Supplemental Information: Using Gaussian Process Regression to Simulate the Vibrational Raman Spectra of Molecular Crystals

Nathaniel Raimbault‡, Andrea Grisafi‡, Michele Ceriotti, Mariana Rossi

1 Fritz Haber Institute of the Max Planck Society, Faradayweg 4-6, 14195 Berlin, Germany
2 École Polytechnique Fédérale de Lausanne, Route Cantonale, 1015 Lausanne, Switzerland

E-mail: michele.ceriotti@epfl.ch and rossi@fhi-berlin.mpg.de

Submitted to: New J. Phys.

‡ These authors contributed equally to the manuscript.
1. Optimizing hyperparameters

In order to find the optimal hyperparameters $\sigma$ and $\eta$ used in GPR, we have used a cross-validation method. We picked a set of structures (here by farthest point sampling) from an NVT trajectory, which we then divided into a training and validation set. We then varied the hyperparameters and calculated the error (see Eq. 6) on the validation set in each case. In Fig. S1 we show an example of a contour plot of $\epsilon_{xx}$ for Paracetamol I. Very similar profiles are obtained for all polarizability components. The values for $\sigma$ and $\eta$ given in the main text have been optimized using random subselections of 900 structures for the training and 300 for the validation set.

A similar procedure has been carried out to optimize the radial cutoff $r_{\text{cut}}$ and the Gaussian width $\sigma$ entering the SOAP-kernel definitions. In this case, random subselections of 1500 configurations out of a total of 2000 has been chosen to train the model while the remaining 500 are selected as a validation set. Both for the monomer and the crystal case, we find that $\sigma = 0.3$ Å and $r_{\text{cut}}=4.0$ Å yield the best learning performances. We multiplied the linear $\lambda$-SOAP kernel by a $\lambda = 0$ scalar kernel with the same parameters, so as to obtain a non-linear kernel that captures higher-order structural correlations without sacrificing the covariant behavior of the model \[1\].

2. Learning curves

2.1. Atomic coordinates

In the main text, Gaussian process regression was systematically used with a density-grid representation. In this section we show the performance of a GPR approach based on unprocessed coordinates in Fig. S2. Note that representing the system through atomic coordinates only entails using the cartesian components as the feature vector $u_A$, thus requiring an alignment to a reference structure. We observe that the learning curve when simply taking raw coordinates to train the model shares a similar behaviour with the atomic density representation, but is systematically a little lower, meaning that it needs more training points to reach the same level of accuracy. Therefore the simplest representation with unprocessed atomic coordinates was not considered in the rest of this article, where GPR is systematically employed in combination with atomic densities.

3. Fully converged Raman spectrum (molecule)

To make the parallel with Figs 10 and 11 in the main text, we show in Fig. S3 the Raman spectrum of the Paracetamol molecule, using SA-GPR trained on different training models, containing each 2000 structures.

We see that with this amount training points, SA-GPR reproduces exactly the \textit{ab initio} spectrum of the Paracetamol molecule; At this level of accuracy, the error bars are barely visible, and low frequencies and high frequencies are equally well described.

Furthermore, comparison with Fig. 10 in particular highlights the difficulty in predicting a molecular crystal rather than one of its molecular constituents: many more training points are needed in the case of Paracetamol I to reach the same level of accuracy as that of the molecule.
4. Variations in polarizability predictions

In the main text, we showed results that were obtained each time with a unique subselection of the training set. To show the dependence of our predicted polarizabilities on the choice of the training data, we chose a fixed validation set, but varied the selection of training data, keeping its size constant. Results are shown in Fig. S4. While some substantial modifications are observed on a point-by-point comparison basis, the general shape and variations are extremely similar for each of the subselections. Consequently, the corresponding Raman spectrum is impacted very little by such variations, as we demonstrate in Fig. S5. It is thus usually sufficient to have a single model to predict new trajectories, as long as one uses enough training points.

5. Molecules in crystal

Let us have a look at the Raman spectrum of a single molecule as behaving inside the crystal, which we shall call "interacting molecule", and compare it to that of the isolated molecule, shown in Fig. S6. We see that the two spectra are very different. The very high-frequency peaks of the isolated molecules (O-H and N-H stretching motions) disappear in the case of the interacting molecule, and also the low-frequency range shows some important changes. Actually, the
interacting-molecule spectrum looks much closer to the full crystal spectrum.

It actually seems somewhat counter-intuitive that the molecular spectrum would differ that much at high frequency. Indeed, the higher-frequency vibrational modes should be governed by intramolecular vibrations, and in this regard are expected to be similar within the crystal. Interestingly now, if we compute directly the autocorrelation function of \( \alpha \Sigma \), i.e., the sum of molecular contributions, and compare it to the full crystal spectrum, we find a very good agreement, as shown on Fig. S7, provided that we apply \textit{a posteriori} a fixed scaling factor for each different frequency range. We see in particular that the three main peaks of the low-frequency range are present in the molecular spectrum, although not with the right intensity. We note that for the latter figure, we used Gaussian process regression for predicting the individual molecular polarizabilities, but we did not use any machine-learning scheme to predict directly the crystal polarizabilities. All the information coming from the interaction between the different molecules is thus absent, and still needs to be grasped by a machine-learning approach.

As illustrated in Fig. S8 for GPR, the prediction accuracy can be improved even further if one scales the molecular polarizability tensors so that their average matches that of the full crystal. The variations of \( \alpha \Sigma \) are very close to that of the full polarizability of the crystal, but both its absolute values and derivatives are considerably smaller. The scaled version reduces these differences.

6. Harmonic vibrational Raman spectra

A more widely used possibility to simulate vibrational Raman spectra in an \textit{ab initio} manner is through the harmonic approximation, in which the potential energy is Taylor-expanded up to second order: Within this approximation, harmonic Raman intensities \( I_H(\omega) \) are proportional to the derivatives of the polarizability tensor with respect to atomic displacements (see e.g., Refs. [2, 3, 4]):

\[
I_H(\omega) \propto \frac{1}{\omega} \sum_{ij} \left\{ \left[ (\alpha'_{xx})^2 + (\alpha'_{yy})^2 + (\alpha'_{zz})^2 \right] - \langle \partial^2 \alpha_{ij} / \partial Q_p \rangle \langle \partial^2 \alpha_{ij} / \partial Q_p \rangle \right\}
\]

where \( \langle \partial^2 \alpha_{ij} / \partial Q_p \rangle \) is the derivative of the \( ij \) component of the polarizability with respect to the displacement of normal mode \( Q_p \). In practice, we evaluate these derivatives by finite differences: we make \( 6n \) (forward and backward, \( n \) being the number of atoms per unit cell) small nuclear displacements in the unit cell around the equilibrium position, on top of which we compute the polarizability tensor with DFPT.

Because in the harmonic approximation, every calculation takes place close to equilibrium, the standard deviation of the polarizability in that case will be very small. The challenge to predict the latter with machine learning is thus slightly different than in the previous case, where polarizabilities can experience large variations. Here the model will have to be sensitive to very small symmetric alterations in the structure.

We show the results for Paracetamol I in Fig. S9 where we train on an NVT trajectory, as described in the main text, and use Eq. 15. We see that the predicted harmonic spectrum reproduces nicely the \textit{ab initio} one. We note that the predicted polarizabilities
experience a small global offset, which does not impact the spectrum.

Beyond this agreement, these results also signify that the computational cost of a harmonic Raman spectrum can be basically reduced to the cost of a simple geometry optimization.

7. Offset in predictions of Paracetamol II

In this section we show in Fig. S10 that the prediction of the individual susceptibility components that enter the Paracetamol II Raman spectrum presented in the main text (based on the GPR trained for form I) show different offsets with respect to the DFPT values.

Figure S10. Polarizability time series for 3 components, obtained either with DFPT or with SA-GPR.

[1] Wilkins D M, Grisafi A, Yang Y, Lao K U, DiStasio R A and Ceriotti M 2019 Proc. Natl. Acad. Sci. USA 116 3401– 3406
[2] Neugebauer J, Reiher M, Kind C and Hess B A 2002 J. Comput. Chem. 23 895–910
[3] Veithen M, Gonze X and Ghosez P 2005 Phys. Rev. B 71(12) 125107
[4] Prosandeev S A, Waghmare U, Levin I and Maslar J 2005 Phys. Rev. B 71 214307