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

Nathaniel Raimbault[1]‡**, Andrea Grisafi**[2]‡**, Michele Ceriotti**[2]**, Mariana Rossi**[1]

[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`

Abstract. Vibrational properties of molecular crystals are constantly used as structural fingerprints, in order to identify both the chemical nature and the structural arrangement of molecules. The simulation of these properties is typically very costly, especially when dealing with response properties of materials to e.g. electric fields, which require a good description of the perturbed electronic density. In this work, we use Gaussian process regression (GPR) to predict the static polarizability and dielectric susceptibility of molecules and molecular crystals. We combine this framework with ab initio molecular dynamics to predict their anharmonic vibrational Raman spectra. We stress the importance of data representation, symmetry, and locality, by comparing the performance of different flavors of GPR. In particular, we show the advantages of using a recently developed symmetry-adapted version of GPR. As an examplary application, we choose Paracetamol as an isolated molecule and in different crystal forms. We obtain accurate vibrational Raman spectra in all cases with less than 1000 training points, and obtain improvements when using a GPR trained on the molecular monomer as a baseline for the crystal GPR models. Finally, we show that our methodology is transferable across polymorphic forms: we can train the model on data for one structure, and still be able to accurately predict the spectrum for a second polymorph. This procedure provides an independent route to access electronic structure properties when performing force-evaluations on empirical force-fields or machine-learned potential energy surfaces.

‡ These authors contributed equally to the manuscript.

1. Introduction

Machine-learning (ML) models are becoming increasingly popular in the field of atomistic simulations, providing a way to obtain data-driven physical insights [1, 2, 3] and reduce the cost of simulations [4, 5]. Most efforts have been concentrated into predicting total energies and forces from atomic coordinates [6, 7, 8, 5, 9, 10, 11, 12], which are most often the largest cost in a first-principles simulation. More recently, machine-learning models have been also applied to the prediction of response properties of molecules [13, 14, 15, 16]. When dealing with the response of a material to an applied field, the cost of a first-principles calculation is often larger than that of force evaluation. This is thus an area where one can also take advantage of supervised learning techniques in order to reduce the cost *ab initio* simulations that that make use of such response properties.
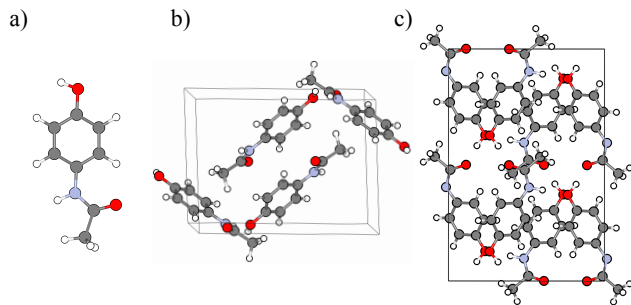
a)   b)   c)

Figure 1. The systems considered in this work for the prediction of vibrational Raman spectra. (a) Isolated Paracetamol molecule. (b) Paracetamol crystal form I (monoclinic). (c) Paracetamol crystal form II (orthorhombic). Atomic color code: hydrogen white, nitrogen blue, carbon grey, oxygen red. The unit cell is drawn in black.

Vibrational Raman spectra are a good example of a property that requires the knowledge of the response of the system to electric field perturbations. The Raman signal is very useful to monitor phase transitions, as well as for the identification of global and local structural patterns [17, 18, 19]. Any technique used to calculate this property requires the calculation of several instances of the polarizability tensor (in molecules) or the dielectric susceptibility (in crystals). Previously, some of the present authors have shown that anharmonic vibrational Raman spectra calculated through a time-correlation formalism can be a powerful tool to identify structural fingerprints in molecular crystals [20, 21]. Within this formalism, it is necessary to calculate *ab initio* molecular dynamics trajectories and compute the response quantities for subsequent atomic configuration, employing, for instance, density-functional perturbation theory (DFPT) [22, 23, 24, 25, 20]. These calculations are computationally demanding, not only because of the tens of thousands of force evaluations that need to be performed to provide sufficient statistical sampling, but also because each DFPT calculation is typically four times more expensive than a force evaluation [21]§. This cost is inherent to the method used and, to complicate matters, there is no known empirical form that can be used to easily fit the polarizability tensor.

We here investigate and propose frameworks to obtain accurate predictions of dielectric response functions for a number of consecutive molecular dynamics (MD) configurations that are necessary to converge simulated vibrational Raman spectra of molecules and molecular crystals. In order to reach this goal, we employ different flavors of Gaussian Process Regression (GPR), which is a method that has already been proven to be efficient in predicting dielectric response properties [13, 15, 26]. In particular, we compare standard GPR schemes with symmetry-adapted GPR (SA-GPR)[27], which is advantageous when describing tensorial quantities. For the former, we exploit the internal structural rigidity of the system in order to model each individual component of the polarizability tensor, therefore remapping the tensor learning problem onto many scalar regression tasks. For the latter, we employ a symmetry-adapted representation of the system in order to learn the irreducible spherical components of the tensor in a covariant fashion. As a trial system we consider the Paracetamol molecule and form I and form II of the Paracetamol crystal, represented in Fig. 1. As we demonstrate below, SA-GPR comes out as the methodology with the best performance. When predicting the Raman spectra of crystals, both methods can benefit from using a GPR trained on the monomer as a baseline. We also find that our model is transferable between different polymorphic forms. Given that empirical and machine-learned potential energy surfaces are becoming more accurate for molecular crystals, the methodology proposed here can be combined in a straightforward manner to such potentials, giving access to the electronic polarization and polarizability of crystals.

In the following, Section 2 introduces the general machinery describing GPR and SA-GPR, as well as the different representations fed to these algorithms. In Section 3, we present applications on the Paracetamol molecule and monoclinic form I crystal, showing for the latter how our models can be refined by the inclusion of molecular polarizability tensors. Finally, we illustrate in Section 4 the transferability of the SA-GPR method, by predicting the Raman spectrum of the orthorhombic form II crystal without prior

§ This number is clearly system- and settings-dependent, and simply represents the estimate reported in Ref.[21].

knowledge of the corresponding values of the dielectric response.

2. Theory

2.1. Vibrational Raman Spectra

The central quantity needed for simulating vibrational Raman spectra is the static polarizability tensor $\boldsymbol{\alpha}$. For simplicity, we will refer to the polarizability tensor for all the rest of the paper, but one should keep in mind that for solids the quantity of interest is rather the electric susceptibility tensor of the system.

As discussed in Ref. [21] and others [28, 29], the vibrational Raman spectrum can be calculated using several approximations, the simplest of which is the harmonic approximation. The framework developed in this paper can be applied to the harmonic case, as shown in the supplemental information, Fig. S8, but we will here focus on the more challenging task of applying it to the linear-response time-correlation formalism, which includes all the anharmonicity of the potential energy surface. In this formalism, vibrational Raman intensities can be obtained from the Fourier transform of the static polarizability autocorrelation function [30] at thermodynamic equilibrium. In particular, the so-called powder spectrum intensity is given by a combination of the anisotropic and isotropic contributions as

$$
\begin{aligned}
I(\omega) \quad &= I_{\text{iso}}(\omega) + \frac{7}{3} I_{\text{aniso}}(\omega) \\
I_{\text{iso}}(\omega) \quad &= \frac{n}{2\pi} \int_{-\infty}^{+\infty} dt e^{-i\omega t} \langle \bar{\alpha}(0)\bar{\alpha}(t)\rangle \\
I_{\text{aniso}}(\omega) &= \frac{n}{2\pi} \int_{-\infty}^{+\infty} dt e^{-i\omega t} \frac{1}{10} \langle \text{Tr}[\tilde{\boldsymbol{\alpha}}(0)\cdot\tilde{\boldsymbol{\alpha}}(t)]\rangle \,,
\end{aligned}
\tag{1}
$$

where n is the number of atoms in the system, the brackets $\langle\cdot\rangle$ denote an ensemble average and Tr represents the trace. $\bar{\alpha}$ and $\tilde{\boldsymbol{\alpha}}$ are, respectively, the isotropic and anisotropic parts of the polarizability tensor, defined as follows,

$$
\begin{aligned}
\boldsymbol{\alpha} &= \bar{\alpha}\boldsymbol{I} + \tilde{\boldsymbol{\alpha}} \\
\bar{\alpha} &= \frac{1}{3}(\alpha_{xx} + \alpha_{yy} + \alpha_{zz}) \,, \ \ \text{Tr}[\tilde{\boldsymbol{\alpha}}] = 0 \,.
\end{aligned}
\tag{2}
$$

2.2. Component-wise Gaussian Process Regression

Gaussian process regression (GPR) is a well-established method based on a kernel function that measures the similarity between structures. In a Bayesian language, a kernel represents the prior distribution for the statistical correlations of the property we aim to predict. In the usual supervised-learning framework, a dataset of structures (i.e., atomic coordinates) and associated polarizabilities is used to train the model. Once the training is complete, the model is tested on a different set of configurations for which the polarizability is in principle not known. If it is possible to align the system to a reference structure, a straight-forward way to apply this procedure to tensorial quantities is to learn each component of the polarizability tensor separately. In particular, a GPR prediction of each individual polarizability component $\alpha_{\gamma\delta}$ reads

$$
\alpha_{\gamma\delta}^{\text{ML}}(\mathcal{A}) = \bar{\alpha}_{\gamma\delta}^{\text{ai}} + \sum_{j=1}^{N} w_j^{\gamma\delta} k(\mathcal{A}, \mathcal{A}_j)\,,
\tag{3}
$$

where N is the number of configurations included in the training set, γ and δ represent Cartesian coordinates, $w_j^{\gamma\delta}$ are the regression weights that need to be determined from the training data for each component, and k is the kernel that couples the target system $\mathcal{A}$, with the training structure $\mathcal{A}_j$. The quantity $\bar{\alpha}_{\gamma\delta}^{\text{ai}}$ is the average (over the training set) of the $\gamma\delta$ polarizability component computed from an *ab initio* method, which effectively allows the training to focus on the fluctuation of the property with respect to a known baseline value.

The kernel entering GPR is based on a Gaussian similarity measure between structures $\mathcal{A}$, given by

$$
k(\mathcal{A}, \mathcal{A}_j) = e^{-\frac{|\boldsymbol{u}(\mathcal{A})-\boldsymbol{u}(\mathcal{A}_j)|^2}{2\sigma^2}}\,,
\tag{4}
$$

with σ being a hyperparameter that controls the magnitude of the correlation between training points. The value of σ which needs to be tuned by cross-validation to maximize the prediction accuracy (see SI, section 1 for further details). $\boldsymbol{u}$ is a vector that has the role of mapping the atomic coordinates of the structure $\mathcal{A}$ to a given representation of dimension M.

The regression weights $\mathbf{w}^{\gamma\delta}$ are obtained by minimizing a loss function regularized by an L2-norm [31] over the training set. This procedure leads to the following expression

$$
\mathbf{w}^{\gamma\delta} = (\mathbb{K} + \eta\mathbb{1})^{-1} \cdot \Delta\boldsymbol{\alpha}_{\gamma\delta}\,,
\tag{5}
$$

where $\mathbb{1}$ is the identity matrix, $\mathbb{K}$ is the $N \times N$ kernel matrix associated with the reference structures, such that $\mathbb{K}_{ij} = k(\mathcal{A}_i, \mathcal{A}_j), i,j = 1, ..., N$, and η is the regularization parameter which controls to which extent the fitted data can deviate from the training points. The quantity $\Delta\boldsymbol{\alpha}_{\gamma\delta}$ represents the vector containing all N entries in the training set of $\Delta\alpha_{\gamma\delta}^{j} = \alpha_{\gamma\delta}^{\text{ai}}(\mathcal{A}_j) - \bar{\alpha}_{\gamma\delta}^{\text{ai}}$.

The efficiency of any GPR model strongly depends on the quality of the representation that is encoded in $\boldsymbol{u}(\mathcal{A})$ (see Eq. 3). For a representation to be efficient, it should contain the least possible number of elements to express the unicity of a given structure and avoid redundant information. Among the vast choice of representations one could think of, we will mainly employ two different ones; one connected with the GPR presented above, namely an atomic-density

grid representation (AD), and one connected to the symmetry-adapted Gaussian process regression (SA-GPR) technique [27] presented in the next section, namely λ-SOAP kernels.

Atomic-density grids consist of a conceptually simple representation. The procedure follows choosing a body-fixed Cartesian reference frame centered on the system under study, and defining a 3-dimensional grid around it. For each grid-point $\mathbf{r}$, we calculate the atomic density distribution, defined as

$$\rho_s(\mathbf{r}) = \sum_{i \in s} \exp\left(-\frac{|\mathbf{r} - \mathbf{r}_i|^2}{2\gamma_s^2}\right), \qquad (6)$$

where s identifies a specific atom type, $\mathbf{r}_i$ are the nuclear positions and γ_s is an adjustable parameter. The feature vector is given by $\boldsymbol{u}(\mathcal{A}) = (\rho_s(\mathbf{r})), s = 1, ..., N_s$, where N_s is the number of different atomic species. For the applications shown in this paper, we have chosen the same $\gamma_s = 0.5$ Å for all species. Such a representation is illustrated in Fig. 2, where a 2-dimensional cut of the paracetamol molecule density distribution is presented.
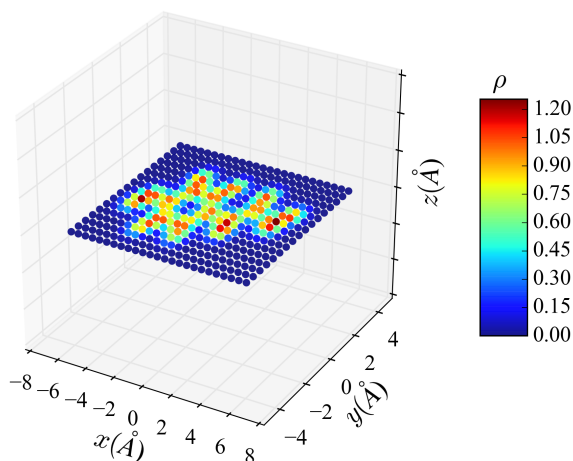

Figure 2. 2D view of the nuclear density distribution of the Paracetamol molecule for a given structure. Each grid point is represented by a colored sphere. Blue (red) indicates a low (high) density.

We have used a grid of evenly-spaced points spanning the maximal extension of the system, although we note that more refined methods could be utilized to define physically-motivated grid points, based on the possible directions of vibrations of atomic species [32].

Given that the components of the polarizability tensor are not invariant to rotations in cartesian space, a cartesian space representation of this quantity, like the atomic densities, requires an alignment to a reference structure. To do so, we have used the Kabsch algorithm considering only heavy atoms [33,

34]. We note that this alignment procedure is not applicable to all systems, in particular for very flexible molecules. For simple relatively rigid molecules, it is known to work well, as has been shown for the case of hyperpolarizabilities of water molecules before [13]. We will show below that even for a more complex and relatively flexible molecule like paracetamol, this representation still yields accurate predictions.

2.3. Symmetry-adapted Gaussian Process Regression with λ-SOAP Kernels

SA-GPR [27] represents a generalization of the GPR formalism, where the tensorial nature of the target property, together with its covariant 3D transformations, are naturally incorporated within the regression algorithm. This technique removes the need of often arbitrary alignment procedures of the systems (like the one described above) in order to predict tensorial quantities of any rank. As such, the model focuses only on the portion of variability of the tensor connected with an internal structural distortion of the molecular geometry, greatly improving the regression performances.

A simplification of the learning problem can be obtained if the target property is first decomposed in its irreducible spherical tensor components. The static polarizability (or the static susceptibility), in particular, being a symmetric rank-2 tensor, can be formally decomposed in an isotropic contribution that transforms as a spherical harmonic of angular momentum $\lambda = 0$, and an anisotropic contribution that transforms as a spherical harmonic of angular momentum $\lambda = 2$. The former contribution, $\alpha_0^{(0)} = \bar{\alpha}$, being directly proportional to the trace of the tensor (which is rotationally and translationally invariant), can be learned in the usual manner by a standard GPR. The entire tensorial nature is instead condensed in the $\lambda = 2$ contribution, $\boldsymbol{\alpha}^{(2)} = (\alpha_{-2}^{(2)}, \alpha_{-1}^{(2)}, \alpha_0^{(2)}, \alpha_1^{(2)}, \alpha_2^{(2)})$, which is related to the anisotropic part of the Cartesian $\boldsymbol{\alpha}$-tensor $\tilde{\boldsymbol{\alpha}}$ of Eq. 2 by a linear transformation. Within SA-GPR, the prediction of this contribution is carried out by making use of a tensorial kernel function $\boldsymbol{k}^{(2)}(\mathcal{A}, \mathcal{B})$, which is a matrix of size 5×5, that describes at the same time the structural similarity between molecular configurations and the tensorial geometric relationship of order $\lambda = 2$ between these configurations, as detailed in Ref. [27]. In general terms, such a kernel can be thought of as a generalization of a scalar kernel function such as the one introduced in Eq. 4, which makes use of the following covariant integration,

$$\boldsymbol{k}^{(2)}(\mathcal{A}, \mathcal{B}) = \int d\hat{R}\, \boldsymbol{D}^{(2)}(\hat{R})\kappa(\mathcal{A}, \hat{R}\mathcal{B}), \qquad (7)$$

where $\hat{R}$ represents the rotation operator and $\boldsymbol{D}^{(2)}(\hat{R})$

the associated Wigner-D matrix which has the role of expressing the rotation of $\lambda = 2$ spherical harmonics. In this definition, the scalar kernel $\kappa(\mathcal{A}, \mathcal{B})$ only needs to be invariant under rigid translations and rotations of the laboratory reference frame with respect to which both $\mathcal{A}$ and $\mathcal{B}$ are defined. In the present work, $\kappa(\mathcal{A}, \mathcal{B})$ is given by a superposition of atom-centered Gaussian densities, equivalent to the one used in the popular Smooth Overlap of Atomic Position (SOAP) kernel [35]. As such, the kernel of Eq. (7) represents the tensorial generalization of SOAP, usually called λ-SOAP, which recovers the scalar case of $\lambda = 0$ as a special limit [14].

Formally, a covariant $\lambda = 2$ prediction performed with a symmetry-adapted kernel function of the same order reads

$$
\begin{aligned}
\boldsymbol{\alpha}^{(2)}(\mathcal{A}) &= \sum_{i=1}^{N} \boldsymbol{k}_{\mathcal{A},i}^{(2)} \, \boldsymbol{w}_i^{(2)} \\
&= \sum_{i=1}^{N} \boldsymbol{k}_{\mathcal{A},i}^{(2)} \sum_{j=1}^{N} \left[\mathbb{K}^{(2)} + \eta \mathbb{1} \right]_{ij}^{-1} \boldsymbol{\alpha}_j^{(2)} ,
\end{aligned}
\tag{8}
$$

where i and j run over the N reference configurations used to train the regression model and η is the regularization parameter. The set of tensorial regression weights $\{\boldsymbol{w}_i^{(2)}\}$ are determined by inverting the $(5 \times N)^2$ kernel matrix $\mathbb{K}^{(2)}$ associated with the training structures, and projecting it on the vector of reference tensors $\{\boldsymbol{\alpha}_j^{(2)}\}$. In doing so, the statistical average of $\{\boldsymbol{\alpha}_j^{(2)}\}$ over the training set is assumed to be zero by symmetry. As such, no baseline of the anisotropic part of the polarizability is adopted when doing the regression.

2.4. Errors and uncertainty estimations

In order to gauge the accuracy of the machine-learned polarizability components, we calculate the root mean square error (RMSE) normalized by the standard deviation (STD) of the set we want to evaluate the error on,

$$
\epsilon_{\gamma\delta}[\sigma, \eta] = 100 \times \sqrt{\frac{\sum_j^N \left(\alpha_{\gamma\delta}^{\mathrm{ML}}(\mathcal{A}_j) - \alpha_{\gamma\delta}^{\mathrm{ai}}(\mathcal{A}_j) \right)^2}{\sum_i^N (\alpha_{\gamma\delta}^{\mathrm{ai}}(\mathcal{A}_j) - \bar{\alpha}_{\gamma\delta}^{\mathrm{ai}})^2}} ,
\tag{9}
$$

where j represents the jth configuration, N is the amount of points in the data set we consider, and the bar here denotes the average of the polarizability component over the dataset of interest.

Whenever the reference properties of the testing data are not available, one cannot make use of Eq. (9) to evaluate the error of the predicted polarizabilities. In these circumstances, one typically needs to estimate the error made on the predicted properties by making use of some *a priori* probabilistic criteria. In the

particular case of GPR, the expected error associated with a testing structure $\mathcal{A}$, can be computed as $\varepsilon(\mathcal{A}) = k(\mathcal{A}, \mathcal{A}) - \sum_{IJ} k(\mathcal{A}, I) k^{-1}(I, J) k(J, \mathcal{A})$, where I and J run over the training structures. As detailed in Ref. [36], this strategy is however not very practical because of its computational expense, so that other kind of methods such as bootstrapping or subsampling can rather be used to estimate the prediction errors. In addition, in the particular case of the present work one would like to propagate the error that occurs in the prediction of $\boldsymbol{\alpha}$ to the Raman spectrum. While this error propagation would be difficult to carry out on top of the GPR intrinsic covariance, it comes naturally from a bootstrapping or subsampling scheme.

In this work, several subselections of the training dataset are considered to obtain a propagated error of the predicted Raman spectrum. In particular, given N_{RS} the number of training subselections chosen, N_{RS} Raman spectra are computed from the predicted polarizabilities. Then, the average and the standard deviation of the predicted spectra over the m subselections give the final Raman spectrum prediction and the propagated estimated error respectively. The downside of this approach is that this model works under the assumption of an uncorrelated subsampling. This is of course not true in general, so that one needs to correct the model to take into account for the underlying correlations. Following Ref. [36], a maximum likelihood recipe can be adopted to linearly scale the variance of the predictions by a constant factor ν^2. The calibration of this scaling factor is carried out by computing the actual prediction errors of the polarizabilites over a suitably selected validation set N_{val}, for which the reference polarizabilities are known, and then considering

$$
\nu^2 = \frac{1}{N_{\mathrm{val}}} \sum_{j=1}^{N_{\mathrm{val}}} \frac{||\boldsymbol{\alpha}_{\mathrm{pred}}(j) - \boldsymbol{\alpha}_{\mathrm{ref}}(j)||^2}{\sigma^2(j)}
\tag{10}
$$

where $\sigma^2(j)$ are the variances of the predicted polarizabilities. Once the value of ν has been determined, each polarizability prediction of a given training model k can be updated by considering

$$
\boldsymbol{\alpha}_k' = \bar{\boldsymbol{\alpha}} + \nu(\boldsymbol{\alpha}_k - \bar{\boldsymbol{\alpha}})
\tag{11}
$$

where $\bar{\boldsymbol{\alpha}}$ is the predicted polarizability averaged over the training models. This definition guarantees that upon computing the Raman spectrum for each model k, the propagated uncertainty estimation associated with the predicted Raman spectra will automatically take into account the calibration of the variance previously considered.

2.5. Simulation and training details

To perform the *ab initio* calculations, we used the FHI-aims [37] program package with *light* settings for

all atomic species. We obtained aiMD trajectories using the PBE functional with many-body dispersion (MBD) corrections [38, 39], and employing a time step of 0.5 fs. The polarizabily/susceptibility was instead computed every 1 fs. Most of the data used here was already available from references [20] and [21]. For each system, we had 20 picoseconds of simulation in the NVT ensemble at 300 K. From this trajectory, a few thousands of configurations were selected to define the training and test set of our model. A full trajectory of 15 picoseconds in the NVE ensemble was instead considered as our validation set to assess the quality of our predictions, by comparing the predicted Raman spectra to the ones obtained with *ab initio* polarizabilities. We chose to train our model on the NVT ensemble and predict on the NVE ensemble. This choice is justified by the fact that the portion of configurational space sampled in the canonical ensemble is supposed to be larger than that sampled in the microcanonical ensemble.

3. Results

In the following, we shall apply the previously described methods to the Paracetamol system. Each time we mention GPR, it is used in combination with AD. Similarly, each time a mention we mention SA-GPR, it is associated to λ-SOAP.

3.1. Paracetamol molecule

We first consider the case of the isolated Paracetamol monomer (Fig. 1 (a)). We constructed the training dataset by selecting 2000 structures with farthest point sampling (FPS) using the scalar SOAP metric from the full NVT trajectory. For GPR, the three-dimensional density field was constructed within a box of $6 \times 4 \times 2.5$ Å, where the molecule had its longest axis along the x direction and the equilibrium geometry lied on the xy plane, and the grid spacing was $d\mathbf{r} = 0.5$ Å. The GPR was computed using $\sigma = 10$ and $\eta = 10^{-3}$. Details about the optimization of the hyperparameters are given in the supplemental information. Regression performances are reported in Fig. 3 (a), where the error ϵ, given by Eq. 9, computed on 500 randomly selected test structures (absent from the training set) out of the total of 2000 is shown as a function of the number of training monomers. ‖

Figure 3(a) shows that the learning capability of the model does not reach saturation within the training set sizes explored. The learning of all the polarizability components follow a similar slope, but they are predicted with different accuracy because of

‖ Learning performance when using only unprocessed atomic coordinates are also shown in the SI, Fig. S2.

the strong anisotropy of the paracetamol molecule, as follows. Because of the high π-conjugation of the system in the molecular plane, the system is much more polarizable along the x-axis rather than along other directions in this particular alignment, making it harder for the learning algorithm to capture the corresponding variations across the dataset. The α_{xx} component presents the largest error, going from about 40% with 300 training points to 17% with 1500 training points. The best learning performance is instead obtained for the α_{yz} component, where the prediction error can be brought down to about 6%.

In order to validate the model, we show in Fig. 4 the correlation between predicted and computed polarizability components on the validation set composed by the full independent NVE trajectory, for a representative training size of 900 molecular configurations, including the ϵ for each component. Although we observe a worsening of the predictions when compared to the previous case, where we trained and predicted on the same ensemble (and trajectory), the predicted polarizabilities are still well correlated to the reference values. The remaining question is how these errors translate to the actual prediction of the vibrational Raman spectrum.

In Fig 5, we show a machine-learned Raman spectrum averaged over 16 subselections of the training set of 900 configurations each, along with its standard deviation (see Sec. 2.4 for a detailed discussion about this procedure), and compare it to the one calculated from fully *ab initio* data. We find that the estimated variance has to be scaled by a factor of $\nu^2 = 2.0$. Despite a relatively small amount of training points, the agreement with the reference spectrum is excellent in the entire frequency range. As shown in Fig. 3, increasing the number of training points in the model would decrease errors even further.

From Figs. 3 and 5, it is clear that the error we make on the polarizability components does not translate directly into an error of similar magnitude on the spectrum. This is a consequence of the fact that the Raman intensities depend on the derivatives of the polarizability components with respect to atomic coordinates, and not on their absolute value.

Such a simple procedure is able to reproduce almost perfectly a reference Raman spectrum with less than 1000 training points on a desktop computer in just a few minutes. Given the clear advantages of SA-GPR to predict tensorial quantities, we compare the effectiveness of this method, for which λ-SOAP kernels were constructed using a Gaussian width of 0.3Å and an environment cutoff of 4.0Å, to GPR with AD representation. The corresponding learning curve is shown in Fig. 3(b). The improvement over a standard GPR scheme is systematic at any training
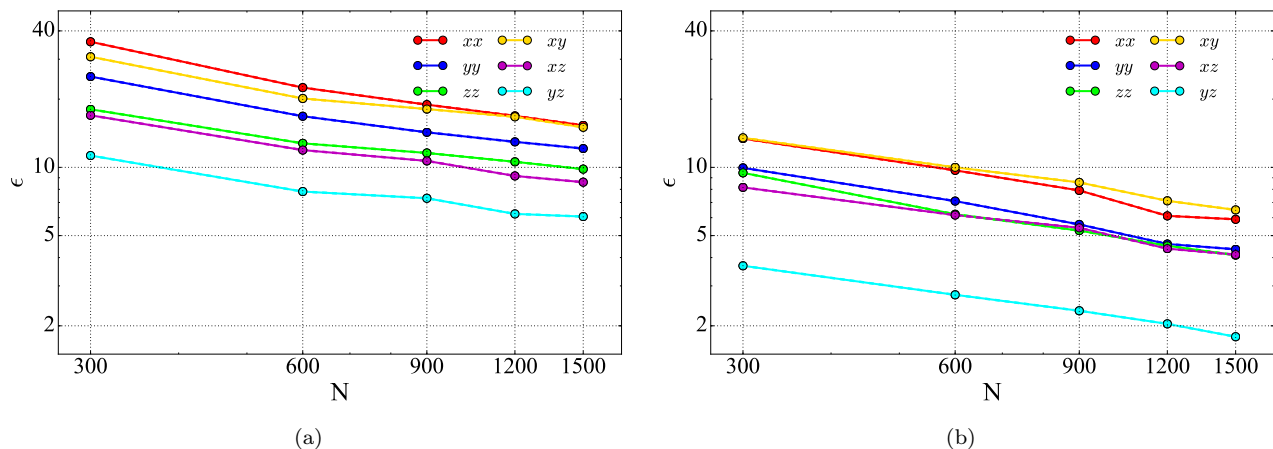
(a)

(b)

Figure 3. Prediction error (as defined by Eq. 9) on each component of the polarizability tensor of the paracetamol molecule with different models. (a) Learning curves from GPR using atomic densities as a representation. (b) Learning curves from SA-GPR with λ-SOAP kernels.
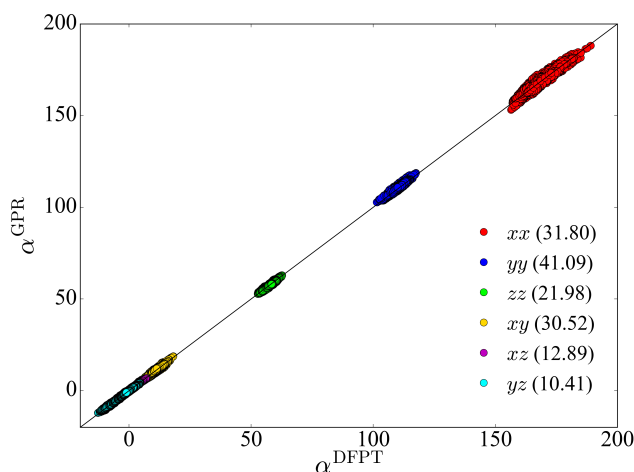

Figure 4. GPR polarizability tensor components versus DFPT *ab initio* ones. The components were trained on 900 configurations coming from an NVT trajectory. The test set contains 20000 configurations coming from an NVE trajectory. Numbers in brackets indicate ϵ for a given component.

set size, underlining the importance of automatically incorporating the $\mathcal{O}(3)$-covariance of the tensor at the local scale. Equivalently to GPR, with 300 training points, the α_{xx} component presents the largest error (about 14%), and it shrinks to only 6% with 1500 training structures. The best learning performance is again obtained for the α_{yz} component, for which the prediction error remarkably goes below 2%.

As shown in Fig. 5, SA-GPR reproduces extremely accurately also the *ab initio* Raman spectrum with only 300 training points. Obviously, as reflected by the learning curves, increasing the amount of points gets rid of the errors everywhere, as exemplified in Fig. S4 of the SI, where the machine-learned and *ab initio* spectra

are virtually indistinguishable.

3.2. Paracetamol crystal

We now turn our attention to the first crystalline form of Paracetamol, containing four individual Paracetamol molecules per unit cell, as shown in Fig. 1 (b).

3.2.1. Direct approach Since we are now dealing with a periodic system, any of the aforementioned approaches are applied by first building a crystal supercell. Then, the same procedure discussed in the previous section is applied. A value of $\sigma = 40$ has been selected to build the Gaussian kernel that enters the GPR formalism, and $\eta = 10^{-4}$. The three-dimensional density field was constructed within a box of $12 \times 14 \times 20$ Å^3 and a grid spacing of $d\mathbf{r} = 0.75$ Å¶. For SA-GPR, the λ-SOAP kernels were constructed using the same parameters as before.

The training set is built by considering a random selection of 2500 configurations belonging to a NVT trajectory. A full NVE trajectory is once again considered to test the quality of the predicted Raman spectrum. Learning curves for both regression models are shown in Fig. 6 (solid lines). When comparing the two methods, we always use the same configurations in both cases.

We observe that both the learning capability of GPR and SA-GPR do not reach saturation when increasing the number of training data, going from 81% (respectively 73% with SA-GPR) of error with 25 training points to 17% (respectively 11%) with 2000 of them; Again, making use of a kernel that is

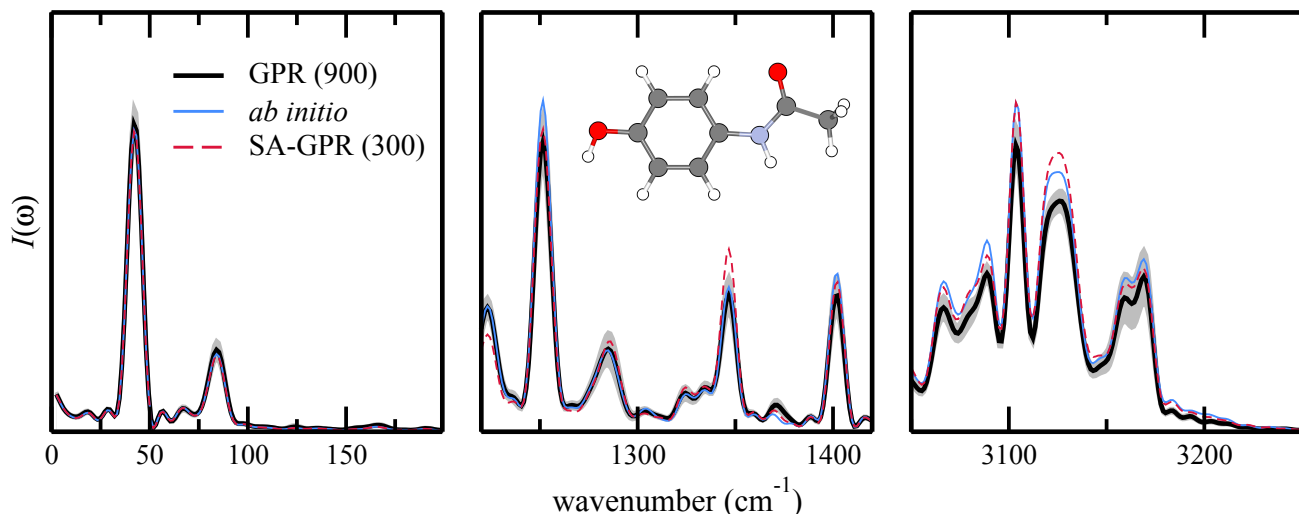¶ No noticeable improvement was observed when using $d\mathbf{r} = 0.5$ Å.

Figure 5. (*black line*) Raman spectrum GPR prediction of the Paracetamol molecule averaged over 16 different training models. Each training model is obtained by a random subselection of 900 configurations over a total of 1100, while the prediction was made on 20000 structures. (*shaded area*) standard deviation of the predicted spectra over the 16 models. (*blue line*) reference *ab initio* Raman spectrum. (*red dotted line*) single SA-GPR prediction using 300 training points.
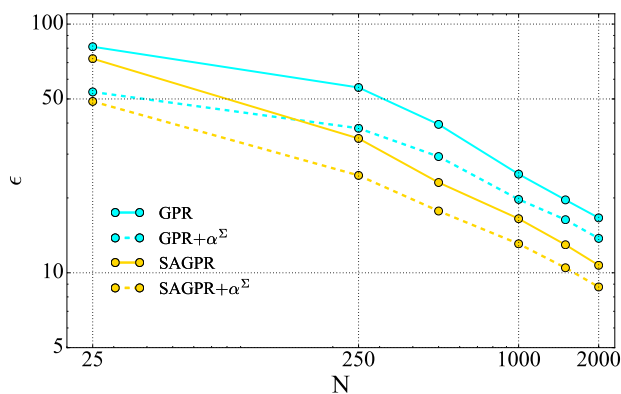

Figure 6. Learning curves for the crystal polarizability tensor on an NVT trajectory, using different approaches. Here the mean error over all components is represented. Including molecular polarizability greatly improves the model, both with GPR and SA-GPR, especially with few training data.

built on a symmetry-adapted comparison between local environments brings a large improvement. Overall, however, for the same amount of training points, the errors are much (typically between two and three times) larger than for the monomer case.

3.2.2. Incorporating molecular polarizability To improve our results, we refined our models by using the predictions of the non-interactig monomers included in the molecular crystal.

Suppose we have a molecular crystal made of N_{mol} molecular units (in the case of Paracetamol I, $N_{\mathrm{mol}} = 4$, while $N_{\mathrm{mol}} = 8$ for Paracetamol II). Since we have already learnt the polarizability tensors of

the individual molecules, we can grasp most of the polarizability tensor of the crystal by summing up the single monomers predictions. Eq. 3 is modified according to

$$\alpha_{\gamma\delta}^{\mathrm{ML}}(\mathcal{A}) = \bar{\alpha}_{\gamma\delta}^{\mathrm{ai,crys}} + \alpha_{\gamma\delta}^{\Sigma}(\mathcal{A}) - \bar{\alpha}_{\gamma\delta}^{\Sigma} + \sum_{j=1}^{N} w_j k(\mathcal{A}, \mathcal{A}_j) \, , (12)$$

where $\boldsymbol{\alpha}^{\Sigma}$ denotes the sum of the molecular polarizability tensors. Specific details about the procedure are explained in Appendix A. An analogous expression to Eq. 12 is obtained for SA-GPR.

Figure 6 shows the advantage of using the molecular baseline for the regression models (dashed lines). The improvement is most noticeable for models based on few training points. Molecular baselining leads to a large decrease in error of about 25% for both GPR and SA-GPR. Upon increasing the number of training points, the difference diminishes, but an improvement remains visible. It is worth noticing that, from 250 training structures on, the direct application of SA-GPR (without any molecular baseline) performs better than the GPR scheme with the baseline. Note that 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, as illustrated in the SI, Fig. S9.

Figure 7 shows the effect of the baselining procedure on the predicted SA-GPR Raman spectrum, when one increases the amount of training points. Several observations can be made. First, just like for the monomer, high frequencies require more training points to be reproduced, be it with a direct prediction or by including molecular polarizability
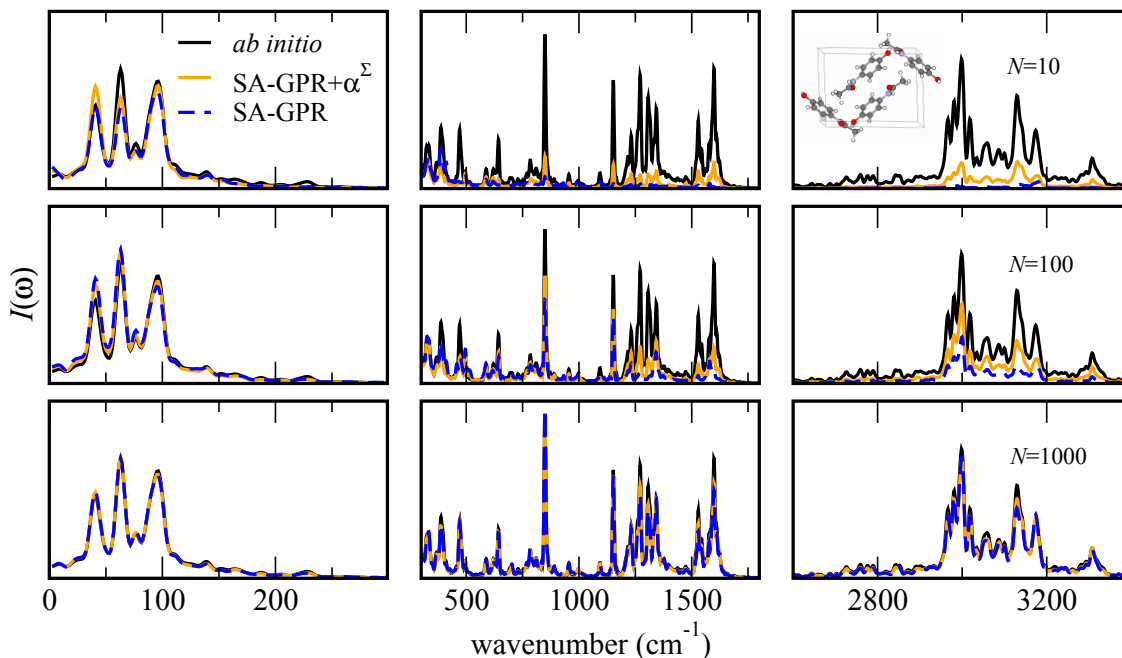
Figure 7. Raman spectrum of paracetamol I coming from an NVE trajectory, using either directly λ-SOAP SA-GPR, or augmenting this description by including molecular polarizabilities. The training was done on NVT structures, the number of which is indicated in each row, while the spectrum was computed over 15000 consecutive NVE configurations.

tensors. Second, including $\boldsymbol{\alpha}^\Sigma$ greatly enhances the spectrum intensity accuracy when few training points are used. This is especially true at high frequencies, where the improved model already gives the right structure of the peaks, albeit not with the right intensity, while the direct learning does not show any peak in this region. Overall, the predicted spectrum is extremely well reproduced when employing enough training points.

Figure 8 shows the (calibrated) estimated error and the average Raman spectrum prediction. In this case, different learning models have been first defined by considering 16 random subselections, each of them made of 80% of the training dataset. Then, for each of these learning models, the polarizabilities of the full NVE trajectory have been predicted and the associated Raman spectra have been computed. We then estimated the standard deviation of these predictions according to the procedure detailed in Sec. 2.4. According to the assumed strong correlations between the random subselections performed, we find that the estimated variance has to be increased by roughly an order of magnitude, i.e., $\nu^2 = 10.9$. One can observe that the excellent agreement between the reference and predicted spectrum at low frequencies is consistent with a negligible estimated error, while larger discrepancies and error bars can be observed in the high-frequency domains.

4. Extrapolation on other polymorphic forms

Within the λ-SOAP formalism, the polarizability of the system is effectively decomposed in local atom-centered contributions that are summed in order to obtain the final predicted value of $\boldsymbol{\alpha}$. This implies that the information is learned at the local scale and, as such, can be transferred across systems that share a similar chemical nature. In the case of paracetamol polymorphs, one can think of predicting the polarizability of the the form II crystal (Fig. 1 (c)) with the model trained on form I only. Since different polymorphic forms are mainly distinguished by the different intermolecular interactions, major difficulties in this extrapolation procedure are expected to be associated with the low-frequency (intermolecular) modes of the molecular crystal.

As detailed in the SI (Fig. S10), we observed a small offset in the time series of some of the polarizability components. In particular, the predicted time series is very well reproduced for α_{xx}, while there is a systematic underestimation of α_{yy} and overestimation (by the same amount) of α_{zz}. Remarkably, because of the error cancellation between the α_{yy} and the α_{zz} predictions, the trace of the tensor is reproduced with about the same level of accuracy as the first crystal form. All the complexity of the extrapolation procedure is thus condensed in predicting the anisotropic part of $\boldsymbol{\alpha}$, highlighting the importance of conformational arrangements of the
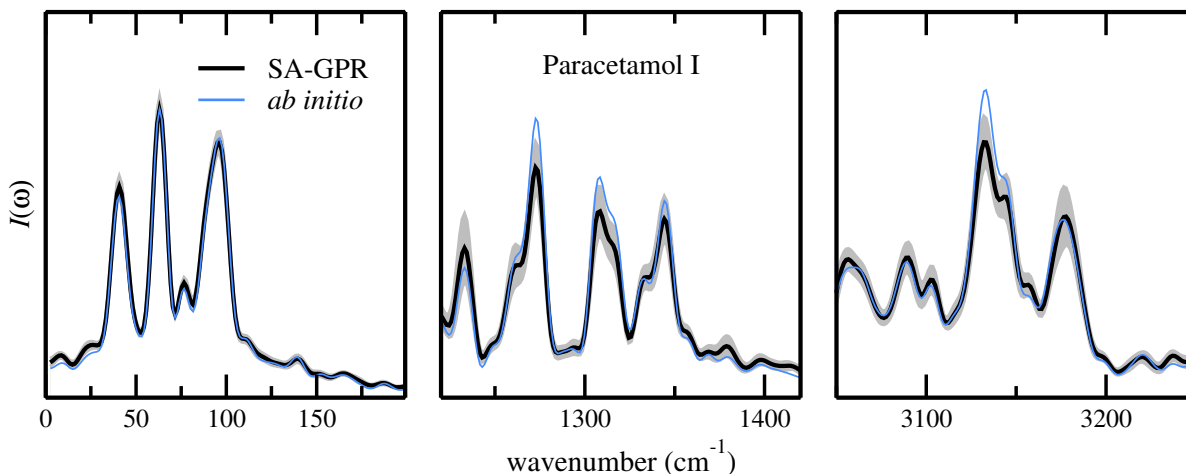
Figure 8. (*black line*) Raman spectrum prediction of paracetamol form-I averaged over 16 different training models. Each training model is obtained by a random subselection of 2000 configurations over a total of 2500. (*shaded area*) standard deviation of the predicted spectra over the 16 models. (*blue line*) reference *ab initio* Raman spectrum.

monomers within the crystal in characterizing the nature of the polymorphic form.

Fig. 9 shows the predicted Raman spectrum, when training SA-GPR on 2000 structures already used for the first polymorphic form, averaged again over 16 subselections. Although a few discrepancies in terms of intensity can be observed, and the standard deviation is overall higher than for the direct prediction of the first polymorph, the general lineshape is excellent, and all the main features of the *ab initio* spectrum are reproduced. We underline the difference in behaviour in contrast to Fig. 8: now, high frequencies are better described and errors are more pronounced at low frequencies. This suggests indeed that the training basis is missing some environments specific to form II, which are necessary to fully reproduce the low frequency features.

5. Conclusions

In this work, we proposed GPR models to predict vibrational Raman spectra, based on learning polarizability and susceptibility tensors obtained from density-functional perturbation theory. As an example, we applied our methodology to predict anharmonic Raman spectra of the Paracetamol molecule and two polymorphs of the Paracetamol crystal. The methodology also works for simpler harmonic Raman intensities. The use of an ensemble of models to estimate the uncertainty in the polarizability tensors allows us to propagate the error estimation from the ML prediction of the polarizability tensors to the vibrational spectra, by generating an ensemble of spectra out of which it is simple to compute frequency-dependent confidence intervals.

We showed that for the molecule, a standard GPR scheme that takes as input a nuclear density representation on a 3D grid works extremely well and enables one to reproduce Raman spectra almost perfectly with a low amount of training points. For the crystal, such a scheme, albeit possible, is more difficult to apply for several reasons: The difficulty to compare crystal structures with different unit cell sizes, the redundancy of information contained in a fixed grid-based representation, and the volumetric increase of grid-points with system size. We have shown that a smoother solution is to use a symmetry-adapted GPR scheme, used here in combination with the λ-SOAP representation [27]. Such a scheme yields accurate predictions for molecules and crystals, due to its capability to better capture the local structural information in a covariant fashion. Moreover, since λ-SOAP is a local descriptor, it is easy to treat larger systems sizes and even transfer models to other polymorphs. We have shown this transferrability by successfully predicting the Raman spectrum of Paracetamol II with a model trained only on Paracetamol form I. This suggests the possibility of predicting Raman spectra of any polymorphic form, as long as a model trained in one of them is available. In addition, for all models presented, we observe a considerable improvement when using previously-trained GPR models for the molecular units as a baseline for the crystal prediction, thus reducing the amount of more costly calculations that must be available for training the model. In a similar manner, it is also straightforward to extend this framework to other ensembles (e.g. NPT) or path-integral molecular dynamics simulations, which include the quantum nature of the nuclei.

The models we presented regard the electronic electric-field response properties, and can be extended
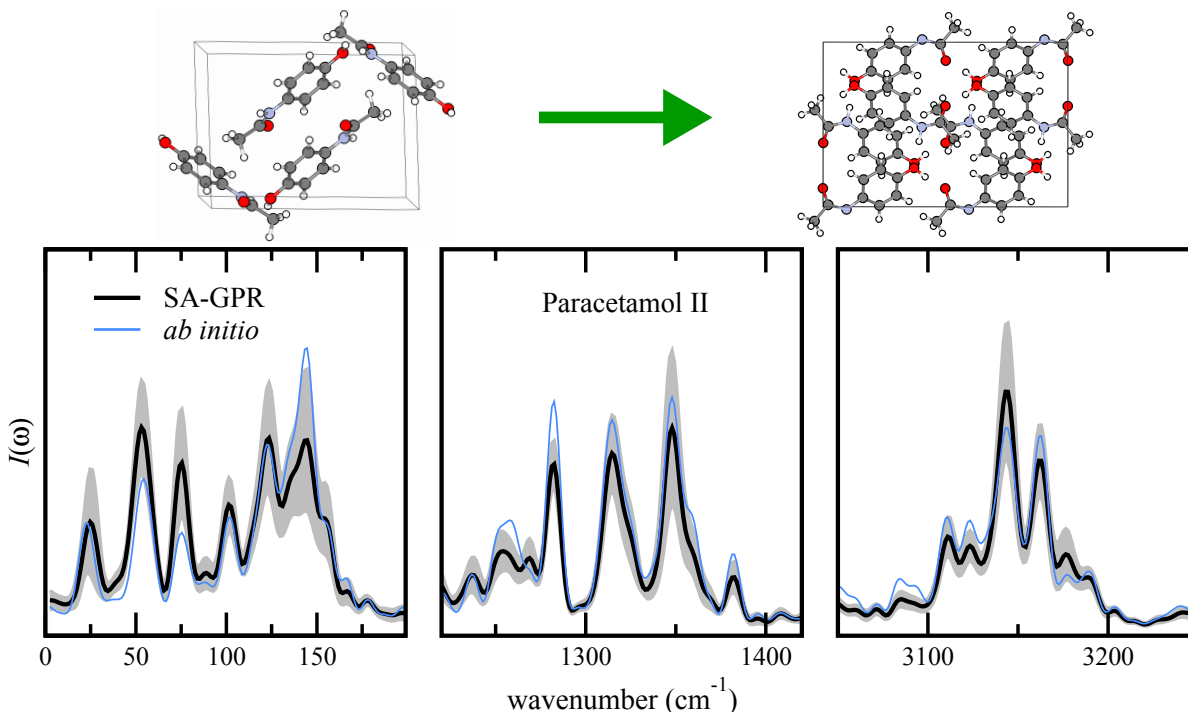
Figure 9. (*black line*) average Raman spectrum prediction of paracetamol form-II associated with the same 16 training models already used for the prediction of paracetamol form-I. (*shaded area*) standard deviation of the predicted spectra over the 16 models. (*blue line*) reference *ab initio* Raman spectrum.

to dipoles and higher-order responses. They can thus be seamlessly combined with empirical potentials or other machine-learned potentials that give access to forces. This presents an alternative route to including the training of such quantities directly into these potentials [40], which can present a higher level of complexity.

Finally, we remark that even though we applied our framework to polarizabilities and Raman spectroscopy, applying it to any other kind of spectroscopy, like infrared, sum-frequency generation, etc., would be straightforward, as long as the electronic-structure data is available.

6. Acknowledgments

Appendix A. Molecular baselining with GPR

When trying to predict the polarizability tensor of the crystal from its individual molecular components within the GPR and AD representation scheme, one should consider that each molecule is oriented in a specific direction that differs from the one we trained

the molecule on. In order for the regression model to recognize the orientation at hand, we first need to find the relationship between the orientation of the molecules in the crystal and the one used in the molecular GPR procedure described in 2.2. The scheme is depicted in Fig. A1.
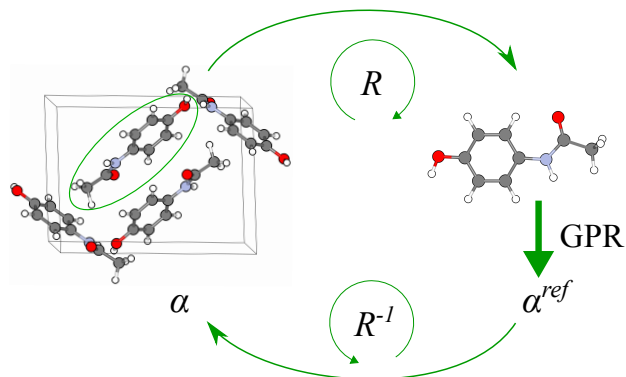

Figure A1. Schematic explanation of the molecular baselining process for the GPR approach. An individual molecule of the crystal is first rotated by a rotation matrix R to match the alignment of a reference structure, on top of which we calculate the polarizability using Eq. 3, and we then rotate back. We repeat this process for every molecule in the unit cell, and sum the resulting polarizability tensors to obtain α^Σ.

Each geometry G_{im} corresponding to the ith molecule of the mth structure is thus rotated by a

rotation matrix R_{im} as

$$G_{im}^{\text{ref}} = R_{im}G_{im} \,, \tag{A.1}$$

where $i = 1 \cdots N_{\text{mol}}$. Finally, once a molecular polarizability α_i^{ref} is predicted we rotate the tensor back to its original orientation inside the crystal, i.e.,

$$\boldsymbol{\alpha}_{i,m}^{\text{mol}} = R_{im}^{-1}\boldsymbol{\alpha}_{i,m}^{\text{ref,mol}} R_{im} \,. \tag{A.2}$$

Having defined the sum of molecular polarizabilities $\boldsymbol{\alpha}^{\Sigma}$ as

$$\boldsymbol{\alpha}^{\Sigma} = \sum_{i=1}^{N_{\text{mol}}} \boldsymbol{\alpha}_i^{\text{mol}} \,, \tag{A.3}$$

we consider the regression target

$$\Delta\alpha_{\gamma\delta} = \alpha_{\gamma\delta}^{\text{crys}} - \bar{\alpha}_{\gamma\delta}^{\text{crys}} - (\alpha_{\gamma\delta}^{\Sigma} - \bar{\alpha}_{\gamma\delta}^{\Sigma}) \,, \tag{A.4}$$

where, once again, the bar denotes the average over the training set. Equation 3 becomes

$$\alpha_{\gamma\delta}^{\text{ML}}(\mathcal{A}) = \bar{\alpha}_{\gamma\delta}^{\text{ai,crys}} + \alpha_{\gamma\delta}^{\Sigma}(\mathcal{A}) - \bar{\alpha}_{\gamma\delta}^{\Sigma} + \sum_{j=1}^{N} w_j k(\mathcal{A}, \mathcal{A}_j) \,. \tag{A.5}$$

An analogous expression is obtained for SA-GPR, but the rotation and alignments previously described do not need to be carried out explicitly since the rotational covariance of the tensor is built in the structure of the method.

[1] Alberi K *et al* 2019 *J. Phys. D* **52** 013001
[2] Ghiringhelli L M, Vybiral J, Levchenko S V, Draxl C and Scheffler M 2015 *Phys. Rev. Lett.* **114**(10) 105503
[3] Ceriotti M 2019 *arXiv e-prints* arXiv:1902.05158 (*Preprint* 1902.05158)
[4] Butler K T, Davies D W, Cartwright H, Isayev O and Walsh A 2018 *Nature* **559** 547–555
[5] Behler J 2016 *J. Chem. Phys.* **145** 170901
[6] Bartók A P, Payne M C, Kondor R and Csányi G 2010 *Phys. Rev. Lett.* **104**(13) 136403
[7] Jain A, Ong S P, Hautier G, Chen W, Richards W D, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G and Persson K A 2013 *APL Mater.* **1** 011002
[8] Calderon C E, Plata J J, Toher C, Oses C, Levy O, Fornari M, Natan A, Mehl M J, Hart G, Nardelli M B and Curtarolo S 2015 *Comput. Mater. Sci.* **108** 233 – 238
[9] Ward L and Wolverton C 2017 *Curr. Opin. Solid State Mater. Sci.* **21** 167–176
[10] Li Z, Kermode J R and De Vita A 2015 *Phys. Rev. Lett.* **114**(9) 096405
[11] Glielmo A, Sollich P and De Vita A 2017 *Phys. Rev. B* **95** 214302
[12] Glielmo A, Zeni C and De Vita A 2018 *Phys. Rev. B* **97**(18) 184307
[13] Liang C, Tocci G, Wilkins D M, Grisafi A, Roke S and Ceriotti M 2017 *Phys. Rev. B* **96**(4) 041407
[14] Grisafi A, Wilkins D M, Csányi G and Ceriotti M 2018 *Phys. Rev. Lett.* **120**(3) 036002
[15] 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
[16] Christensen A S, Faber F A and von Lilienfeld O A 2019 *J. Chem. Phys.* **150** 064105
[17] Agrawal K V, Shimizu S, Drahushuk L W, Kilcoyne D and Strano M S 2016 *Nature Nanotech.* **12** 267–273
[18] Heiner Z, Zeise I, Elbaum R and Kneipp J 2018 *J. Biophotonics* **11** e201700164
[19] Monserrat B, Drummond N D, Dalladay-Simpson P, Howie R T, López Ríos P, Gregoryanz E, Pickard C J and Needs R J 2018 *Phys. Rev. Lett.* **120**(25) 255701
[20] Shang H, Raimbault N, Rinke P, Scheffler M, Rossi M and Carbogno C 2018 *New J. Phys.* **20** 073040
[21] Raimbault N, Athavale V and Rossi M 2019 *Phys. Rev. Mater.* **3**(5) 053605
[22] Gerratt J and Mills I M 1968 *J. Chem. Phys.* **49** 1719–1729
[23] Gonze X and Lee C 1997 *Phys. Rev. B* **55**(16) 10355–10368
[24] Putrino A, Sebastiani D and Parrinello M 2000 *J. Chem. Phys.* **113** 7102–7109
[25] Baroni S, de Gironcoli S, Corso A D and Giannozzi P 2001 *Rev. Mod. Phys.* **73** 515–562
[26] Bereau T, Andrienko D and von Lilienfeld O A 2015 *J. Chem. Theory Comput.* **11** 3225–3233
[27] Grisafi A, Wilkins D M, Csányi G and Ceriotti M 2018 *Phys. Rev. Lett.* **120**(3) 036002
[28] Putrino A and Parrinello M 2002 *Phys. Rev. Lett.* **88**(17) 176401
[29] Pagliai M, Cavazzoni C, Cardini G, Erbacci G, Parrinello M and Schettino V 2008 *J. Chem. Phys.* **128** 224514 URL https://doi.org/10.1063/1.2936988
[30] Mukamel S 1995 *Principles of nonlinear optical spectroscopy* Oxford series in optical and imaging sciences (Oxford University Press) ISBN 9780195092783
[31] Vu K, Snyder J C, Li L, Rupp M, Chen B F, Khelif T, Mller K R and Burke K 2015 *Int. J. Quantum Chem.* **115** 1115–1128
[32] Zauleck J P P and de Vivie-Riedle R 2018 *J. Chem. Theory Comput.* **14** 55–62
[33] Kabsch W 1976 *Acta Crystallogr. Sect. A* **32** 922–923
[34] Kabsch W 1978 *Acta Crystallogr. Sect. A* **34** 827–828
[35] Bartók A P, Kondor R and Csányi G 2013 *Phys. Rev. B* **87**(18) 184115
[36] Musil F, Willatt M J, Langovoy M A and Ceriotti M 2019 *J. Chem. Theory Comput.* **15** 906–915
[37] Blum V, Gehrke R, Hanke F, Havu P, Havu V, Ren X, Reuter K and Scheffler M 2009 *Comput. Phys. Commun.* **180** 2175–2196
[38] Tkatchenko A, DiStasio R A, Car R and Scheffler M 2012 *Phys. Rev. Lett.* **108**(23) 236402
[39] Ambrosetti A, Reilly A M, DiStasio R A and Tkatchenko A 2014 *J. Chem. Phys.* **140** 18A508
[40] Gastegger M, Behler J and Marquetand P 2017 *Chem. Sci.* **8**(10) 6924–6935