Re-evaluation of the Hydra results for Method LS2 (Δ -learning)

Luis Itza Vazquez-Salazar and Markus Meuwly*

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

E-mail: m.meuwly@unibas.ch

After the publication of the final results, we revisited the procedure for obtaining the frequencies using the Δ -learning approach, ¹ referred to as LS2 in the main manuscript. This was primarily motivated by the observation that the predictions from Δ -learning were unexpectedly inferior to all other similar approaches. For the rest of the document, we will distinguish three data sets: a) the "training" data set for which anharmonic experimental frequencies are provided for given structures in the Hydra challenge, see Table 1; b) the "challenge" data set for which structures are given but no frequencies; predicting the (anharmonic) frequencies for these molecules is the task; c) the "harmonic" dataset with frequencies for molecules from the GDB-11 dataset which are referred to as the "harmonic" dataset. This set contains 213 molecules, and the frequencies are used to train a first model to predict harmonic frequencies for molecules with given structure.

In the following, two models are required which are referred to as "Model 1" (M1) and "Model 2" (M2). The first model uses harmonic frequencies determined at the B3LYP/aug-cc-pVTZ+D3 level of theory for 213 molecules taken from GDB-11 and the "training" set form the challenge. The second model carries out the Δ -learning step using the "training" data provided in the hydra challenge.

Model 1: For training the first model, all frequencies ω_i of the "harmonic" dataset were first normalized (standardized) according to

$$\hat{\omega}_i = \frac{\omega_i - \mu}{\sigma} \tag{1}$$

The values $\hat{\omega}_i$ were used in fitting the kernel, ω_i is the frequency of the *i*-th molecule, $\mu = \sum \omega_i / N$ is the average frequency, and σ their standard deviation. For the 'harmonic' set, the corresponding values were $\mu_{\text{harm}} = 3633.8 \text{ cm}^{-1}$ and $\sigma_{\text{harm}} = 76.7 \text{ cm}^{-1}$. These harmonic frequencies $\hat{\omega}_i$ need to be transformed to anharmonic frequencies according to

$$\mathbf{v}_j = \sigma_{\mathbf{v}} \hat{\boldsymbol{\omega}}_j + \boldsymbol{\mu}_{\mathbf{v}} \tag{2}$$

This is necessary in order to have the values on a scale that corresponds to the anharmonic experimental frequencies which are known to be at lower wavenumbers compared with the harmonic frequencies. In the original submission, the values used for this inversion were μ_{harm} and σ_{harm} instead of σ_v and ω_v , which is incorrect. With the known experimental anharmonic frequencies from the "training" set, the values are $\mu_v = 3557.3 \text{ cm}^{-1} \sigma_v = 59.1 \text{ cm}^{-1}$. Using these values in equation 2, makes the predicted values for the kernel comparable with the experimental results, which otherwise will be systematically overestimated (See Figure 1) due to neglecting anharmonicity.

Model 2: Based on the anharmonic frequencies $v_{tset}^{(M1)}$ from evaluating "model 1" for the "training" dataset and the given experimental values $v_{tset}^{(Exp)}$ from the hydra challenge, a second model is trained on the difference $\Delta = v_{tset}^{(M1)} - v_{tset}^{(Exp)}$, which is the Δ -learning step.

For predicting anharmonic frequencies for molecules *j* from the "challenge" dataset, first "model 1" is evaluated to yield $\hat{\omega}_j$ from which $v_j^{(M1)}$ is obtained and "model 2" provides the correction due to Δ -learning. In other words, both models need to be evaluated to obtain a prediction. The

final prediction will be:

$$v_j = v_j^{(M1)} + \Delta^{(M2)}$$
 (3)

where $v_j^{(M1)}$ corresponds to the predicted anharmonic frequency from harmonic frequencies according to Eq. 2 (model 1), and $\Delta^{(M2)}$ is the correction from Δ -learning (model 2).



Figure 1: Scatter plot of the experimental frequencies vs the predicted frequencies (see Table 1) for the training set of the hydra challenge. The trend line corresponds to the linear regression for the provided points. Meanwhile, the individual points are individual values. The MAE was 96 cm⁻¹ and 23 cm⁻¹ for initial and corrected submissions respectively. Orange: original submission; blue: corrected procedure as outlined here. The black diagonal line is the 1:1 correlation between experiment and prediction. Note that the blue and orange trend lines diverge towards higher frequency which correctly indicates increasing anharmonicity and the blue line is considerably closer to the 1:1 correlation.

Using the correct normalization values (μ_v and σ_v) the new results for the training set of the challenge are shown in Table 1 and Figure 1 (blue). This considerably reduces the MAE between predicted and experimentally measured frequencies from 96 cm⁻¹ (original submission, orange) to 23 cm⁻¹ (present work) because the correct normalization scale was used.

]	Molecule	Exp	$\omega_{\rm init}$.	$\Delta \omega_{\rm init}$	$v_{\rm new}$	$\Delta v_{\rm new}$
	288-32-4	3458	3575.9	-117.9	3512.7	-54.7
	67-64-1	3538	3621.7	-83.7	3548.2	-10.2
	98-86-2	3536	3608.6	-72.6	3538.0	-2.0
	62-53-3	3524	3598.8	-74.8	3530.4	-6.4
	98-85-1	3620	3768.8	-148.8	3662.0	-42.0
2	2406-25-9	3484	3481.4	2.6	3439.6	44.4
	327-54-8	3647	3782.7	-135.7	3672.8	-25.8
1	191-95-3	3548	3630.6	-82.6	3555.1	-7.1
	132-64-9	3623	3756.5	-133.5	3652.5	-29.5
	611-20-1	3595	3698.8	-103.8	3607.8	-12.8
	MAE			96		23

Table 1: Frequencies for the molecules in the training set of the hydra challenge. All the values are in cm^{-1} .

The values for the challenge set follow a similar trend. Using Eq. 2 the MAE reduces to 37 cm^{-1} compared with 55 cm⁻¹ from the original submission and the RMSE improves to 54 cm⁻¹ compared with 74 cm⁻¹, see Table 2.

Table 2: Predicted anharmonic frequencies for the molecules from the challenge set. Values for 'model 1' and the final prediction are also shown. All the values are in cm^{-1} .

Molecule	$v^{(Exp)}$	$\omega_{ m init}$	$\Delta \omega_{\rm init}$	$v_{\rm new}^{(\rm M1)}$	$v_{\rm new}^{\rm Final}$	$\Delta v_{\rm new}$
125132-75-4	3597	3547.6	49.4	3620.7	3585.4	11.6
434-45-7	3611	3587.2	23.8	3603.0	3589.4	21.6
110-01-0	3507	3594.6	-87.6	3616.7	3585.5	-78.5
50-0-0	3591	3586.6	4.4	3602.3	3570.3	20.7
547-64-8	3524	3425.4	98.6	3510.3	3479.7	44.3
75-89-8	3649	3482.6	166.4	3578.9	3524.0	125.0
110-86-1	3454	3388.7	65.3	3440.7	3405.5	48.5
502-49-8	3503	3542.9	-39.9	3514.5	3541.6	-38.6
80-73-9	3492	3506.5	-14.5	3548.8	3509.0	-17.0
109-99-9	3491	3492.4	-1.4	3543.9	3511.6	-20.6
MAE			55			37
RMSE			74			54
		'	. ,			

Up to this point the kernel was trained on the 'training' set of the challenge and the information from the 213 molecules from GDB-11. A final model was trained by adding the harmonic frequencies for the molecules of the 'challenge' set to the pool of molecules used to obtain the kernel. This is akin to method LS3 (Transfer Learning). For this case, the values for σ_v and μ_v used for the anharmonic frequency prediction were the same as in the previous test; on the other hand, the values of the harmonic dataset were $\mu_{harm} = 3634.1 \text{ cm}^{-1}$ and $\sigma_{harm} = 77.0 \text{ cm}^{-1}$. It was observed that the predictions (see Table 3 for molecules in the 'challenge' dataset) improved by reducing the MAE from 37 cm⁻¹ to 26 cm⁻¹ and the RMSE from 54 cm⁻¹ to 49 cm⁻¹ with respect to the experimental values compared with an RMSE of ~ 12 cm⁻¹ for the predictions from LS3.

		1		1	
Molecule	Exp	$\omega_{\rm init}$.	$\Delta \omega_{\rm init}$	<i>v</i> _{new}	$\Delta v_{\rm new}$
125132-75-4	3597	3547.6	49.4	3568.0	29.0
434-45-7	3611	3587.2	23.8	3601.8	9.2
110-01-0	3507	3594.6	-87.6	3492.9	14.1
50-0-0	3591	3586.6	4.4	3596.4	-5.4
547-64-8	3524	3425.4	98.6	3532.6	-8.6
75-89-8	3649	3482.6	166.4	3500.7	148.3
110-86-1	3454	3388.7	65.3	3446.8	7.2
502-49-8	3503	3542.9	-39.9	3470.5	32.5
80-73-9	3492	3506.5	-14.5	3503.4	-11.4
109-99-9	3491	3492.4	-1.4	3480.1	10.9
MAE			55		26
RMSE			74		49
	I	1	1	1	1

Table 3: Predicted anharmonic frequencies for the molecules in the challenge set considering all the molecules ('harmonic', 'training', and 'challenge'). All the values are in cm^{-1} .

As a final comment on the present application it should be mentioned that we used the hyperparameters of FCHL19² that were optimized for predicting total energies. This implies that the hyperparameters are best suited for fitting energy functions. Further improvements of the present approach may be obtained by re-optimizing the hyperparameters of the representation used for training harmonic frequencies. Re-optimization of hyperparameters for the particular application at hand is likely to yield further improvements also because the original FCHL model had prediction errors of the order of 50 cm⁻¹ to 300 cm⁻¹ for the highest harmonic frequency.³ Additional improvements can also be expected from using operator-based machine learning.⁴ Therefore, it must be stated that the used Δ -learning method needs to be adjusted for the specific problem at hand.

References

- 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. 2015, 11, 2087–2096.
- (2) Christensen, A. S.; Bratholm, L. A.; Faber, F. A.; Anatole von Lilienfeld, O. FCHL revisited: Faster and more accurate quantum machine learning. *J. Chem. Phys.* 2020, *152*, 044107.
- (3) Faber, F. A.; Christensen, A. S.; Huang, B.; Von Lilienfeld, O. A. Alchemical and structural distribution based representation for universal quantum machine learning. *J. Chem. Phys.* 2018, *148*, 241717.
- (4) Christensen, A. S.; Faber, F. A.; von Lilienfeld, O. A. Operators in quantum machine learning: Response properties in chemical space. *J. Chem. Phys.* 2019, *150*, 064105.