# First-principles NMR of oxide glasses boosted by machine learning

Thibault Charpentier

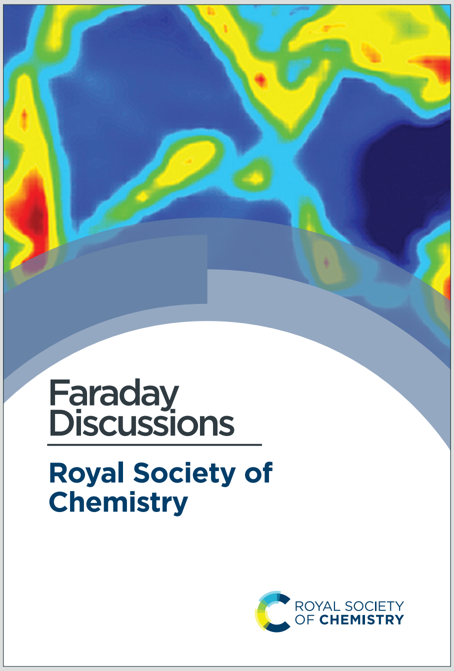
## To cite this version:

Thibault Charpentier. First-principles NMR of oxide glasses boosted by machine learning. Faraday Discussions, 2024, 10.1039/d4fd00129j . cea-04626737

**HAL Id: cea-04626737**
**https://cea.hal.science/cea-04626737**

Submitted on 27 Jun 2024

Faraday Discussions



This article can be cited before page numbers have been issued, to do this please use: T. Charpentier, *Faraday Discuss.*, 2024, DOI: 10.1039/D4FD00129J.

ROYAL SOCIETY OF CHEMISTRY

www.rsc.org/faraday_d

# First-principles NMR of oxide glasses boosted by machine learning

View Article Online
DOI: 10.1039/D4FD00129J

Thibault Charpentier[1]

Université Paris-Saclay, CEA, CNRS, NIMBE, 91191 Gif-sur-Yvette cedex, France.

## Abstract

Solid-State NMR has established itself as a cutting-egde spectroscopy for elucidating the structure of oxide glasses thanks to several decades of methodological and instrumental progresses. First-principles calculations of NMR properties combined with molecular dynamics (MD) simulations provides a powerful complementing approach for the interpretation of the NMR data although they still suffer from limitations in terms of size, time and high consumption of computational resources. We address this challenge by developing a machine-learning framework to boost predictive modelling of NMR spectra. We use kernel ridge regression techniques (least-square support vector regression and linear ridge regression) combined with the smooth overlap of atomic position (SOAP) atom-centered descriptors to efficiently predict NMR interactions: the isotropic magnetic shielding and the electric field gradient (EFG) tensor. As illustrated in this work, this approach enables the simulation of MAS and MQMAS NMR spectra of very large models (more than 10000 atoms) and an efficient averaging of NMR properties over MD trajectories of nanoseconds for incorporating finite temperature effects, at computational cost of classical MD simulation. We illustrate these advances on sodium silicate glasses ($SiO_2$-$Na_2O$). NMR parameters (isotropic chemical shift and electric field gradient) could be predicted with an accuracy of 1 to 2% in terms of the total span of the NMR parameter values. To include vibrational effects, an approach is proposed by scaling the EFG tensor in NMR simulations with a factor obtained from the time auto-correlation functions computed on MD trajectories.

## Introduction

Glass are ubiquitous materials in our modern life because of their low-cost but good performances for a high diversity of usage.[1–5] Glass indeed offers an infinite combination of compositions (but within a glass formability window) to match the properties of interest. The importance of glass has been recognized in 2022 by UNESCO declared as the International Year of the Glass.[2] The fundamental challenge in glass science remains the establishment and the knowledge of the relationships between the composition-structure to the properties for the design of new formulations for specific applications. To this aim, molecular dynamics (MD) is the reference technique for building atomistic structural models of glasses and calculating the properties of interest.[6–9] Spectroscopic data are however needed to assess the MD structural models which have been mostly provided by neutron and X-Ray diffusion experiments for decades.[10–12] In this regard and in the context of oxide glasses, NMR remains underexploited and has been generally limited to its ability

---

[1] Université Paris-Saclay, CEA, CNRS, NIMBE, 91191 Gif-sur-Yvette cedex, France. thibault.charpentier@cea.fr

[2] https://www.iyog2022.org/

View Article Online
DOI: 10.1039/D4FD00129J

of providing quantitative information on the structural motifs building the glass network (such as $SiO_4$, $AlO_4$, $BO_3$, $BO_4$ ) featuring the so-called short-range order (SRO).

NMR has clearly proven over the two last decades to be a key spectroscopy for deciphering glass structure at various atomic length scale order (from short to intermediate range order).[13–22] Beside the direction of development of advanced radio-frequency pulse sequences for manipulating NMR interactions with an extreme high degree of accuracy[19, 21], a second direction was proposed with the introduction of accurate and robust DFT computation methodologies, based on the popular DFT-GIPAW method.[23–25] These two directions are of course complementary but both address the essential difficulty encountered in studying amorphous materials such as oxide glasses: the chemical and geometrical disorder introduces broadening of the lines resulting in a strong overlap between NMR peaks which can persist despite usage of sophisticated 2D high-resolution techniques. This spectral broadening is engendered by the NMR parameter distribution, a salient feature of glassy materials that cannot be ignored when fitting experimental data.[16, 26–31] With the help of relationships established either from known crystalline samples or DFT computations, NMR parameter distributions can be inverted to a distribution of a geometrical parameter such as inter-atomic distances, bond lengths or bond-angle distribution (BAD).[31–34] However such approaches are limited to simple glass compositions and are difficult to be extended when several parameters interplay. A direct link between a 3D structural model and the NMR spectrum that can provide such information is therefore necessary.

The continuous increase of computational resources combined with the efficiency and accuracy the DFT-GIPAW method[35–37] has enabled its fruitful combination with MD so that a *first-principles NMR* approach has emerged as a new tool for investigating glasses.[38–41] Within this framework, one can validate the interpretation of experimental data by prediction of NMR fingerprints of atomic species[42, 43] or assess the quality of MD structural models with a direct comparison with NMR data.[44–46] According to our own experience, modern computational resources enable GIPAW-DFT calculations with MD models up to ~800 atoms to be performed. DFT inherently limits applicabilition to representative models of glass of several thousands of atoms (a standard size is 5000-10000 atoms for MD studies). In addition, investigations of dynamical effects (such as vibrations or diffusion of atoms) on NMR[47–50] necessitates long MD trajectories of the order of ns to be considered. Consequently, the recent emergence of machine learning (ML) methodologies in atomistic modelling[51] offers appealing perspectives for accessing larger system size and time scale. In solid-state NMR, kernel ridge regression (KRR) techniques were applied to predict NMR shifts in organic molecular solids by Paruzzo et al.[52], and in organic molecules by Rupp et al.[53] The first application to silicate glasses was provided by Cuny et al.[54] using a neural network and by Chaker et al.[55] in sodium aluminosilicate glasses with linear ridge regression (LRR). Gaumard et al.[56] studied the performances of different kernel regression in zeolites. Recently, positions of cesium in clays could be refined by Ohkubo et al.[57] by using predicted NMR chemical shifts predicted by LRR. In the context of amorphous molecular solids, Cordova et al.[58, 59] combined MD and KRR-SOAP predicted shifts (ML-Shift),[52] to match structural models to experimental MAS NMR spectra. In the context of oxide glasses where quadrupolar nuclei are predominant, prediction of the EFG tensor is needed.

For atomistic modelling by ML, the so-called atom-centered descriptors (ACD) play a central role. In the ML context, they were introduced in the seminal work of Behler et al.[60] and Bartòk et al.[61] The variety of descriptors that have been proposed in literature is too vast to be covered here (see for example in Refs[62–64]). ACD must provide a faithful and symmetry adapted representation of the local environment of a central atom (chemically and geometrically): they must be invariant to translations and permutations, and equivariant to rotations and inversions in regards of nature of the property of interest. Indeed, for prediction of a scalar property (such as the isotropic magnetic shielding, a charge or an energy), they must be invariant.[61] In contrast, for tensorial properties (in our case the electric field gradient (EFG) tenso), they must respect the rotational properties of second rank tensors, as was first discussed by Grifasi et al.[65] In this work, we have chosen the popular smooth overlap of atomic positions (SOAP) descriptors which have extensively studied[61, 66–69] and shown to have excellent performances on NMR shifts prediction.[52, 55, 56]

View Article Online
DOI: 10.1039/D4FD00129J

We present a fully-integrated methodology based on ML kernel tools (linear and kernel ridge regression, least-square support vector regression, kernel density estimation, dimensional reduction with incomplete Cholesky decomposition of the kernel Gram matrix) for the simulation of NMR spectra of structural models containing up to several thousands of atoms (calculations were performed on a standard workstation). Our strategy for building a database for ensuring a good transferability of the ML predictions (between various glass compositions and MD temperatures) is presented, as well as ideas for application to the study of the impact of local mobilities (vibrations, or diffusion of atoms) on the NMR spectrum. This opens appealing perspective for the investigation of ionic conduction in glasses by NMR. Sodium-silicate glasses $Si_2O$-$Na_2O$ have been chosen as a representative model of oxide glasses: they contains both I=1/2 ($^{29}Si$) and quadrupolar nuclei ($^{23}Na$ and $^{17}O$); one with a high local mobility ($Na^+$). Experimental data were taken from previous studies[27, 29].

## Theory and Methods

### SOAP descriptors for representing the local environment

To represent the local environment of an atom, the SOAP descriptors were chosen.[61, 66] They are constructed from the expansion of the smoothed atomic density $\rho_i(r)$ of a central atom *i* as:

$$\rho_i(r) = \sum_{j \in N_i} g_\sigma(r - r_{ij}) f_c(r_{ij}) = \sum_{\mu,n,l,m} c^{i,\mu}_{n,l,m} Y_{lm}(\grave{r}) R_{nl}(r) \quad (1)$$

$N_i$ is the set of atoms *j* that are in the neighbourhood within a cutoff radius $r_{cut}$. $g_\sigma$ is a (3d) Gaussian function of width $\sigma$ and $f_c(r) = (1/2)\{cos(\pi\, r/r_{cut}) + 1\}$ is the function used to smooth the density at the cutoff radius. $Y_{lm}(\grave{r})$ are spherical harmonics ($\grave{r} = (\theta,\phi)$ are the polar and azimuthal angles) and $R_{nl}(r)$ are radial functions. $\mu$ is an index that runs on the various atomic species (here, Na, Si and O). In this work, we employ real spherical harmonics (RSH) $Y_{lm}(\grave{r})$ and spherical Bessel functions $R_{nl}(r) = j_l(\alpha_{nl} r/r_{cut})$. $\alpha_{nl}$ is the n^th^ root of the Bessel function $j_l()$. This choice ensures that the $R_{nl}(r)$ form an orthonormal set on the segment $[0,r_{cut}]$ (see section S4 of the ESI). With eqn (1), the local environment $\rho(r)$ of the central atom is represented by the set of parameters $c^{\mu}_{n,l,m}$. For computations, the expansion eqn (1) is truncated to values $n \in [0,N_{max}]$ and $l \in [0,L_{max}]$. Thus, $(r_{cut},N_{max},L_{max})$ which are called hyper-parameters, have to be optimized during the training of the ML algorithm. The $c^{\mu}_{n,l,m}$ are invariant to permutations of the neighbouring atoms and to translations, and they behave like a $l$-rank tensor under rotations. They represent the local environment of a central atom X in terms of two body interactions. Indeed, for $Na_2O$-$SiO_2$ glasses, the expansion on the index $\mu$ yields X-Si, X-O and X-Na terms.

For the prediction of the *scalar* isotropic magnetic shielding value $\sigma_{iso}$ (or equivalently, the isotropic chemical shift), we need rotation-invariant descriptors. The symmetry adapted descriptors are therefore reduced to $c^{\mu}_{n,0,0}$ values, which severally limits the number of descriptors for an environment that can contains around ten atoms (in practice $N_{max}$ ranges from 2 to 12). To overcome this limitations in their seminal work, Bartòk et al.[61] introduced the *powerspectrum*, a set of descriptors that combine the $c^{\mu}_{n,l,m}$ into a sum of rotationally invariant products:

$$p_{n_1,n_2,l}^{\mu_1,\mu_2} = \sum_m c_{n_1,l,+m}^{\mu_1} c_{n_2,l,-m}^{\mu_2} \qquad (2)$$

View Article Online
DOI: 10.1039/D4FD00129J

The powerspectrum contains three body-terms which accounts for the distribution of angles around the central atom (i.e., Si-X-Si, Si-X-O, Si-X-Na, O-X-O, O-X-Na and Na-O-Na) as was explicitly shown in Ref.[67] For tensorial properties of rank $\lambda$(i.e., using the familiar notation in NMR $T_{\lambda,m}$), eqn (2) can be easily generalized to become equivariant:

$$\left(q_{n_1,n_2,l_1,l_2}^{\mu_1,\mu_2}\right)_{\lambda,m} = \sum_{m_1,m_2} \left(\lambda,m\middle|l_1 m_1 l_2 m_2\right) c_{n_1,l_1,m_1}^{\mu_1} c_{n_2,l_2,m_2}^{\mu_2} \qquad (3)$$

where $\left(\lambda,m\middle|l_1 m_1 l_2 m_2\right)$are the Clebsh-Gordan coefficients. These descriptors, referred to as λ-SOAP, were introduced by Grifasi et al.[65], generalizing thereby the SOAP powerspectrum to represent a tensorial quantity. One can easily recognise that eqn (2) is a special case of eqn (3) with the choice λ=0 (then $l_1=l_2$) In the case of EFG or chemical shift anisotropy (CSA, symmetric part) tensor, λ=2. The tensor (we consider only the EFG is this work) is then represented by 5-dimensional vector denoted $V_m$ in its spherical form (see section S6 of the ESI), which is the form which used in our LRR algorithms. Note that the dimension of λ-SOAP descriptors can be very high (in our case, from $10^2$ to $10^4$), thus necessitating a large dataset for training the LRR algorithm.

**Least-square support vector and linear ridge regression**

The kernel ridge regression (KRR) is a generalization of the popular linear ridge regression (LRR) by introducing non-linearities through a kernel function which can be seen as a measure of the similarity between two environments, here denoted$\chi$. For the isotropic magnetic shielding $\sigma_{iso}(\chi)$, the prediction is performed by a linear combination of similarities between the new environment $\chi_{new}$ and the *training* ones $\chi_i$ as follows:

$$\sigma_{iso}(\chi_{new}) = \sum_{i \in train} K(\chi_{new},\chi_i)\alpha_i \qquad (4)$$

where $K(\chi_i,\chi_j)$ is the kernel function. In LRR, the kernel is the scalar product $K(\chi_i,\chi_j) = \chi_i \cdot \chi_j$. A standard choice is the Gaussian function (in this case KRR shares similarities with Gaussian Processes[70])$K(\chi_i,\chi_j) = exp\{-\theta(\chi_i - \chi_j)^2\}$where $\theta$ is a (hyper-)parameter that needs to be optimized. Note that considering all points in the dataset, eqn (4) can be rewritten in a matrix as $\delta = K\alpha$. Consequently, the determination of the regression parameters $\alpha$(i.e., the training phase) are obtained from $\alpha = (K + \epsilon I)^{-1}\delta$. $\epsilon$is the ridge-parameter that controls the norm of the regression parameters $\alpha$in order to prevent the ML predictor to overfit the training data. $\epsilon$has therefore to be optimized from a second set of independent data (the validation set) in order to ensure a good transferability of the ML prediction to new data (the testing set).

Because of the high dimensionality of the training set, the resolution of the linear system eqn (4) can be cumbersome. The idea of the least-square support vector regression (LSSVR)[71] is to use a reduced set of representative data, denoted $\xi$ , generally referred to as the inducing points or the landmark points. Restricted to this small set, the inversion of the kernel matrix $(K + \epsilon I)^{-1}$is then tractable. From a NMR perspective, it can be easily understood that many environments in the database are similar (and thus their NMR parameters are also very close), so that fewer points are necessary to support the linear regression eqn (4) (*the support vectors*). Mathematically, the kernel matrix is then approximated by $K_{\chi,\chi} \approx K_{\chi,\xi} K_{\xi,\xi}^{-1} K_{\xi,\chi}$ with

View Article Online
DOI: 10.1039/D4FD00129J

$N_\xi \approx 10^3$ whereas $N_\chi \approx 10^5$in our study. This is the Nyström approximation. Resolution of eqn (4) then only requires the diagonalization of$K_{\xi,\xi}$, as detailed in Ref.[71], but the determination of the regression parameters $\alpha$ proceeds so as to accounts for all samples of the training set. To determine the inducing points $\xi_i$ from the training set, we found that the incomplete cholesky decomposition (ICD) of the kernel matrix $K_{\chi,\chi}$[72] was very efficient and informative as will be discussed below.

Concerning the EFG tensor, it was predicted in spherical form using the λ-SOAP descriptors eqn (3) with a linear ridge regression (LRR):

$$V_m = \sum_{\mu_1,\mu_2,n_1,n_2,l_1,l_2} \alpha^{\mu_1,\mu_2}_{n_1,n_2,l_1,l_2} \left( q^{\mu_1,\mu_2}_{n_1,n_2,l_1,l_2} \right)_{2,m} \quad (5)$$

Note that the summation over $l_1$ and $l_2$ is limited to values such that $\left(\lambda = 2,m \middle| l_1 m_1 l_2 m_2\right) \neq 0$. In our work with $L_{max}$=4, from the 25 $(l_1,l_2)$pairs, only 16 contributes in eqn (5).

## NMR simulations with kernel density estimation

From DFT-GIPAW computations or ML predictions, the NMR parameters $x = (\delta_{iso},C_Q,\eta_Q)$ for each atom (here $^{29}$Si, $^{23}$Na, and $^{17}$O) are obtained. A simple approach for simulating the NMR spectrum consists in summing the individual NMR spectra for each atom. This would be however quite ineffective and time-consuming. As was discussed in earlier works,[41, 73] a better strategy for glasses consists in first reconstructing the NMR parameter distribution. The latter can of course be multi-modal if various species are present; in this case, a clustering algorithm can advantageously help to identify the speciations, an option which is left for future studies. Here a kernel density estimation (KDE)[74] approach is adopted for building the NMR parameter distribution $p(\delta_{iso},C_Q,\eta_Q)$on a 3d grid of pre-computed NMR spectra. In the case of silicon-29, quadrupolar parameters can be replaced by CSA values for an accurate modelling but here a simple 1d approach is adopted so that only $p(\delta_{iso})$is considered (for the reconstruction of anisotropic/isotropic 2D correlation with KDE, see for example Ref[41]). In this work, we focus on MQMAS reconstruction of the two quadrupolar nuclei of interest, $^{17}$O and $^{23}$Na.

In the KDE formalism, the value of the distribution is estimated for each grid point $x_g$ from the database points $x_i$ as:

$$p(x_g) = \frac{1}{N_i} \sum K_\Sigma (x_g - x_i) \qquad (6)$$

where the kernel $K_\Sigma$ is chosen as a Gaussian distribution which shape and width is controlled by the covariance matrix $\Sigma$(also referred to as the bandwidth matrix) which calculated from the (training) $x_i$ points. This matrix is essential to account for the correlation effects that exists between the NMR parameters, as was observed in oxygen-17 MQMAS NMR of silicate glasses.[27, 32, 33] To lower the computational cost of eqn (6), numerous strategies exist to reduce the number of distance evaluations, such as approximate nearest neighbour search based on graph.[75]

From $p(x_g)$ and precomputed spectra on the 3d grid, the simulation of the MAS and MQMAS spectra can then be reduced to simple matrix multiplications. In the case of a MQMAS simulation, equations were given in Ref.[41] The pre-computed spectra can optionally account for the finite pulse width effects (intensity and lineshape distortion) and finite spinning rate. Denoting $I(\nu;\delta_{iso},C_Q,\eta_Q)$ the precomputed spectra on the grid and neglecting offset effects so that $I(\nu;\delta_{iso},C_Q,\eta_Q) \approx I(\nu - \delta_{iso};0,C_Q,\eta_Q)$, only a 2d grid of $(C_Q,\eta_Q)$

parameters is necessary. Typical grids have the following resolution (boundary values depend on the nucleus): $C_Q$0.1 MHz; $\eta_Q$ 0.05, and $\nu$(or equivalently $\delta_{iso}$) 0.2 ppm.

View Article Online
DOI: 10.1039/D4FD00129J

This KDE approach underlines that in the specific case of glasses, the primary goal of the ML prediction is to reconstruct the NMR parameter distribution $p(\delta_{iso},C_Q,\eta_Q)$ from the set of environments generated with the MD model(s). Thus in addition to the quantification of prediction error for each point of the dataset, it is important to compare the reconstructed NMR parameter distribution from predicted ML values (and simulated NMR spectra) with the one computed from the DFT values. Accordingly, many (small) structural model are used to provide a set of sufficiently densely distributed point to estimate $p(\delta_{iso},C_Q,\eta_Q)$. The alternative of using larger models offered by ML prediction enables to investigate size effects (such a spurious correlation effect induced by the periodic boundary conditions in MD simulations).

**Optimisation of the Hyper-parameters: k-Fold Cross-Validation**

Beside the determination of the regression parameters in eqns (4) and (5), SOAP, kernel and ridge parameters need to be optimized. They are generally referred to as hyper-parameters (with respect to the regression parameters $\alpha$in eqn (4)). We use a k-Fold cross-validation approach which consists in splitting the database into k subsets (in practice, k=5): the testing set, the validation set and the reminder subsets are in the training set. In LSSVR and LRR, the regression parameters are determined with the training set and the ridge regression parameter is optimized by minimizing the error on the validation set. Note that the inducing points are restricted to be in the training set. The reported error is computed with the testing set. In order to have each point tested, the procedure is repeated k-times by shifting the k-folds. With an initial random shuffling of the database points, the whole procedure can be repeated so that each point can be a testing points more than once, and it provides a robust value of the error standard deviation value. The (hyper-)parameters of the descriptors and kernel width are optimized in an outer loop. Two errors are reported: the mean absolute error (MAE) and root mean square error (RMSE). The latter is sensitive to outliers so that both values a complementary.

## Design of the NMR database

Structural models of $Na_2O$-$SiO_2$ glasses with %mol $Na_2O$ ranging from 10% to 50% (by step of 10%) were generated with classical MD simulations, as detailed in sections S1 and S2 of the ESI. For each composition, 20 independent small models of 300 atoms and 2 supplementary models of 600 atoms were generated for testing the transferability of ML algorithms to larger models when trained on small models. To investigate increase the geometrical diversity, structures were extracted at 300K, 1000K, (both forming the reference set), 1500K and 2000K. Long MD trajectories (up to 1 ns) were simulated to study the impact of vibrations on the NMR spectra. Additionally, for each composition a model of 14400 atoms was generated to demonstrate the applicability of ml-NMR simulations to very large system in short CPU times (here, ~2s).

Part of database was recently used for training a machine learning potential (MLP).[76] Two compositions, denoted NS22.5 and NS43.1 (22.5 and 43.1 %mol $Na_2O$, respectively) were chosen from this work ( ~700 atoms, see Table S2 of the ESI) for the comparison of the ml-NMR spectra with experimental data ($^{29}Si$, $^{23}Na$ and $^{17}O$).[27] To this end, ab-initio MD simulations at 300K were performed (with CP2K[77]) for incorporating effects of vibrations at 300K in the NMR simulations through the computation. About 41 structures were extracted every 50 fs to built a NMR database from the first 2ps of the trajectory. In the last 10ps, 8 structures were extracted every 500 fs to check the accuracy of ML prediction.

View Article Online
DOI: 10.1039/D4FD00129J

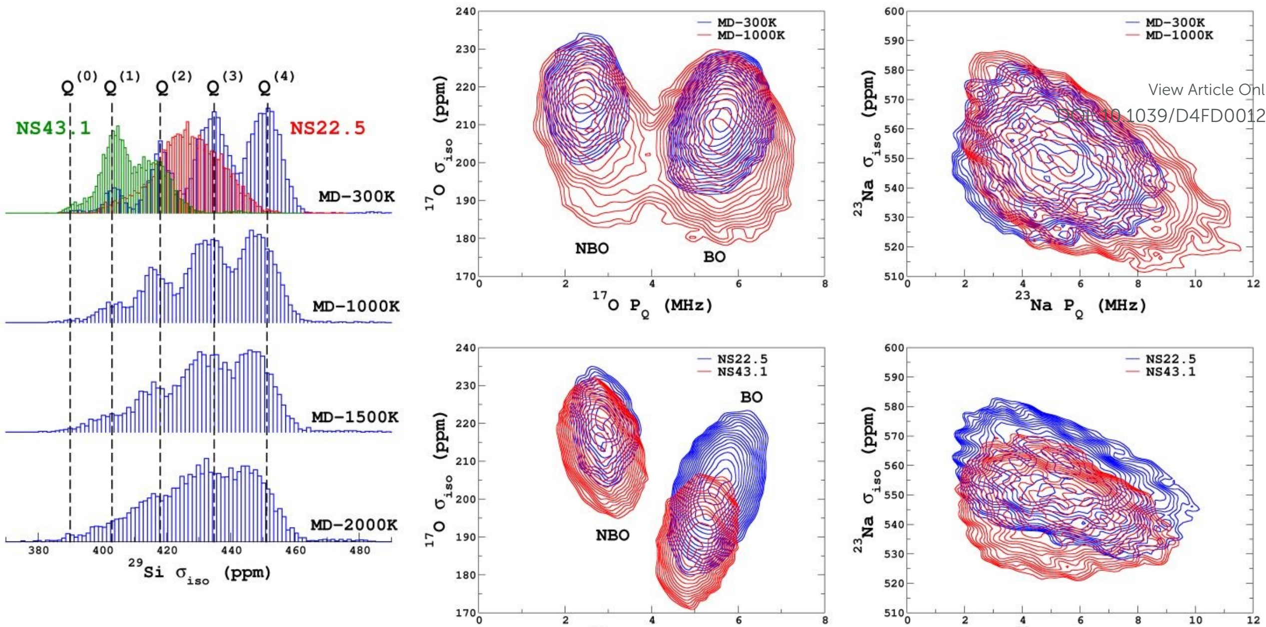
NMR properties, the magnetic shielding and electric field gradient tensors, were calculated with the DFT-GIPAW method[23] as implemented in VASP (version 5.3.x).[35] For referencing the NMR-GIPAW outputs $\sigma_{iso}$ to isotropic chemical shifts values $\delta_{iso}$ of $SiO_2$ (cristobalite and quartz), $Na_2SiO_3$ and α- and β-$Na_2Si_2O_5$ were used but in a consistent approach with the MD models: for each system, a supercell of ~ 300 atoms was built, its geometry (atomic positions and unit cell parameters) was optimized with CP2K and the NMR-GIPAW values computed with VASP. Details are given in section S.3 of the ESI. Notably, the calibration parameters (α and $\sigma_{REF}$) of the linear regression $\delta_{iso} = -\alpha(\sigma_{iso} - \sigma_{REF})$ are given in Table S4 of the ESI.

## Results and Discussion

### Examination of the NMR database

We propose to examine the diversity of environments in the database be visualizing their respective NMR parameter distribution. For the quadrupolar nuclei, the NMR parameters were efficiently represented with the 2D distribution $p(\delta_{iso}, P_Q)$ using the quadrupolar product $P_Q^2 = C_Q^2(1+\eta^2/3)$. Indeed, $P_Q$ (in contrast to $C_Q$and $\eta$) is invariant to both the orientation and the ordering principal values of the EFG tensor ($V_{XX}$,$V_{YY}$,$V_{ZZ}$). This parameter should be therefore be preferred to represent the EFG tensor strength than $C_Q$in glasses and during MD trajectories. As shown in Figure 1, the broadening effect with increasing the MD temperature is clearly seen on all data (resulting from a higher geometrical diversity). The aiMD data shows narrower distributions with more pronounced effect of correlation between the NMR parameters for oxygen-17 (note the opposite sign of the linear correlation between NBO and BO). As MD and aiMD distributions do not overlap, it is therefore expected that MD data cannot predict well aiMD data. This also points out that, even if less accurate, classical MD produces more disordered structures thus more suitable for building database with high transferability (specialized database can be of course generated depending on the aim of the ML prediction). All datasets of the NMR database are given in Table S3 in the ESI.

View Article Online
DOI: 10.1039/D4FD00129J



$^{29}$Si $^{17}$O $^{23}$Na

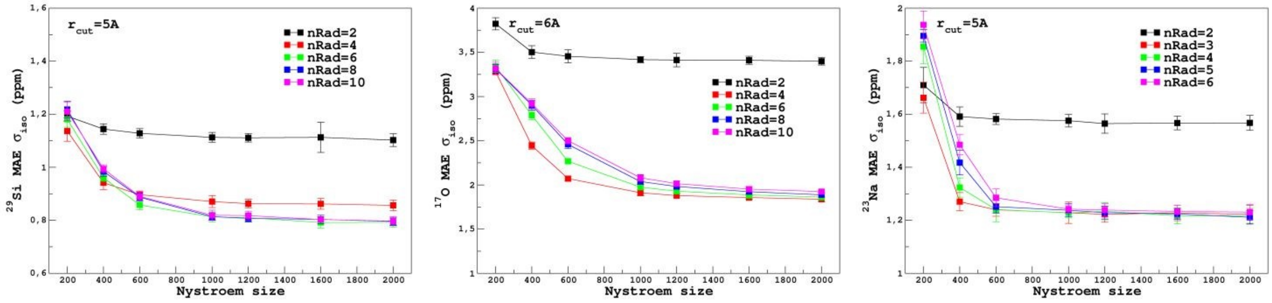
*Figure 1. $^{29}$Si, $^{17}$O and $^{23}$Na NMR parameter distributions of datasets from the database. MD-xxx (xxx=300K, 1000K, 1500K and 2000K) are classical MD models, NS22.5 and NS43.1 are MLP models extracted from aiMD data at 300K. For $^{17}$O and $^{23}$Na, $p(\delta_{iso}, P_Q)$ was calculated using KDE eqn (6).*

## Learning the isotropic magnetic shielding

For the prediction of $\sigma_{iso}$ (or equivalently $\delta_{iso}$), we performed a systematic grid search on the SOAP hyper-parameters as follows: $\sigma_{SOAP}$in [0.2, 1] Å (step 0.2Å ); $r_{cut}$=3, 4, 5 and 6Å; $N_{max}$from 2 to 12, $L_{max}$ was fixed to 4. Details on the computation and optimization of the SOAP descriptors for the MD-300K and MD-1000K datasets are given in sections S4 and S5, respectively, of the ESI. For each $(\sigma_{SOAP}, r_{cut}, N_{max})$ values, the Nyström size (denoted $N_\xi$) was incremented until the MAE has converged (denoted $N_\xi^{opt}$). $N_\xi$ strongly depends on the variety of environments present in the training set and $N_\xi^{opt}$ can be therefore considered as a measure of the diversity of the database. An optimal value of $\sigma_{SOAP}$ of 0.4 Å was obtained for all studied nuclei ($^{29}$Si, $^{17}$O and $^{23}$Na). Note that no attempt was made to a set different value for each atom. Representative MAE convergence curves are shown in Figure 2 using the MD-300K as training sets. We observe that increasing the number of radial functions (to better captures the diversity of the environment) effectively requires an increase of $N_\xi^{opt}$. The final optimal values that were determined from the dataset MD-300K + MD-1000K (the sets used to built the final LSSVR predictors) are given in Table **Error! Reference source not found.** and convergence curves (MAE and RMSE) are shown in Figure 3. Note that a higher value of $N_\xi^{opt}$ is necessary with this composite set versus the individual set (for example, $N_\xi^{opt} = 1200$ for $^{17}$O in MD-300K and -1000K, see section S5 of the ESI) showing the complementarity of the two datasets.

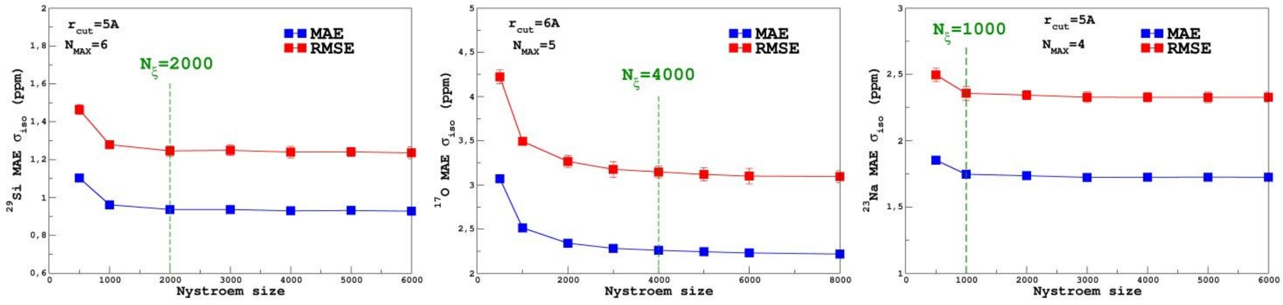
*Table 1. Parameters of SOAP descriptors $c^{\mu}_{n,l,m}$ eqn.(1), $\sigma_{iso}$ mean absolute error (MAE) and root mean square error (RMSE) (with standard deviations values in parentheses).*

| atom | $\sigma_{SOAP}$ (Å) | $r_{cut}$ (Å) | $L_{max}$ | $N_{max}$ | $N^{opt}_{\xi}$ (MD-300K+1000K) | MAE (ppm) | RMSE (ppm) |
|---|---|---|---|---|---|---|---|
| Si | 0.4 | 5 | 4 | 4 | 2000 | 0.93 (0.02) | 1.25 (0.03) |
| Na | 0.4 | 5 | 4 | 6 | 1000 | 1.39 (0.02) | 1.79 (0.03) |
| O | 0.4 | 6 | 4 | 5 | 4000 | 2.25 (0.02) | 3.15 (0.07) |



$^{29}$Si $^{17}$O $^{23}$Na

*Figure 2. Convergence of $\sigma_{iso}$ MAE with respect to the Nyström size $N_{\xi}$and the number of radial functions ($nRad = N_{max}$) in the SOAP descriptors for the MD-300K training set (other SOAP parameters given in Table **Error! Reference source not found.**)*



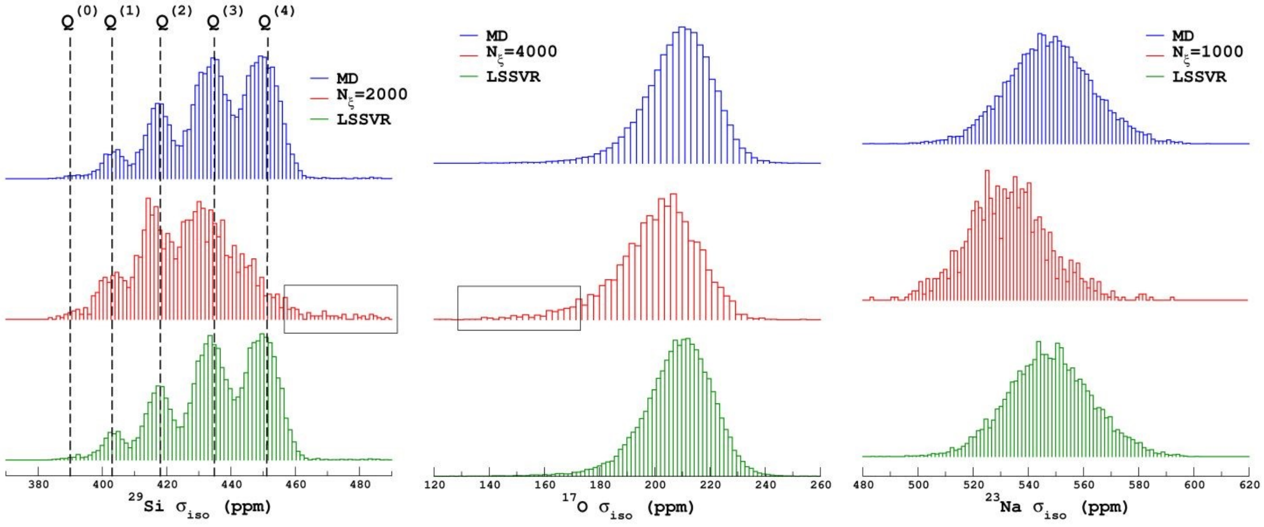
$^{29}$Si $^{17}$O $^{23}$Na

*Figure 3. Convergence of the $\sigma_{iso}$MAE with respect to the Nyström size $N_{\xi}$for MD-300K+1000K training sets (SOAP parameters given in Table **Error! Reference source not found.**)*

Similarly to Figure 1, the selection of the inducing points $\xi$ by ICD can be effectively visualized by examination of the NMR parameter distribution as shown in Figure 4. We first note the excellent prediction by LSSVR. Interestingly, ICD produces a distribution that differs from the original datasets by enhancing for example regions in the tails. Taking an as example $^{29}$Si data, clearly the Q$^{(3)}$ and Q$^{(2)}$ regions are significantly enhanced (which thus suggests a higher local structural diversity) versus the (more regular) Q$^{(4)}$ region. Note that the weak region Q$^{(0)}$ is well retained by ICD. These results also confirm that the SOAP descriptors combined with a kernel-ICD approach are providing an efficient procedure for extracting the representative set of environments in a database. Attempts to force the selection of inducing points uniformly distributed on $\sigma_{iso}$ did not improved the LSSVR predictions.

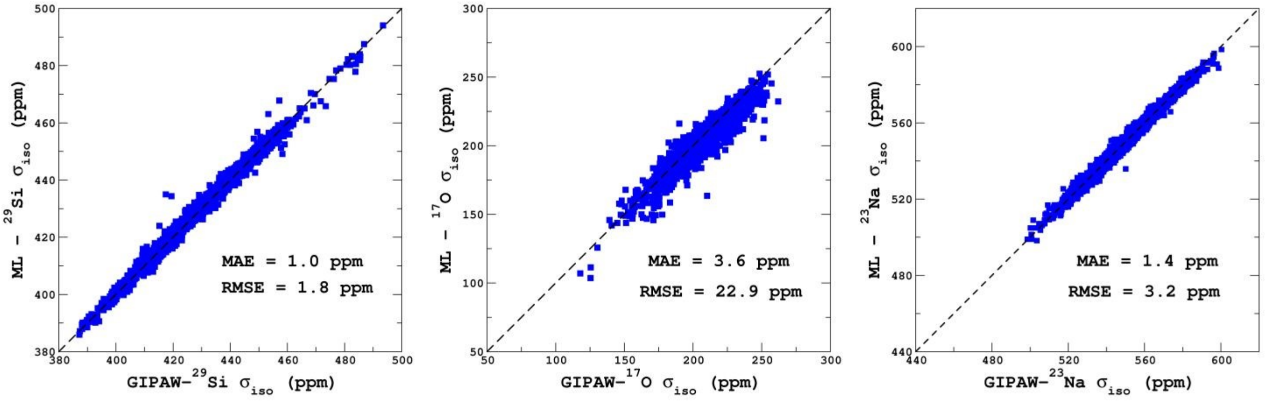
View Article Online
DOI: 10.1039/D4FD00129J

*Figure 4. Distribution of isotropic magnetic shielding $\sigma_{iso}$values from MD-300K+1000K datasets, inducing points selected by ICD and the LSSVR predictions. Distribution are normalized to the same area. Black boxes highlight the weak intensity regions well captured by ICD.*



$^{29}$Si $^{17}$O $^{23}$Na

Transferability test between the different training sets were performed and are reported in section S5 of the ESI. As expected, high-temperature sets gives better accuracy when tested on lower temperature than the opposite case. Transferability to larger systems (i.e. from 300 to 600 atoms) were also excellent and are illustrated in Figure 5.



$^{29}$Si $^{17}$O $^{23}$Na

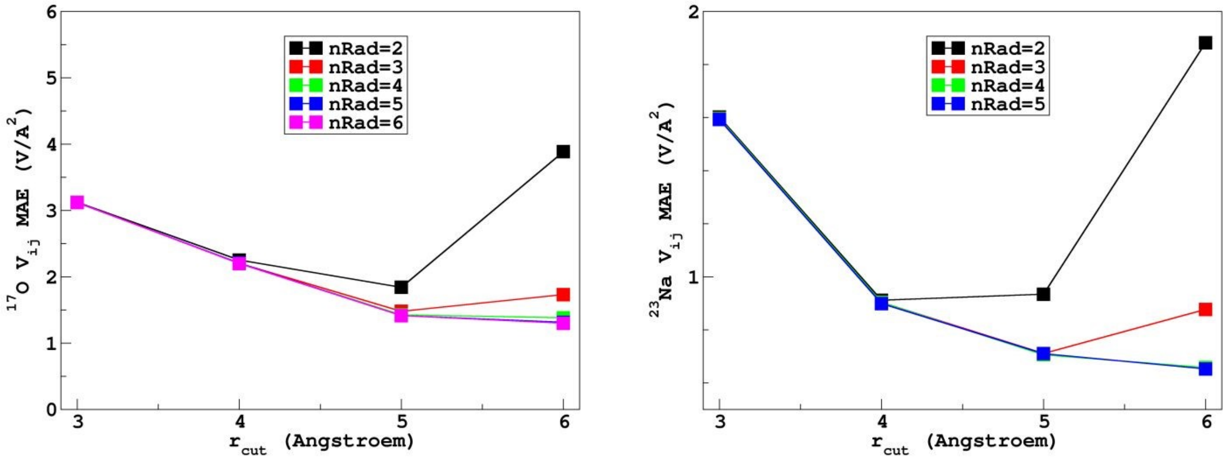
*Figure 5. LSSVR versus DFT-GIPAW $\sigma_{iso}$ values for $^{29}$Si, $^{17}$O and $^{23}$Na in the 600 atoms MD datasets (300K+1000K). Training was performed on the 300 atoms data.*

## Learning the EFG tensor

View Article Online
DOI: 10.1039/D4FD00129J

It would tempting to use LSSVR to predict the *scalar* quadrupolar parameters $(C_Q,\eta)$. However both quantities depend on the ordering of the eigenvalues of the EFG tensor and are therefore discontinuous. Effectively, they are very poorly predicted by LSSVR as shown in section S6 of the ESI. Other ML algorithms (not investigated here) that are robust to discontinuities (such as neural networks or random forests) could be better adapted but their investigation is out of the scope of the present study. Better LSSVR results are obtained for $P_Q$(which is invariant to rotation) but still showing some discrepancies. For these reason, the option of predicting the full EFG tensor appeared as the best one.

Optimizations of λ-SOAP descriptors for $^{23}$Na and $^{17}$O EFG tensors are summarized in Figure 6. Despite contributions from long-range Coulombic interactions, the full EFG tensor is very well captured by the short-range SOAP descriptors (with $r_{cut} \geq 5$Å). As was observed for LSSVR, increasing $r_{cut}$ requires a larger number of radial functions. For sake of simplicity and efficiency, values of Table **Error! Reference source not found.** were chosen (so that SOAP descriptors need only to be computed once for predicting the three NMR parameters). Such good performances of λ-SOAP can be understood as resulting from the correlation that exists between the short- and long-range contributions to the EFG, analogous to the Sternheimer approximation.[78] Such correlations were observed in water for quadrupolar nuclei.[79] With the MD-300K+1000K datasets, LRR predictions MAE (RMSE) for the $^{17}$O and $^{23}$Na EFG components are 1.6 V/Å$^2$ (2.0 V/Å) and 0.8 V/Å$^2$ (1.0 V/Å) respectively, representing ~1% of their respective total spans. Using the 600 atoms MD models (300K+1000K), the same numbers were obtained showing therefore an excellent transferability, as shown in Figure 7. $^{23}$Na and $^{17}$O quadrupolar parameter distributions are very well predcited by LRR λ-SOAP as shown in Figure 8.



$^{17}$O $^{23}$Na

*Figure 6. Variations of the LRR λ-SOAP mean absolute error (MAE) of the EFG tensor components in the MD-300K datasets.*

View Article Online
DOI: 10.1039/D4FD00129J

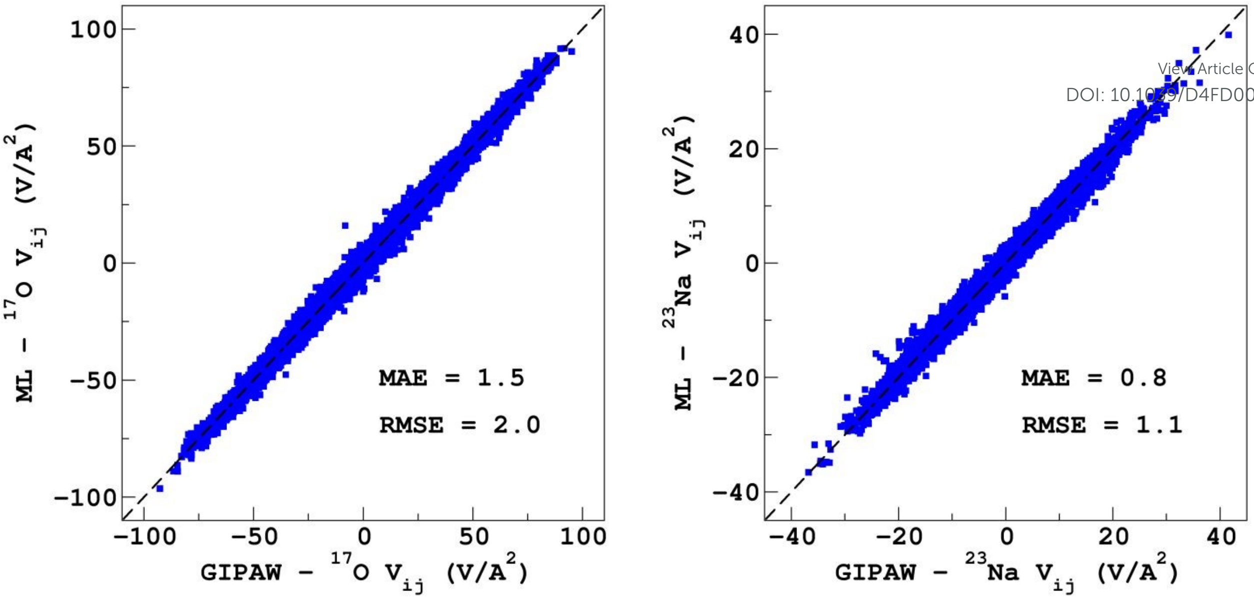


*Figure 7. LRR λ-SOAP versus DFT-GIPAW EFG tensor components of the 600 atoms MD datasets (300K+1000K).*

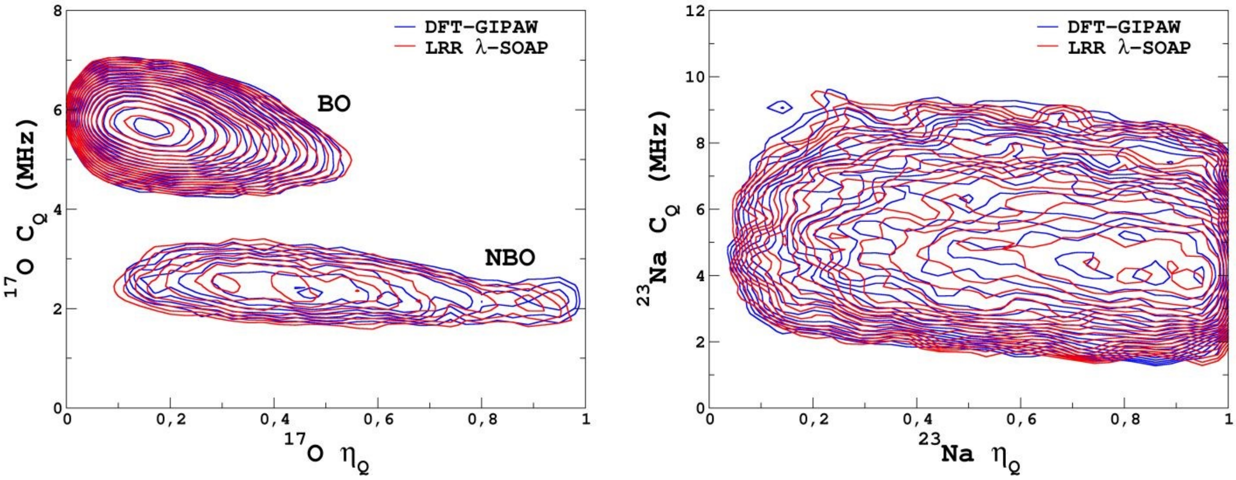


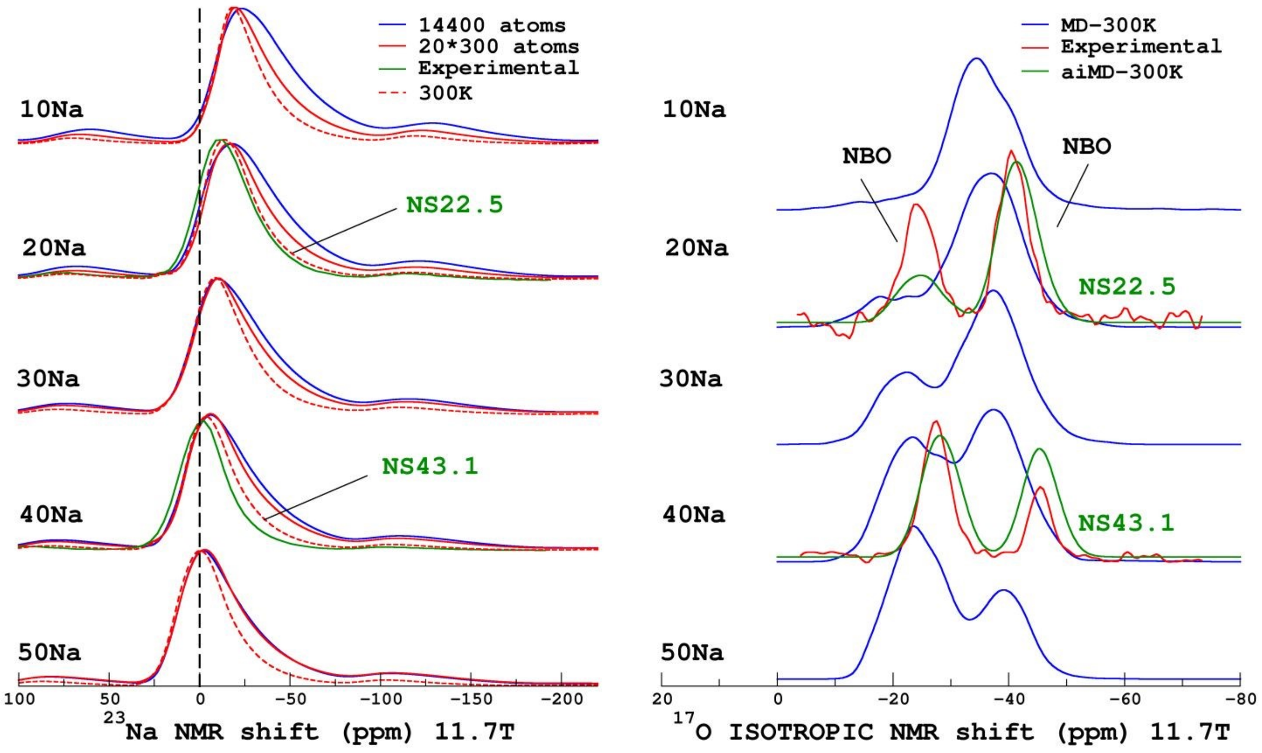
*Figure 8. LRR λ-SOAP versus DFT-GIPAW NMR parameter distribution $p(C_Q,\eta)$ for the MD-300K+1000K datasets.*

## KDE simulation of the MAS and MQMAS NMR spectra

The original motivation for the ML prediction of NMR parameters was to enable the modelling of NMR spectra from structural models of large size (here, 14440 atoms) as with modern HPC resources DFT-GIPAW typically addresses 800 atoms with VASP. All subsequent simulations were performed on a standard single processor (Intel CORE i7). Calculations for the large models took around 2 seconds; the most time consuming part was the calculation of the SOAP descriptors (see Figure S5 in the ESI). $^{23}$Na and $^{17}$O MAS NMR spectra (for the latter we show the isotropic projections of the MQMAS spectrum) are shown in Figure 9 (MQMAS spectra are given in sections S8 and S9 of the ESI).

Focusing first on the $^{23}$Na NMR data and classical MD models, we observe a discrepancy between the large and small models spectra. This is indicative of the effect of the PBC on the structure of the small models but without impacting the transferability of the ML predictors (Figure 5 and Figure 7). The comparison with experimental data shows a good agreement, except for the width of the spectra which is due to an overestimated mean values of the quadrupolar coupling constant, as already observed.[29] As will be shown below, accounting for short-times scale averaging by vibrations significantly reduce the discrepancies. Considering the $^{17}$O NMR data, we note a strong deviation of the MD NMR models spectra from the experimental data. This is due to approximate Si-O and Si-O-Si bond angle values predicted in classical MD, in contrast to aiMD (DFT) data which yields an excellent agreement in the position of the NBO and BO peaks (intensities differs because impact of MQMAS pulse sequence was not taken into account for sake of simplicity). Those simulated spectra can be useful for many purposes. One of them is the assessment of analytical models of NMR parameter distribution, such as the Gaussian Isotropic Model[80, 81] (GIM) for fitting the $^{23}$Na NMR data for example.

View Article Online
DOI: 10.1039/D4FD00129J



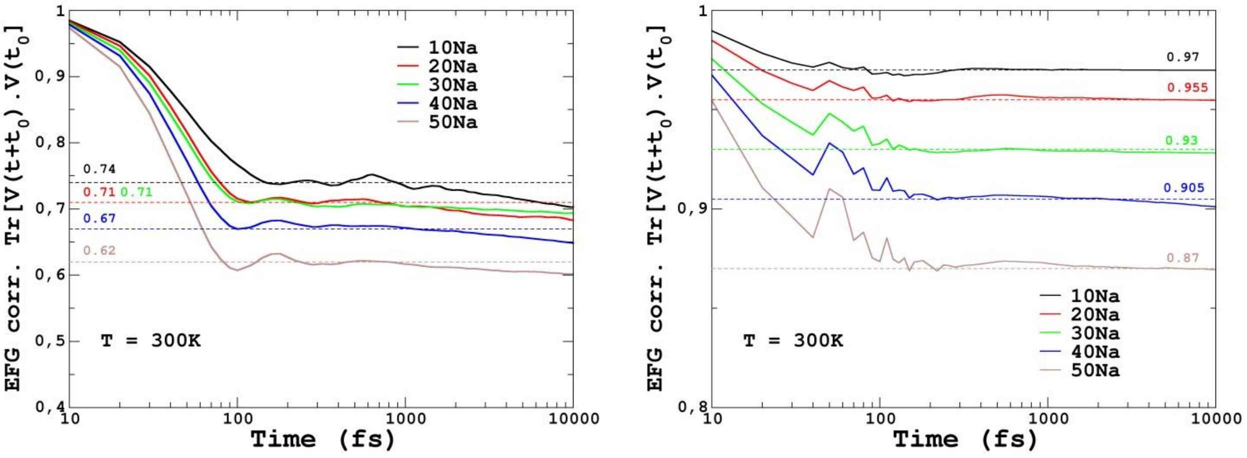
$^{23}$Na $^{17}$O

*Figure 9. ML simulations of $^{23}$Na (left) and $^{17}$O (right) MAS and isotropic (isotropic projection of the MQMAS spectra) NMR spectra of the $SiO_2$-$Na_2O$ glasses from 10% to 50% mol. $Na_2O$ (denoted 10Na to 50Na). For purpose of comparison, experimental data from Ref.[27] are shown (22.5% and 43.1% mol. $Na_2O$, denoted NS22.5 and NS43.1 respectively). For $^{17}$O, ML simulations from aiMD structural models (aiMD-300K) are also shown.*

To quantify the impact of atomic vibrations on the NMR parameters, we investigated the auto-correlation functions of the two NMR interaction of interest in this work, $G_{\delta}(\tau) = \langle \delta(\tau + t_0)\delta(t_0) \rangle_{t_0}$ and $G_{EFG}(\tau) = \langle V(\tau + t_0)V(t_0) \rangle_{t_0}$, calculated on MD trajectories at 300K. $\langle\rangle_{t_0}$ denotes the ensemble average that includes here the averaging over all initial times $t_0$ and over all atoms. At short-time scale, $G_{\delta}(\tau)$ is a constant function (~1) at 300K, thus meaning a very minimal impact of vibrations. In contrast, $G_{EFG}(\tau)$is

View Article Online
DOI: 10.1039/D4FD00129J

decaying on typical time-scale of vibrations (10-100 fs), with a stronger effect for $^{23}$Na, in contrast to $^{17}$O, as illustrated in Figure 10. After that first decay, the EFG time-correlation function reaches plateau that can be considered, in a first approximation, as the time-averaged EFG that really contribute to the NMR spectra (whereas the initial decaying part controls the relaxation times[50] but this will be not discussed here). Using the plateauing value as a scaling factor of the *frozen* EFG tensors (we assume an isotropic scaling) vibrations can then be simply incorporated into the NMR spectra, as shown for $^{23}$Na in (left panel, dashed lines). This clearly improves the simulations (see especially 20Na). We note that for 40Na, the experimental width is still overestimated. Indeed, longer time-scale would be needed to account for the diffusion of Na atom at 300K (Na rich glasses show a higher Na mobility[82]). The time-scale of the MD simulations is still well below the typical Larmor period for $^{23}$Na ( here ~ 8ns at 11.7T). Interestingly, aiMD simulations gave very close curves for the EFG correlation functions, as shown in the section S10 of the ESI. This means that classical MD potentials are able to capture the vibrational averaging with a good accuracy.



$^{23}$Na $^{17}$O

*Figure 10. $^{23}$Na and $^{17}$O EFG correlation functions $G_{EFG}(\tau) = \langle V(\tau + t_0)V(t_0)\rangle_{t_0}$. In left and right panels, plateauing values used as EFG scaling factor for finite temperature simulations are indicated.*

## Conclusion

We have described in this paper new computational methodologies for modelling NMR spectra of oxide glasses, combining prediction of NMR properties by ML kernel methods with efficient simulations using KDE of the NMR parameter distribution. We have shown that the SOAP descriptors are very efficient and symmetry adapted for representing the local environment for the prediction of NMR interactions, be it a scalar (isotropic magnetic shielding value) or a matrix (the EFG second-rank symmetric tensor). Our strategy to build NMR database that embraces a sufficiently large variety of environments (in term of both geometrical and chemical disorder) was based on MD simulations at various temperatures and for various glass compositions. It was shown that small models were suitable to build ML predictors transferable to much larger systems (here, more than 10 000 atoms). Most representative environments could be extracted with incomplete Cholesky decomposition of the kernel Gram matrix of the dataset, providing an efficient tool for analysis of the database, as confirmed by the examination of the NMR parameter distributions. Appealing perspectives for an easy incorporation of finite temperature effects (vibrations) in NMR simulations were presented. Because of its fundamental importance in glass science, ML-NMR will clearly enable NMR investigations of melts (with high temperature NMR[20]) to be now more closely connected to MD simulations by computation of the underlying correlation functions of NMR observations. The next step is the incorporation of NMR spectra as direct constraints in the reconstruction of 3d glass structure of glasses, in Reverse Monte-Carlo simulations widely used in glass science.[11, 12, 83]

View Article Online
DOI: 10.1039/D4FD00129J

## Conflicts of Interest

There are no conflicts to declare.

## Data Availability Statement

The data that support the findings of this study (DFT-GIPAW data for structures of the NMR database) are available at 10.5281/zenodo.12314395 from the author upon reasonable request.

## Acknowledgements

This work was granted access to the HPC resources of TGCC under the allocation DARI-A0110906303, DARI-A0130906303, DARI-A0150906303 attributed by GENCI (Grand Equipement National de Calcul Intensif).

## Notes and References

1. Youngman RE. 22 Silicate Glasses and Their Impact on Humanity. *22 Silic. Glas. Their Impact Humanity*. De Gruyter; 2022:1015–1038. https://doi.org/10.1515/9781501510939-023
2. Mauro JC, Philip CS, Vaughn DJ, Pambianchi MS. Glass Science in the United States: Current Status and Future Directions. *Int J Appl Glass Sci*. 2014;5(1):2–15. https://doi.org/10.1111/ijag.12058
3. Li L, Lin H, Qiao S, *et al.* Integrated flexible chalcogenide glass photonic devices. *Nat Photonics*. 2014;8(8):643–649. https://doi.org/10.1038/nphoton.2014.138
4. Gin S, Abdelouas A, Criscenti LJ, *et al.* An international initiative on long-term behavior of high-level nuclear waste glass. *Mater Today*. 2013;16(6):243–248. https://doi.org/10.1016/j.mattod.2013.06.008
5. Varshneya AK, Mauro JC. Fundamentals of Inorganic Glasses. 3 edition. Elsevier; 2019
6. Liu H, Zhao Z, Zhou Q, *et al.* Challenges and opportunities in atomistic simulations of glasses: a review. *Comptes Rendus Géoscience*. 2022;354(S1):35–77. https://doi.org/10.5802/crgeos.116
7. Takada A. Atomistic Simulations of Glass Structure and Properties. 1st ed. In: Richet P, Conradt R, Takada A, Dyon J, eds. *Encycl. Glass Sci. Technol. Hist. Cult.*, 1st ed. Wiley; 2021:221–232. https://doi.org/10.1002/9781118801017.ch2.8
8. Du J. Molecular Dynamics Simulations of Oxide Glasses. In: Musgraves JD, Hu J, Calvez L, eds. *Springer Handb. Glass*. Cham: Springer International Publishing; 2019:1131–1155. https://doi.org/10.1007/978-3-319-93728-1_32
9. Atomistic Simulations of Glasses: Fundamentals and Applications – ICG | International Commission on Glass. 2022.
10. Hannon AC. Neutron Diffraction Techniques for Structural Studies of Glasses. *Mod. Glass Charact.* John Wiley & Sons, Ltd; 2015:1–83. https://doi.org/10.1002/9781119051862.ch5
11. Evrard G, Pusztai L. Reverse Monte Carlo modelling of the structure of disordered materials with RMC++ : a new implementation of the algorithm in C++. *J Phys Condens Matter*. 2005;17(5):S1. https://doi.org/10.1088/0953-8984/17/5/001
12. McGreevy RL. Reverse Monte Carlo modelling. *J Phys Condens Matter*. 2001;13(46):R877. https://doi.org/10.1088/0953-8984/13/46/201
13. Edén M. Chapter Four - 27Al NMR Studies of Aluminosilicate Glasses. In: Webb GA, ed. *Annu. Rep. NMR Spectrosc.* Vol. 86. Academic Press; 2015:237–331.
14. Edén M. NMR studies of oxide-based glasses. *Annu Rep Sect C Phys Chem*. 2012;108(1):177–221. https://doi.org/10.1039/C2PC90006H

15. Massiot D, Messinger RJ, Cadars S, *et al.* Topological, Geometric, and Chemical Order in Materials: Insights from Solid-State NMR. *Acc Chem Res*. 2013;46(9):1975–1984. https://doi.org/10.1021/ar3003255
16. Edén M. Probing oxide-based glass structures by solid-state NMR: Opportunities and limitations. *J Magn Reson Open*. 2023;16–17:100112. https://doi.org/10.1016/j.jmro.2023.100112
17. MacKenzie KJD, Smith ME. Multinuclear Solid-State Nuclear Magnetic Resonance of Inorganic Materials, Volume 6. 1st ed. Pergamon; 2002
18. Eckert H. Structural characterization of noncrystalline solids and glasses using solid state NMR. *Prog Nucl Magn Reson Spectrosc*. 1992;24(3):159–293. https://doi.org/10.1016/0079-6565(92)80001-V
19. Deschamps M, Fayon F, Montouillout V, Massiot D. Through-bond homonuclear correlation experiments in solid-state NMR applied to quadrupolar nuclei in Al–O–P–O–Al chains. *Chem Commun*. 2006;(18):1924–1925. https://doi.org/10.1039/B600514D
20. Massiot D, Fayon F, Montouillout V, *et al.* Structure and dynamics of oxide melts and glasses: A view from multinuclear and high temperature NMR. *J Non-Cryst Solids*. 2008;354(2–9):249–254. https://doi.org/10.1016/j.jnoncrysol.2007.06.097
21. Deschamps M, Fayon F, Hiet J, *et al.* Spin-counting NMR experiments for the spectral editing of structural motifs in solids. *Phys Chem Chem Phys*. 2008;10(9):1298–1303. https://doi.org/10.1039/B716319C
22. Eckert H. Advanced Dipolar Solid State Nmr Spectroscopy of Glasses. *Mod. Glass Charact.* John Wiley & Sons, Ltd; 2015:1–46. https://doi.org/10.1002/9781119051862.ch9
23. Pickard CJ, Mauri F. All-electron magnetic response with pseudopotentials: NMR chemical shifts. *Phys Rev B*. 2001;63(24):245101. https://doi.org/10.1103/PhysRevB.63.245101
24. Charpentier T. The PAW/GIPAW approach for computing NMR parameters: A new dimension added to NMR study of solids. *Solid State Nucl Magn Reson*. 2011;40(1):1–20. https://doi.org/10.1016/j.ssnmr.2011.04.006
25. Bonhomme C, Gervais C, Babonneau F, *et al.* First-Principles Calculation of NMR Parameters Using the Gauge Including Projector Augmented Wave Method: A Chemist's Point of View. *Chem Rev*. 2012;112(11):5733–5779. https://doi.org/10.1021/cr300108a
26. Massiot D, Fayon F, Capron M, *et al.* Modelling one- and two-dimensional solid-state NMR spectra. *Magn Reson Chem*. 2002;40(1):70–76. https://doi.org/10.1002/mrc.984
27. Angeli F, Villain O, Schuller S, Ispas S, Charpentier T. Insight into sodium silicate glass structural organization by multinuclear NMR combined with first-principles calculations. *Geochim Cosmochim Acta*. 2011;75(9):2453–2469. https://doi.org/10.1016/j.gca.2011.02.003
28. Gambuzzi E, Pedone A, Menziani MC, Angeli F, Florian P, Charpentier T. Calcium environment in silicate and aluminosilicate glasses probed by 43Ca MQMAS NMR experiments and MD-GIPAW calculations. *Solid State Nucl Magn Reson*. 2015;68–69:31–36. https://doi.org/10.1016/j.ssnmr.2015.04.003
29. Gambuzzi E, Charpentier T, Menziani MC, Pedone A. Computational interpretation of 23 Na MQMAS NMR spectra: A comprehensive investigation of the Na environment in silicate glasses. *Chem Phys Lett*. 2014;612:56–61. https://doi.org/10.1016/j.cplett.2014.08.004
30. Angeli F, Gaillard M, Jollivet P, Charpentier T. Contribution of 43Ca MAS NMR for probing the structural configuration of calcium in glass. *Chem Phys Lett*. 2007;440(4–6):324–328. https://doi.org/10.1016/j.cplett.2007.04.036
31. Angeli F, Villain O, Schuller S, *et al.* Effect of temperature and thermal history on borosilicate glass structure. *Phys Rev B*. 2012;85(5):054110. https://doi.org/10.1103/PhysRevB.85.054110
32. Clark TM, Grandinetti PJ, Florian P, Stebbins JF. Correlated structural distributions in silica glass. *Phys Rev B*. 2004;70(6):064202. https://doi.org/10.1103/PhysRevB.70.064202
33. Charpentier T, Kroll P, Mauri F. First-Principles Nuclear Magnetic Resonance Structural Analysis of Vitreous Silica. *J Phys Chem C*. 2009;113(18):7917–7929. https://doi.org/10.1021/jp900297r
34. Angeli F, Charpentier T, Faucon P, Petit J-C. Structural Characterization of Glass from the Inversion of 23Na and 27Al 3Q-MAS NMR Spectra. *J Phys Chem B*. 1999;103(47):10356–10364. https://doi.org/10.1021/jp9910035
35. Vasconcelos F, Wijs GA de, Havenith RWA, Marsman M, Kresse G. Finite-field implementation of NMR chemical shieldings for molecules: Direct and converse gauge-including projector-augmented-wave methods. *J Chem Phys*. 2013;139(1):014109. https://doi.org/10.1063/1.4810799

36. Pickard CJ, Mauri F. Nonlocal Pseudopotentials and Magnetic Fields. *Phys Rev Lett*. 2003;91(19):196401. https://doi.org/10.1103/PhysRevLett.91.196401
37. Joyce SA, Yates JR, Pickard CJ, Mauri F. A first principles theory of nuclear magnetic resonance J-coupling in solid-state systems. *J Chem Phys*. 2007;127(20):204107. https://doi.org/10.1063/1.2801984
38. Charpentier T, Ispas S, Profeta M, Mauri F, Pickard CJ. First-Principles Calculation of 17O, 29Si, and 23Na NMR Spectra of Sodium Silicate Crystals and Glasses. *J Phys Chem B*. 2004;108(13):4147–4161. https://doi.org/10.1021/jp0367225
39. Charpentier T, Menziani MC, Pedone A. Computational simulations of solid state NMR spectra: a new era in structure determination of oxide glasses. *RSC Adv*. 2013;3(27):10550–10578. https://doi.org/10.1039/C3RA40627J
40. Pedone A. Recent advances in solid-state NMR computational spectroscopy: The case of alumino-silicate glasses. *Int J Quantum Chem*. 2016;n/a-n/a. https://doi.org/10.1002/qua.25134
41. Pedone A, Charpentier T, Menziani MC. Multinuclear NMR of CaSiO3 glass: simulation from first-principles. *Phys Chem Chem Phys*. 2010;12(23):6054. https://doi.org/10.1039/b924489a
42. Pedone A, Charpentier T, Malavasi G, Menziani MC. New Insights into the Atomic Structure of 45S5 Bioglass by Means of Solid-State NMR Spectroscopy and Accurate First-Principles Simulations. *Chem Mater*. 2010;22(19):5644–5652. https://doi.org/10.1021/cm102089c
43. Chakraborty S, Bobela DC, Taylor PC, Drabold DA. Voids in Hydrogenated Amorphous Silicon: A Comparison of ab initio Simulations and Proton NMR Studies. *MRS Proc*. 2008;1066:1066-A11-02. https://doi.org/10.1557/PROC-1066-A11-02
44. Bertani M, Bisbrouck N, Delaye J-M, Angeli F, Pedone A, Charpentier T. Effects of magnesium on the structure of aluminoborosilicate glasses: NMR assessment of interatomic potentials models for molecular dynamics. *J Am Ceram Soc*. 2023;106(9):5501–5521. https://doi.org/10.1111/jace.19157
45. Ishii Y, Salanne M, Charpentier T, Shiraki K, Kasahara K, Ohtori N. A DFT-Based Aspherical Ion Model for Sodium Aluminosilicate Glasses and Melts. *J Phys Chem C*. 2016;120(42):24370–24381. https://doi.org/10.1021/acs.jpcc.6b08052
46. Harper AF, Emge SP, Magusin PCMM, Grey CP, Morris AJ. Modelling amorphous materials via a joint solid-state NMR and X-ray absorption spectroscopy and DFT approach: application to alumina. *Chem Sci*. 2023;14(5):1155–1167. https://doi.org/10.1039/D2SC04035B
47. Dumez J-N, Pickard CJ. Calculation of NMR chemical shifts in organic solids: Accounting for motional effects. *J Chem Phys*. 2009;130(10):104701-104701–8. https://doi.org/doi:10.1063/1.3081630
48. Schmidt J, Hutter J, Spiess H-W, Sebastiani D. Beyond Isotropic Tumbling Models: Nuclear Spin Relaxation in Liquids from First Principles. *ChemPhysChem*. 2008;9(16):2313–2316. https://doi.org/10.1002/cphc.200800435
49. Sen S, Mukerji T. A molecular dynamics simulation study of ionic diffusion and NMR spin–lattice relaxation in Li2Si4O9 glass. *J Non-Cryst Solids*. 2001;293–295:268–278. https://doi.org/10.1016/S0022-3093(01)00679-2
50. Badu S, Truflandier L, Autschbach J. Quadrupolar NMR Spin Relaxation Calculated Using Ab Initio Molecular Dynamics: Group 1 and Group 17 Ions in Aqueous Solution. *J Chem Theory Comput*. 2013;9(9):4074–4086. https://doi.org/10.1021/ct400419s
51. Schütt KT, Chmiela S, Von Lilienfeld OA, Tkatchenko A, Tsuda K, Müller K-R, editors. Machine Learning Meets Quantum Physics. Cham: Springer International Publishing; 2020 https://doi.org/10.1007/978-3-030-40245-7
52. Paruzzo FM, Hofstetter A, Musil F, De S, Ceriotti M, Emsley L. Chemical Shifts in Molecular Solids by Machine Learning. 2018.
53. Rupp M, Ramakrishnan R, von Lilienfeld OA. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *J Phys Chem Lett*. 2015;6(16):3309–3313. https://doi.org/10.1021/acs.jpclett.5b01456
54. Cuny J, Xie Y, Pickard CJ, Hassanali AA. Ab Initio Quality NMR Parameters in Solid-State Materials Using a High-Dimensional Neural-Network Representation. *J Chem Theory Comput*. 2016. https://doi.org/10.1021/acs.jctc.5b01006
55. Chaker Z, Salanne M, Delaye J-M, Charpentier T. NMR shifts in aluminosilicate glasses via machine learning. *Phys Chem Chem Phys*. 2019. https://doi.org/10.1039/C9CP02803J

56. Gaumard R, Dragún D, Pedroza-Montero JN, *et al.* Regression Machine Learning Models Used to Predict DFT-Computed NMR Parameters of Zeolites. *Computation*. 2022;10(5):74. https://doi.org/10.3390/computation10050074
57. Ohkubo T, Takei A, Tachi Y, *et al.* New Approach To Understanding the Experimental 133Cs NMR Chemical Shift of Clay Minerals via Machine Learning and DFT-GIPAW Calculations. *J Phys Chem A*. 2023;127(4):973–986. https://doi.org/10.1021/acs.jpca.2c08880
58. Cordova M, Balodis M, Hofstetter A, *et al.* Structure determination of an amorphous drug through large-scale NMR predictions. *Nat Commun*. 2021;12(1):2964. https://doi.org/10.1038/s41467-021-23208-7
59. Cordova M, Moutzouri P, Nilsson Lill SO, *et al.* Atomic-level structure determination of amorphous molecular solids by NMR. *Nat Commun*. 2023;14(1):5138. https://doi.org/10.1038/s41467-023-40853-2
60. Behler J, Parrinello M. Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces. *Phys Rev Lett*. 2007;98(14):146401. https://doi.org/10.1103/PhysRevLett.98.146401
61. Bartók AP, Kondor R, Csányi G. On representing chemical environments. *Phys Rev B*. 2013;87(18):184115. https://doi.org/10.1103/PhysRevB.87.184115
62. Schmidt J, Marques MRG, Botti S, Marques MAL. Recent advances and applications of machine learning in solid-state materials science. *Npj Comput Mater*. 2019;5(1):1–36. https://doi.org/10.1038/s41524-019-0221-0
63. Langer MF, Goeßmann A, Rupp M. Representations of molecules and materials for interpolation of quantum-mechanical simulations via machine learning. *Npj Comput Mater*. 2022;8(1):1–14. https://doi.org/10.1038/s41524-022-00721-x
64. Deringer VL, Bernstein N, Bartók AP, *et al.* Realistic Atomistic Structure of Amorphous Silicon from Machine-Learning-Driven Molecular Dynamics. *J Phys Chem Lett*. 2018;9(11):2879–2885. https://doi.org/10.1021/acs.jpclett.8b00902
65. Grisafi A, Wilkins DM, Csányi G, Ceriotti M. Symmetry-Adapted Machine Learning for Tensorial Properties of Atomistic Systems. *Phys Rev Lett*. 2018;120(3):036002. https://doi.org/10.1103/PhysRevLett.120.036002
66. Kocer E, Mason JK, Erturk H. Continuous and optimally complete description of chemical environments using Spherical Bessel descriptors. *AIP Adv*. 2020;10(1):015021. https://doi.org/10.1063/1.5111045
67. Jinnouchi R, Karsai F, Kresse G. On-the-fly machine learning force field generation: Application to melting points. *Phys Rev B*. 2019;100(1):014105. https://doi.org/10.1103/PhysRevB.100.014105
68. Jinnouchi R, Karsai F, Verdi C, Asahi R, Kresse G. Descriptors representing two- and three-body atomic distributions and their effects on the accuracy of machine-learned inter-atomic potentials. *J Chem Phys*. 2020;152(23):234102. https://doi.org/10.1063/5.0009491
69. Willatt MJ, Musil F, Ceriotti M. Atom-density representations for machine learning. *J Chem Phys*. 2019;150(15):154110. https://doi.org/10.1063/1.5090481
70. Rasmussen CE, Williams CKI. Gaussian Processes for Machine Learning. The MIT Press; 2005 https://doi.org/10.7551/mitpress/3206.001.0001
71. De Brabanter K, De Brabanter J, Suykens JAK, De Moor B. Optimized fixed-size kernel models for large data sets. *Comput Stat Data Anal*. 2010;54(6):1484–1504. https://doi.org/10.1016/j.csda.2010.01.024
72. Bach FR, Jordan MI. Predictive low-rank decomposition for kernel methods. *Proc. 22nd Int. Conf. Mach. Learn.* New York, NY, USA: Association for Computing Machinery; 2005:33–40. https://doi.org/10.1145/1102351.1102356
73. Soleilhavoup A, Delaye J-M, Angeli F, Caurant D, Charpentier T. Contribution of first-principles calculations to multinuclear NMR analysis of borosilicate glasses. *Magn Reson Chem*. 2010;48(S1):S159–S170. https://doi.org/10.1002/mrc.2673
74. Silverman BW. Density Estimation for Statistics and Data Analysis by B. W. Silverman (1986) Hardcover. n.d.
75. Fu C, Cai D. EFANNA : An Extremely Fast Approximate Nearest Neighbor Search Algorithm Based on kNN Graph. 2016. https://doi.org/10.48550/arXiv.1609.07228
76. Bertani M, Charpentier T, Faglioni F, Pedone A. Accurate and Transferable Machine Learning Potential for Molecular Dynamics Simulation of Sodium Silicate Glasses. *J Chem Theory Comput*. 2024;20(3):1358–1370. https://doi.org/10.1021/acs.jctc.3c01115

View Article Online
DOI: 10.1039/D4FD00129J

77. Kühne TD, Iannuzzi M, Del Ben M, *et al.* CP2K: An electronic structure and molecular dynamics software package - Quickstep: Efficient and accurate electronic structure calculations. *J Chem Phys*. 2020;152(19):194103. https://doi.org/10.1063/5.0007045

78. Schmidt PC, Sen KD, Das TP, Weiss A. Effect of self-consistency and crystalline potential in the solid state on nuclear quadrupole Sternheimer antishielding factors in closed-shell ions. *Phys Rev B*. 1980;22(9):4167–4179. https://doi.org/10.1103/PhysRevB.22.4167

79. Carof A, Salanne M, Charpentier T, Rotenberg B. Accurate Quadrupolar NMR Relaxation Rates of Aqueous Cations from Classical Molecular Dynamics. *J Phys Chem B*. 2014;118(46):13252–13257. https://doi.org/10.1021/jp5105054

80. Caër GL, Bureau B, Massiot D. An extension of the Czjzek model for the distributions of electric field gradients in disordered solids and an application to NMR spectra of 71Ga in chalcogenide glasses. *J Phys Condens Matter*. 2010;22(6):065402. https://doi.org/10.1088/0953-8984/22/6/065402

81. Czjzek G, Fink J, Götz F, *et al.* Atomic coordination and the distribution of electric field gradients in amorphous solids. *Phys Rev B*. 1981;23(6):2513–2530. https://doi.org/10.1103/PhysRevB.23.2513

82. Du J, Cormack AN. The medium range structure of sodium silicate glasses: a molecular dynamics simulation. *J Non-Cryst Solids*. 2004;349:66–79. https://doi.org/10.1016/j.jnoncrysol.2004.08.264

83. Hannon AC, Vaishnav S, Alderman OLG, Bingham PA. The structure of sodium silicate glass from neutron diffraction and modeling of oxygen-oxygen correlations. *J Am Ceram Soc*. 2021;104(12):6155–6171. https://doi.org/10.1111/jace.17993

# First-principles NMR of oxide glasses boosted by machine learning

View Article Online
DOI: 10.1039/D4FD00129J

Thibault Charpentier[1]

Université Paris-Saclay, CEA, CNRS, NIMBE, 91191 Gif-sur-Yvette cedex, France.

## Data Availability Statement

The data that support the findings of this study (DFT-GIPAW data for structures of the NMR database) are available at *10.5281/zenodo.12314395* or from the author upon reasonable request.

---

[1] Université Paris-Saclay, CEA, CNRS, NIMBE, 91191 Gif-sur-Yvette cedex, France. thibault.charpentier@cea.fr