimized based on 9 experimental points (Note that the molecule 2406-25-9 was not included in the training and evaluation because, contrary to the other molecules, it is a radical). TL was repeated multiple times on different splits of the experimental data (i.e. different molecules were used as validation set) and the average of the predictions was determined.

1.3 Reproducing Kernel Hilbert Space + DVR3D - Kai Töpfer

QM+Kernel approach: The internal potential of an adsorbed water molecule is most important in determining the eigenvalue of the symmetric stretch vibration. As such, a 3-dimensional potential energy surface (PES) was computed from quantum electronic methods, along waters degrees of freedom, and the vibrational eigenstates were obtained from the Discrete Variable Representation (DVR) method.[14] Using the optimized structures at the B3LYP/aug-cc-pVTZ + D3 level of theory as reference the internal degrees of freedom of the water molecule were sampled by 9 points along both OH bonds equilibrium value r_1^{eq}, r_2^{eq} ($\Delta r = \pm 0.21$ Å) and 7 points along the HOH equilibrium angle θ^{eq} ($\Delta\theta = \pm 24^\circ$). The center of mass and the axes of inertia of water kept fixed. In total, 567 data points were obtained at the B3LYP/aug-cc-pVTZ + D3 level of theory. A 3-dimensional PES V_{PES} was then constructed by the sum of 3 Morse potentials V_{Morse} along each internal water coordinate (r_1, r_2, z). Here, z is a transformation of the HOH bond angle with $z = 0.5 \cdot (1 - \cos(\theta))$ and a range of $z = [0, 1]$ for $\theta = [0^\circ, 180^\circ]$. Even though the Morse potential is not designated to reproduce the potentials angular dependency but it does fit well around the equilibrium angle of water with $\theta^{\mathrm{eq}} \approx 104^\circ$. Finally, a Reproducing Kernel Hilbert Space (RKHS)[15, 16] V_{RKHS} potential was also constructed that contributes a 3-body correction term to fit the reference points.

$$V_{\mathrm{PES}} = V_{\mathrm{Morse}}\left(r_1; r_2^{\mathrm{eq}}, z^{\mathrm{eq}}\right) + V_{\mathrm{Morse}}\left(r_2; r_1^{\mathrm{eq}}, z^{\mathrm{eq}}\right) + V_{\mathrm{Morse}}\left(z; r_2^{\mathrm{eq}}, r_1^{\mathrm{eq}}\right) + V_{\mathrm{RKHS}}\left(r_1, r_2, z\right) \tag{1}$$

In general, the potential energy function reproduces the reference values on the grid points with a root mean square error about $\lessapprox 10^{-4}$ Hartree ($\lessapprox 22$ cm^{-1}). The parameters of the 1-dimensional Morse potential terms were optimized to fit the reference points along one internal coordinate, with the remaining two frozen at their equilibrium values. The PES was used together with the DVR3D program suite[17] to determine the first vibrational eigenvalues of the vibrational ground state, first and second bending mode, and the symmetric and asymmetric stretching mode of water. The DVR3D program suite requires the definition of system sensitive Morse-type parameter that are optimized within reasonable range to fit the water symmetric stretch frequency.

In principle, this *ab initio* method does not require any further parameters other than the hydrated system composition. Expecting to optimize to the corresponding optimized structure as measured in the experiments, it provides an estimation of the vibrational frequencies based on the quantum methods level of theory and quality of the PES representation. In contrast, it lacks the vibrational coupling with the remaining degrees of freedom as the frustrated translation, rotation, and vibrations of the molecule adsorbent. It further requires a significant effort in computational power with respect to the chosen level of theory.

1.4 Kernel Based Prediction - Luis I. Vazquez-Salazar

Here, the predictions of the frequencies were performed using the FCHL19 representation[10] as implemented on the *QML* Python toolkit[11]. The descriptor parameters were configured as recommended in the original paper for use on datasets containing only energies. The size of the descriptor was set to 33. The

width of the kernel was manually adjusted as 1.2. Matrix diagonalization was performed using Cholesky decomposition with a regularization parameter (*L2*) of 10^{-10}.

initially a kernel was trained on the harmonic frequencies of the generated molecules selected from the GDB-11[5, 6] (see Sec. 1.1). The molecules for which there were available experimental values were used to validate the training kernel. The frequencies were standardized by subtracting the mean of the frequencies on the training set and then dividing by the standard deviation of those values. The MAE of the predicted values with respect to the experimental values was 95.6 cm^{-1}.

Next, the difference between the predicted frequencies with the kernel and the experimental values was learned following the Δ-Learning method[18]. A second kernel was trained with difference between the values predicted by the kernel method on the first step and the experimental values. As before the FCHL19 descriptor was used for the molecules on the training set of the hydra challenge with the same configuration as on the previous. Finally, the frequencies obtained for the test molecules with the kernel model trained during the first step were corrected with the predicted values of the difference($\Delta = \nu_{exp} - \omega_{kernel}$) on the second step.

References

[1] A. D. Becke. Becke's three parameter hybrid method using the LYP correlation functional. *J. Chem. Phys.*, 98(492):5648–5652, 1993.

[2] R. A. Kendall, T. H. Dunning Jr, and R. J. Harrison. Electron affinities of the first-row atoms revisited. Systematic basis sets and wave functions. *J. Chem. Phys.*, 96(9):6796–6806, 1992.

[3] S. Grimme, J. Antony, S. Ehrlich, and H. Krieg. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.*, 132(15):154104, 2010.

[4] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, B. Mennucci, G. A. Petersson, H. Nakatsuji, M. Caricato, X. Li, H. P. Hratchian, A. F. Izmaylov, J. Bloino, G. Zheng, J. L. Sonnenberg, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. Bearpark, J. J. Heyd, E. Brothers, K. N. Kudin, V. N. Staroverov, R. Kobayashi, J. Normand, K. Raghavachari, A. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, N. Rega, J. M. Millam, M. Klene, J. E. Knox, J. B. Cross, V. Bakken, C. Adamo, J. Jaramillo, R. Gomperts, R. E. Stratmann, O. Yazyev, A. J. Austin, R. Cammi, C. Pomelli, J. W. Ochterski, R. L. Martin, K. Morokuma, V. G. Zakrzewski, G. A. Voth, P. Salvador, J. J. Dannenberg, S. Dapprich, A. D. Daniels, O. Farkas, J. B. Foresman, J. V. Ortiz, J. Cioslowski, and D. J. Fox. Gaussian 09 Revision E.01. Gaussian Inc. Wallingford CT 2009.

[5] T. Fink, H. Bruggesser, and J.-L. Reymond. Virtual exploration of the small-molecule chemical universe below 160 daltons. *Angew. Chem. Int. Ed.*, 44(10):1504–1508, 2005.

[6] T. Fink and J.-L. Reymond. Virtual exploration of the chemical universe up to 11 atoms of C, N, O, F: assembly of 26.4 million structures (110.9 million stereoisomers) and analysis for new ring systems, stereochemistry, physicochemical properties, compound classes, and drug discovery. *J. Chem. Inf. Model.*, 47(2):342–353, 2007.

[7] G. Landrum et al. RDKit: Open-source cheminformatics. https://www.rdkit.org (Q1 2020) Release, March 2020.

[8] W. Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallogr. A*, 32(5):922–923, Sep 1976.

[9] F. A. Faber, A. S. Christensen, B. Huang, and O. A. Von Lilienfeld. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.*, 148(24):241717, 2018.

[10] A. S. Christensen, L. A. Bratholm, F. A. Faber, and O. Anatole von Lilienfeld. FCHL revisited: Faster and more accurate quantum machine learning. *J. Chem. Phys.*, 152(4):044107, 2020.

[11] A. Christensen, F. Faber, B. Huang, L. Bratholm, A. Tkatchenko, K. Muller, and O. von Lilienfeld. QML: A Python toolkit for quantum machine learning. *URL https://github. com/qmlcode/qml*, 2017.

[12] D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[13] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al. Tensorflow: A system for large-scale machine learning. In *12th USENIX symposium on operating systems Design and Implementation (OSDI 16)*, pages 265–283, 2016.

[14] D. O. Harris, G. G. Engerholm, and W. D. Gwinn. Calculation of Matrix Elements for One-Dimensional Quantum-Mechanical Problems and the Application to Anharmonic Oscillators. *J. Chem. Phys.*, 43(5):1515–1517, 1965.

[15] T. Ho and H. Rabitz. A general method for constructing multidimensional molecular potential energy surfaces from ab initio calculations. *J. Chem. Phys.*, 104(7):2584–2597, 1996.

[16] O. T. Unke and M. Meuwly. Toolkit for the Construction of Reproducing Kernel-Based Representations of Data: Application to Multidimensional Potential Energy Surfaces. *J. Chem. Inf. Model.*, 57(8):1923–1931, 2017.

[17] J. Tennyson, M. A. Kostin, P. Barletta, G. J. Harris, O. L. Polyansky, J. Ramanlal, and N. F. Zobov. DVR3D: a program suite for the calculation of rotation–vibration spectra of triatomic molecules. *Comput. Phys. Commun.*, 163(2):85–116, 2004.

[18] R. Ramakrishnan, P. O. Dral, M. Rupp, and O. A. von Lilienfeld. Big data meets quantum chemistry approximations: the Δ-machine learning approach. *J. Chem. Theory Comput.*, 11(5):2087–2096, 2015.