Check for updates

ChemPhysChem

Research Article
doi.org/10.1002/cphc.202300782

Chemistry Europe
European Chemical Societies Publishing

www.chemphyschem.org

# Accelerating NMR Shielding Calculations Through Machine Learning Methods: Application to Magnesium Sodium Silicate Glasses

Marco Bertani,[a, b] Alfonso Pedone,[a] Francesco Faglioni,[a] and Thibault Charpentier*[b]

In this work, we have applied the Kernel Ridge Regression (KRR) method using a Least Square Support Vector Regression (LSSVR) approach for the prediction of the NMR isotropic magnetic shielding ($\sigma_{iso}$) of active nuclei ($^{17}O$, $^{23}Na$, $^{25}Mg$, and $^{29}Si$) in a series of (Mg, Na)–silicate glasses. The Machine Learning (ML) algorithm has been trained by mapping the local environment of each atom described by the Smooth Overlap of Atomic Position (SOAP) descriptor with isotropic chemical shielding values computed with DFT using the Gauge-Included-Projector-Augmented-Wave (GIPAW) approach. The influence of different training datasets generated through molecular dynamics simulations at various temperatures and with different inter-atomic potentials has been tested and we demonstrate the importance of a wide exploration of the configurational space to enhance the transferability of the ML-regressor. Finally, the trained ML-regressor has been used to simulate the $^{29}Si$ MAS NMR spectra of systems containing up to 20000 atoms by averaging hundreds of configurations extracted from classical MD simulations to account for thermal vibrations. This ML approach is a powerful tool for the interpretation of NMR spectra using relatively large systems at a fraction of the computational time required by quantum mechanical calculations which are of high computational cost.

## 1. Introduction

Solid-state NMR spectroscopy has now established itself as among one of the most powerful experimental techniques for glass structure investigation.[1–11] The NMR fingerprint of an active nucleus strongly depends on its local environment and can therefore give fundamental information on the short and medium-range order in glass structures, like the coordination numbers (CN), network connectivity ($Q^n$: Q stands for quaternary species whereas n is the number of bridging oxygens linked to it), oxygen speciations (bridging and non-bridging among network former cations), and cations intermixing.

However, the amorphous nature of glasses makes the interpretation and deconvolution of NMR spectra a challenging task, possible only for systems with a simple composition since the observed NMR spectrum results from the topological and chemical distributions of the glass constituents. This induces a spectral broadening that leads to a strong overlapping of the contributions stemming from the different environments of the probed nucleus. As a mean to overcome these limitations, Molecular Dynamics (MD) simulations coupled with NMR DFT calculations have become a very important tool for interpreting NMR spectra.[12]

The approach consists of generating glass structures using classical MD simulations, optimizing the structures at the DFT level, and computing the NMR parameters using the gauge-including projector augmented wave (GIPAW)[13–19] method. Unfortunately, the high computational cost of the DFT-NMR calculations limits the application of this so-called MD-GIPAW approach to small system sizes (max ~1000 atoms),[12,20] in addition to the fact that it does not account for dynamical effects due to atomic vibrations at room temperature.[21]

Nowadays, Machine Learning (ML) for the prediction of NMR parameters has emerged as a promising way to overcome these limitations (and more generally those of DFT calculations).[22,23] ML techniques offer the opportunity to predict various properties with near DFT accuracy but saving orders of magnitude of computational time.[24] They were applied to organic molecular solids by Paruzzo et al[25] to predict NMR chemical shift and to determine the structure of molecules by comparing simulated and experimental spectra. The earlier applications of ML to predict chemical shifts in silicate glasses were provided by Cuny et al.[26] who trained a neural network to predict the 29Si isotropic chemical shift in silica glass and by Chaker et al.[27] who performed a detailed analysis of different atomic descriptors and ML methods for sodium silicate and sodium aluminosilicate glasses. Furthermore, Gaumard et al.[28] studied the performances of different ML algorithms for the reproduction of isotropic shielding values in zeolites, and Ohkubo et al.[29] developed a method to investigate the position of $^{133}Cs$ in clay's layers predicting its NMR chemical shift using ML.

[a] M. Bertani, A. Pedone, F. Faglioni
University of Modena and Reggio Emilia, Department of Chemical and Geological Sciences, via Campi 103, Modena, Italy

[b] M. Bertani, T. Charpentier
Université Paris-Saclay, CEA, CNRS, NIMBE, 91191 Gif-sur-Yvette, France
E-mail: thibault.charpentier@cea.fr

Supporting information for this article is available on the WWW under https://doi.org/10.1002/cphc.202300782

© 2024 The Authors. ChemPhysChem published by Wiley-VCH GmbH. This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.



© 2024 The Authors. ChemPhysChem published by Wiley-VCH GmbH

In this work, we develop and apply a Least-Square Support Vector Regression (LSSVR)[30–32] model (a sparse variant of Kernel Ridge Regression, KRR) to predict 29Si, $^{17}O$, $^{23}Na$, and $^{25}Mg$ NMR isotropic magnetic shieldings for a series of (Mg, Na)-silicate glasses. Using a Nyström approximation[33,34] of the Kernel matrix, LSSVR offers a significant dimensional reduction in data space, thus a thorough optimization of the hyperparameters (Kernel and descriptors parameters) is possible, as illustrated in this work.

Magnesium-containing silicates are very important components of geological melts[35,36] and find application in many fields,[37] from medicine[38–41] to technological systems.[42–46] The structural impact of magnesium in silicate glasses and its interaction with other elements are not yet fully understood as they can vary greatly with the chemical composition of the glass[44,47–49] leading to significant variation of properties such as viscosity, glass transition temperature, elastic properties, crystallization behavior, and chemical durability.

Therefore, the availability of fast and reliable ML models to predict NMR parameters of spin active nuclei in Mg-containing oxide glasses would allow to 1) compare spectra simulated using large structural models of statistical relevance with the experimental counterparts and 2) exploit data-driven reverse MD or Monte Carlo approaches to refine glass structures from experimental spectra.[50,51] The latter application will be explored in detail in future works.

In this work, five sodium-magnesium silicate glasses, whose composition is reported in Table 1, were studied and used to build a robust NMR database for machine learning training. The compositions were chosen based on the availability in literature[8,52] of experimental 29Si and/or 17O NMR spectra. However, since we have not yet trained an ML model to predict Electric Field Gradient (EFG) tensors of quadrupolar nuclei (here $^{23}Na$, $^{17}O$, and $^{25}Mg$) only the spectra of $^{29}Si$ nuclei will be reported in this work, while the performance on the $\sigma_{iso}$ prediction for $^{23}Na$, $^{25}Mg$, and $^{17}O$ are detailed in the ESI. The glasses are labeled as NMSX where N=Na, M=Mg, S=Si and X represents the MgO/Na2O ratio.

We investigated the impact of the training database composition on the reliability and transferability of the model including data from MD simulations using i) different interatomic potentials; ii) configurations extracted at various temperatures; and iii) including DFT-optimized structures, and training sub-models using data obtained from specific parts of the database and testing it on the complementary part. Finally, the ML model trained (and tested) on about 950 structures containing 400 atoms was applied to simulate NMR spectra of configurations containing up to 20000 atoms.

**Table 1.** Composition of the glasses studied in this work.[8,52]

| | %mol. MgO | %mol. $Na_2O$ | %mol. $SiO_2$ | $MgO/Na_2O$ |
|---|---|---|---|---|
| NMS0.11 | 3.3 | 30 | 66.7 | 0.11 |
| NMS0.33 | 8.3 | 25 | 66.7 | 0.33 |
| NMS1 | 16.7 | 16.7 | 66.7 | 1 |
| NMS1.25 | 25 | 20 | 55 | 1.25 |
| NMS2 | 30 | 15 | 55 | 2 |

## Computational Approach and ML Procedure

The machine learning procedure followed to predict NMR isotropic magnetic shieldings of spin-active nuclei in oxide glasses consists of several steps that are detailed in this section.

### Construction of the Database

The database of NMR/DFT calculations was built as follows. Ten independent melt-quench[53] classical MD simulations (the quench was performed by decreasing the temperature in steps of 100 K from 3000 K to 300 K, as detailed in Section S4 of the ESI), were performed for each glass composition given in Table 1. Configurations were then extracted from NVE (PMMCS and BMP potentials) or NVT (CS potential) trajectories, following the procedure reported in Sections S3 and S4 of the ESI, at 1000 K, 300 K, and 0 K (the latter is equivalent to an optimized structure with the MD potential). Because of the high computational cost of DFT-GIPAW calculations, most of the simulation boxes were limited in size to 400 atoms, but few larger 800 atoms models were generated to check the transferability of the LSSVR regressions. For assessing the direct application of the trained ML models to MD trajectories (for finite temperature simulations), we also generated larger MD models of 4000 and 20000 atoms (see Tables S1, S2, S3, and S4 of the ESI).

The MD simulations were carried out using DL_POLY_4 v5.0.0 package[54] using three different potentials: the PMMCS,[55] the BMP-shrm[56,57] (here referred to as BMP), and the CS.[58,59] The first two potentials are based on a rigid ionic model and the main difference between them is the presence, in the BMP, of a Si–O–Si three-body term and a Si–Si repulsive interaction, whereas the CS includes the oxygen polarizability through the shell-model.[60,61] A description of the three potentials is reported in Section S3 of the ESI.

Selected structures obtained at 1000 K, 300 K, and 0 K (of the 400 and 800 simulation boxes) were also relaxed at 0 K at the DFT level using CP2 K[62] (details are given in Section S5 of the ESI), first to observe the structural differences with the MD structures and, second, to enlarge the database for NMR predictions.

For simulations of larger systems containing 800, 4000, and 20000 atoms, only the BMP potential was used as the aim was to observe the applicability of the model trained on small systems to larger ones and to understand the effect of the system size on the simulated NMR spectra.

The calculations of the NMR parameters were carried out employing the PBE-GGA DFT functional,[63] using the VASP[62] code (version 5.3) which makes use of plane-wave pseudopotentials[64] and the Gauge Including Projector Augmented Wave (GIPAW) formalism.[13,18] The pseudopotentials parameters used are reported in Table S5 of the ESI (these pseudopotentials are the *standard* PAW potentials provided with the VASP package). Outputs of the DFT-GIPAW calculations were processed with the fpNMR package developed by T. Charpentier.[65] This code also performs statistical analysis of NMR data and their correlation with local structural properties.

To fix the 29Si $\delta_{iso}$ scale when comparing predicted $\sigma_{iso}$ with experimental data, the GIPAW data are transformed as[66]

$$\delta iso = \text{-}0.853^{*}(\sigma\text{-}315.9) \quad (1)$$


© 2024 The Authors. ChemPhysChem published by Wiley-VCH GmbH

14397641, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cphc.202300782 by University Modena, Wiley Online Library on [14/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

All NMR GIPAW calculations were performed at the Γ point with a kinetic energy cutoff of 550.0 eV. Here, a small Gaussian smearing value of 0.01 eV was used; this was found particularly necessary for the high-temperature structures where spurious partial occupation of the Fermi level was observed for some models. This could be removed either with a smaller smearing value or, alternatively and equivalently, by usage of a denser k-grid.

The dataset comprising all the structures is denoted DS_tot. Different sub-datasets were generated to better understand the influence of including different geometric and structural spaces on the accuracy and transferability of the machine learning prediction. These sub-datasets were created as follows:

- Temperature: one dataset for each temperature (0 K, 300 K, and 1000 K) including data from all the potentials (DS_MD-0 K, DS_MD-300 K, DS_MD-1000 K)
- Potential: one dataset for each potential including data at all temperatures (DS_MD-BMP, DS_MD-PMMCS, DS_MD-CS)
- DFT optimization: one dataset formed only by DFT optimized structures and one only with simple MD data (DS_DFT, DS_MD)
- Leave one composition out: datasets formed by all the compositions excluding one (DS-NMS0.11, DS-NMS0.33, DS-NMS1, DS-NMS1.25, and DS-NMS2).

The numbers of structures and atoms in each of these dataset (DS) are reported in Table 2.

Each sub-dataset was used to train an ML model that was then tested on the complementary part of the total database. For example, the model trained on DS_MD-1000 K was tested on 0 K and 300 K data, the model from DS_MD-PMMCS was tested on BMP and CS data, and so on.

### Local Environment Featurization: SOAP Descriptor Using Spherical Bessel Functions

Atom-centered descriptors (ACD) are the key parameters that control the performance of many ML schemes and form the inputs of the algorithm.[67,68] ACD is a vector encoding the local environment of each atom, constructed from the Cartesian coordinates of the central atom and its nearest neighbors within a user-defined cutoff radius (typically 5 Å).

A large number of different atomic descriptors have been developed in the last years[27,67–72] and this is still a subject of intensive research, see for example the recent works of Langer et al.[73] and Deringer et al.[74]

In order to guarantee a faithful representation of the structure, ACD must fulfill mathematical and physical constraints.[70] For the prediction of an isotropic quantity such as the total energy or, as in this work, the isotropic magnetic shielding (or equivalently the isotropic chemical shift obtained from Equation (1)): (i) they must be invariant with respect to rotations, translations, reflections and permutations of the neighboring atoms; (ii) they must vary continuously with the atomic environment, must be differentiable and unique (no identical descriptors for different input samples).

In this work, we used the Smooth Overlap of Atomic Positions (SOAP)[67] descriptors which are among the most popular and best-performing for NMR predictions.[27,75–78] The choice of this descriptor was based on a previous investigation by one of us.[27] where also the Behler-Parrinello Symmetry Functions and the Angular Radial Distribution Functions were tested for the prediction of the $\sigma_{iso}$ of sodium silicate and aluminosilicate glasses, but lead to worse results.

SOAP descriptors are based on a smooth representation of the atomic density $\rho_i(r)$ around a central atom *i* within a cutoff radius $r_c$

$$\rho_i(r) = \sum_{j \in N_j} f_c(r_{ij}) g(r - r_{ij}) \tag{2}$$

where $f_c(r)$ is a smooth cutoff function that brings the density to zero outside the cutoff radius and (equivalently) the summation can therefore be limited to the neighboring atoms of the central atom. Here, the cutoff function as proposed by Behler and Parrinello[68] was used: $f_c(r) = \frac{1}{2} cos\left(\pi \frac{r}{r_c}\right)$ when $r \leqslant \rho_\psi$ and 0 otherwise. distance between the two atoms is denoted as *rij* and $g(\boldsymbol{r})$ is a 3D Gaussian function (note that bold notation is used for vectors):

$$\rho_i(\boldsymbol{r}) = \sum_{j \in N_j} e^{\frac{-(r-r_{ij})^2}{2\sigma^2}} f_c(r_{ij}) \tag{3}$$

**Table 2.** Sub-dataset description: MD temperature or glass composition, number of structures and atoms.

| | N° | Structures | Si | Mg | Na | O |
|---|---|---|---|---|---|---|
| DS_tot | | 900 | 78480 | 21420 | 53640 | 205200 |
| DS_MD-T | 0 K | 150 | 13080 | 3570 | 8940 | 34200 |
| | 300 K | 150 | 13080 | 3570 | 8940 | 34200 |
| | 1000 K | 150 | 13080 | 3570 | 8940 | 34200 |
| DS_MD-FF | BMP | 150 | 13080 | 3570 | 8940 | 34200 |
| | PMMCS | 150 | 13080 | 3570 | 8940 | 34200 |
| | CS | 150 | 13080 | 3570 | 8940 | 34200 |
| DS-DFT/MD | DFT | 450 | 39240 | 10710 | 26820 | 102600 |
| | MD | 450 | 39240 | 10710 | 26820 | 102600 |
| DS-NMSX | NMS0.11 | 760 | 62460 | 20700 | 39240 | 165240 |
| | NMS0.33 | 760 | 62100 | 19440 | 41400 | 164340 |
| | NMS1 | 760 | 61380 | 17100 | 45000 | 162360 |
| | NMS1.25 | 760 | 64080 | 14940 | 43200 | 164700 |
| | NMS2 | 760 | 63900 | 13500 | 45720 | 164160 |


© 2024 The Authors. ChemPhysChem published by Wiley-VCH GmbH

14397641, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cphc.202300782 by University Modena, Wiley Online Library on [14/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

For each kind of atomic neighbor, here denoted $\mu$, the atomic density is expanded on radial basis functions $\chi_{nl}(r)$ and spherical Harmonics $Y_{lm}(\hat{r}) = Y_{lm}(\theta,\phi)$ as follows:

$$\rho_i(\boldsymbol{r}) = \sum \rho_\mu(\boldsymbol{r}) = \sum_{\mu nlm} c^{\mu}_{nlm}\chi_{nl}(r)Y_{lm}(\hat{r}) \tag{4}$$

where $0 \leqslant \nu \leqslant_{\mu\alpha\xi}$ and$0 \leqslant \lambda \leqslant \Lambda_{\mu\alpha\xi}$.

In this work, the spherical Bessel functions $\chi_{nl} = j_l(q_{nl}r)$ were used because of their mutual orthogonality and non-discontinuity at the origin.[24,76] The $c^{\mu}_{nlm}$ values were computed using an in-house code (developed by T. Charpentier) using a real representation of the spherical harmonics (RSH) and taking as an input a structure file (atomic coordinates of all atoms and lattice parameters). Outputs are the descriptors of all atoms calculated with the chosen *hyper*-parameters, here represented by the quadruplet $(r_c, n_{rad}, L_{max}, \sigma_{SOAP})$ where $\sigma_{SOAP}$ is the width of the Gaussian function $g(\boldsymbol{r})$.

The descriptors[67] $p^{\mu\nu}_{nn'l}$are then obtained from the rotational invariant given by the so-called powerspectrum:

$$p^{\mu\nu}_{nn'l} = \sum_m c^{\mu}_{nlm}c^{\nu}_{n'lm} \tag{5}$$

### Least Square Support Vector Regression – Kernel Ridge Regression (LSSVR-KRR)

Similar to Kernel Ridge Regression (KRR), support vector machines (SVM) are a family of techniques that are able to model nonlinear relationships and to deal with high dimensional input vectors.[30–32,79] They were initially developed by Vapnik[80] as a binary classification tool and were subsequently extended to regression tasks and subsequently sped up via a least-squares approach.[30–32] In this family of kernelized methods, a kernel function replaces the Euclidean scalar product (used in linear ridge regression) for comparing two vectors of descriptors (and thus the similarity between two environments). Let $y \cdot x$ be the scalar product, the kernelized scalar product is then defined as $K(x,y) = \phi(y) \cdot \phi(x)$ where the existence of the non-linear map $x \rightarrow \phi(x)$ is mathematically guaranteed by the Mercer Theorem for positive definite Kernel (such as the Gaussian Kernel).[81] The kernel therefore implicitly maps the input data $x$ into a space of higher dimension represented by $\phi(x)$ (possibly of infinite dimension in the case of the Gaussian kernel) where a *simple* linear regression can then be applied.[82] As during the procedure only scalar products are needed, the non-linear featurization is done implicitly, i.e. without the need to know explicitly $\phi(x)$. This is known as the *"kernel trick"*[83] and thus many *linear* algorithms can be *kernelized*.[84]

In this work, the standard Gaussian kernel was used:

$$K(x,y) = exp\left\{-\frac{\|x-y\|^2}{2\sigma^2}\right\} \tag{6}$$

Denoting $\sigma_{iso}(\chi)$ the isotropic magnetic shielding of the local environment $\chi$ represented by the SOAP vector $p$, the KRR aims at predicting $\sigma_{iso}(\chi)$ from the (training) database values $\{\sigma_{iso}(\chi_j), p(\chi_j)\}_{j=1,N}$with the *linear* regression:

$$\sigma_{iso}(\chi) = \sum K\big(\boldsymbol{p}(\chi), \boldsymbol{p}(\chi_j)\big)\alpha_j \tag{7}$$

where the regression parameters $\alpha_j$are calculated by solving the linear system generated from the Equation (7) applied to the training set. In matrix form, the solution is given by:

$$\boldsymbol{\alpha} = (K + \lambda I)^{-1}\boldsymbol{\sigma_{iso}} \tag{8}$$

The idea underlying the LSSVR is to resolve Equation (8) using a (much) smaller set of data (referred to as landmark or inducing points, denoted here $\xi$) using the Nyström[33,34] approximation of the Kernel matrix $K_{\chi\chi} \approx C_{\chi\xi}W^{-1}_{\xi\xi}C^{t}_{\xi\chi}$where $W_{\xi\xi}$is the (small) kernel matrix spanned by the landmark points $(K(p(\xi_i), p(\xi_j)))$ and $C_{\chi\xi}$is the projection kernel matrix between the landmark/inducing points and the training data $(K(p(\chi_i), p(\xi_j)))$. The diagonalization of $W_{\xi\xi}$then allows the system Equation (8) to be solved much more efficiently, see refs [33,34,84] for a detailed description of the algorithm we implemented. The landmark points were obtained from an Incomplete Cholesky Decomposition[33,85,86] (ICD) of the full kernel matrix *K*, using the LAPACK library.[87] These points can be seen as being the most representative (and less redundant) in the database for providing an informative sparse representation of the kernel matrix *K*. We found that alternative approaches such as Kmeans++[88] or entropy maximization[89] for choosing $\xi$did not perform better than ICD, with a much higher computational cost.

The applied ML algorithm will be referred to as LSSVR-KRR (from Least Square Support Vector Regression–Kernel Ridge Regression) hereafter. Albeit other regression methodologies such as Neural Networks[90] exists, these are more time consuming and thus were avoided in this work. Moreover, the advantage of dimensional reduction of the LSSVR method will allow the extension of the model to tensorial quantities such as the computation of electric fields gradients that require tensorial descriptors (λ-SOAP)[24] and thus the handling of large block-matrix Kernels. This will be the focus of future works.

### Optimization of the Hyperparameters, Training, and Validation of the LSSVR-KRR

The LSSVR-KRR model was trained and optimized using a 5-fold Cross-Validation (CV) technique which proceeds as follows. The data are first randomly shuffled and divided into five parts. Three parts are used for solving Equation (8) (the training set), one for optimization of the hyperparameters (the validation set), and the last for assessing the transferability to unseen data (the testing set) and computation of the error. This procedure was repeated so that each point of the database set is in the testing set at least once (or more time by repeating the initial random shuffling). To speed up the optimization of the hyperparameters, the whole procedure (hereafter referred to as training) was performed on the DS_MD–T (T = 0, 300, 1000 K) databases independently. Each led to very close optimal hyperparameters.

The number of radial basis functions ($n_{rad}$), the cut-off radius ($r_{cut}$), and the $\sigma$ value of Equation (3) ($\sigma_{SOAP}$) are the SOAP's parameters that need to be optimized. $L_{max}$ was chosen according to the previous work of Chaker et al.[27] and set to $L_{max} = 4$ ($L_{max} = 3$ led to worse results and no improvement was yielded by $L_{max} = 5$). As for the LSSVR-KRR, the $\sigma$ of the Gaussian Kernel in Equation (6) ($\sigma_{Kernel}$ from hereafter) and the ridge parameter $\lambda$ in Equation (8) ($\lambda_{ridge}$ from hereafter) were optimized, as well as the number of landmark points (Nyström size, $r_{size}$).

During the 5-fold CV described above, taking into consideration the mean absolute error (MAE) or the root mean square error (RMSE) calculated on $\sigma_{iso}$, values of the validation set did not led to change the results for the different nuclei studied ($^{17}$O, $^{23}$Na, $^{25}$ Mg, and $^{29}$Si) so selecting these yields robust regressors.


© 2024 The Authors. ChemPhysChem published by Wiley-VCH GmbH

14397641, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cphc.202300782 by University Modena, Wiley Online Library on [14/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

A systematic grid-search approach was applied, following the scheme reported in Figure 1, to obtain the best parameters. Specifically, a first grid was obtained varying $\lambda_{ridge}$ so that $log_{10}(\lambda_{ridge})$ ranges from −12 to −1 with steps of 1, while, for *σKernel*, *log2(σKernel)* = *0.25, 0.5, 1, 2, 4, 6, 8, 10*. The optimal region was found, for all the nuclei, to *log2(σKernel)* from 2 to 8 and $log_{10}(\lambda_{ridge})$ from −10 to −8 and in this region, the chosen values were *log2(σKernel)* = *5*, and $log_{10}(\lambda_{ridge}) = -8$ (See Figure S1 of the ESI).

Once the best KRR hyperparameters were found, a second grid search was performed (in an outer loop) to optimize the SOAP parameters and the number of landmark points forming the support vectors of the LSSVR. In particular, the *σSOAP* parameter which controls the broadening of the Gaussian smoothing function, was varied at 3 levels with values of 0.5, 1, and 2. As for the other parameters, the tested values are $r_{cut}$ = *3, 4, 5*, and *6 Å*, *nrad* = *4, 6, 8, 10*, and *12* radial basis functions, and *rsize* = *100, 200, 400, 800, 1200, 1600*, and *2000* landmark points. Figure 2 reports the MAE of the

**1° GRID SEARCH ($\sigma_{kernel}$, $\lambda_{ridge}$)**

For $log_2(\sigma_{kernel})$ in [0.25, 0.5, 1, 2, 4, 6, 8, 10]:
  For $log_{10}(\lambda_{ridge})$ in range (-12, -1):
    Train LSSVR-KRR model using 5-CV
    Test trained model on Test-Set
    If *MAE* < *min_MAE* and *RMSE* < *min_RMSE*:
      *min_MAE* = *MAE*
      *min_RMSE* = *RMSE*
      $\sigma_{kernel\_OPT} = \sigma_{kernel}$
      $\lambda_{ridge\_OPT} = \lambda_{ridge}$

**2° GRID SEARCH ($\sigma_{SOAP}$, $r_{cut}$, $n_{rad}$, $r_{size}$)**

For $\sigma_{SOAP}$ in [0.5, 1, 2]:
  For $r_{cut}$ in [3, 4, 5, 6]:
    For $n_{rad}$ in [4, 6, 8, 10, 12]:
      For $r_{size}$ in [100, 200, 400, 800, 1200, 1600, 2000]:
        Train LSSVR-KRR model using 5-CV
        Test trained model on Test-Set
        If *MAE* < *min_MAE* and *RMSE* < *min_RMSE*:
          *min_MAE* = *MAE*
          *min_RMSE* = *RMSE*
          $\sigma_{SOAP\_OPT} = \sigma_{SOAP}$
          $r_{cut\_OPT} = r_{cut}$
          $n_{rad\_OPT} = n_{rad}$
          $r_{size\_OPT} = r_{size}$

**Figure 1.** Pseudocode representation of the optimization process via a systematic grid (or loop) search of the hyperparameters.

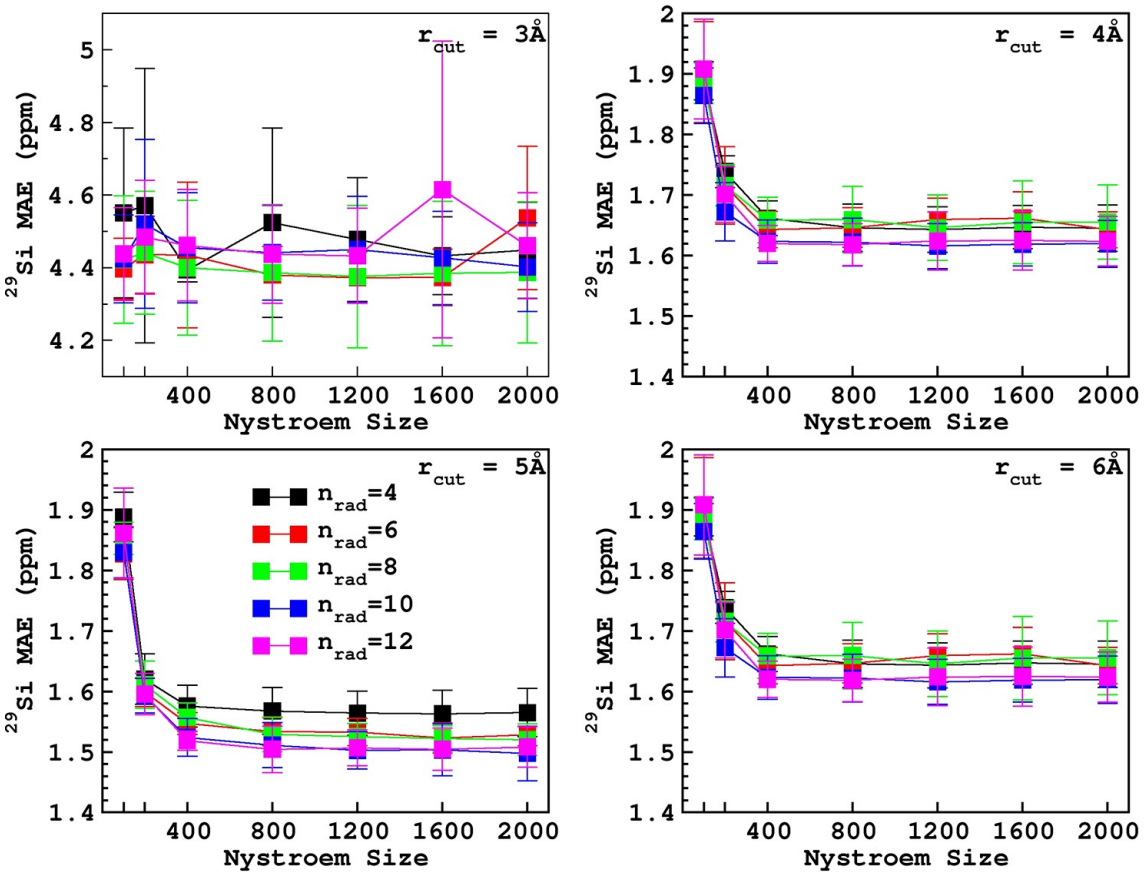



**Figure 2.** Variation of the MAE for the prediction of the $^{29}$Si NMR $\sigma_{iso}$ in the MD-300 K set with respect to the Nyström size (LSSVR), number of radial basis functions ($n_{rad}$) and for various values of the SOAP cutoff radius vales. $L_{MAX}$ = 4 for all points. The error bars describe the uncertainty of the MAE (standard deviation values) derived from the different validation sets of the 5-fold CV.


© 2024 The Authors. ChemPhysChem published by Wiley-VCH GmbH

14397641, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cphc.202300782 by University Modena, Wiley Online Library on [14/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

predicted $^{29}$Si NMR shielding of structures at 300 K (the same was done for the other temperatures and nuclei and results are reported in Figures S2 to S6 of the ESI) varying $r_{size}$, at different $n_{rad}$ and $r_{cut}$. For the sake of simplicity, this figure refers to data obtained using the level of $\sigma SOAP = 0.5$ (which gives the best results) but analogous plots were obtained for all three $\sigma SOAP$ values tested, giving the same optimal parameters. The optimal parameters obtained from the described procedure are $r_{cut} = 5$ Å for all the nuclei except $^{17}$O, for which 6 Å gives better results, $n_{rad} = 8$, and $r_{size} = 800$ for all the studied nuclei, except $^{25}$Mg, for which the optimal $r_{size}$ is *600*.

It is evident that a cut-off radius lower than 4 Å and a number of landmark points lower than 400 lead to a poor description of the silicon environment. At the same time, the use of more than 800 landmark points and 8 radial basis functions increases the computational cost but does not improve the accuracy significantly.

Figure 3 reports the predicted $\sigma_{iso}$ against DFT values on the test set of the DS_tot database made with the model trained at the optimized conditions for the studied nuclei.

The obtained MAE and RMSE confirm the good reproduction of the DFT data with the trained LSSVR-KRR model. The percentage of error with respect to the range of $\sigma_{iso}$ values observed for each element was also calculated to give an idea of the impact of the estimation error on the prediction of the spectra. The resulting % MAE are 1.9, 1.8, 3.3, and 2.0% for $^{17}$O, $^{23}$Na, $^{25}$Mg, and $^{29}$Si, respectively, while the %RMSE are 2.5, 2.3, 4.2, and 2.6%. These values show that, even if $^{17}$O seems not to be estimated as accurately as $^{29}$Si and $^{23}$Na at first glance (from MAE and RMSE), this is mainly due to a larger domain of $\sigma_{iso}$ as the percentage error is similar to that of $^{29}$Si and $^{23}$Na (also consistent with the need of a large cutoff radius values). For $^{25}$Mg, the %errors are slightly larger, but still acceptable considering the poor resolution of experimental $^{25}$Mg NMR MAS spectra.[1]

## 2. Results and Discussion

### 2.1. Transferability to Larger Structures

The regressors, trained using dataset generated with structures containing 400 atoms, were used to simulate the 29Si MAS NMR spectrum of the NMS1.25 and NMS2 glasses containing 800 atoms obtained with the BMP potential and relaxed to the DFT level and compared with experiments in Figure 4. The DFT data of 800 atom systems were not included into the training database but only used for validation. Since the DFT NMR calculations on such a large system are quite expensive only one simulation for each glass was performed.

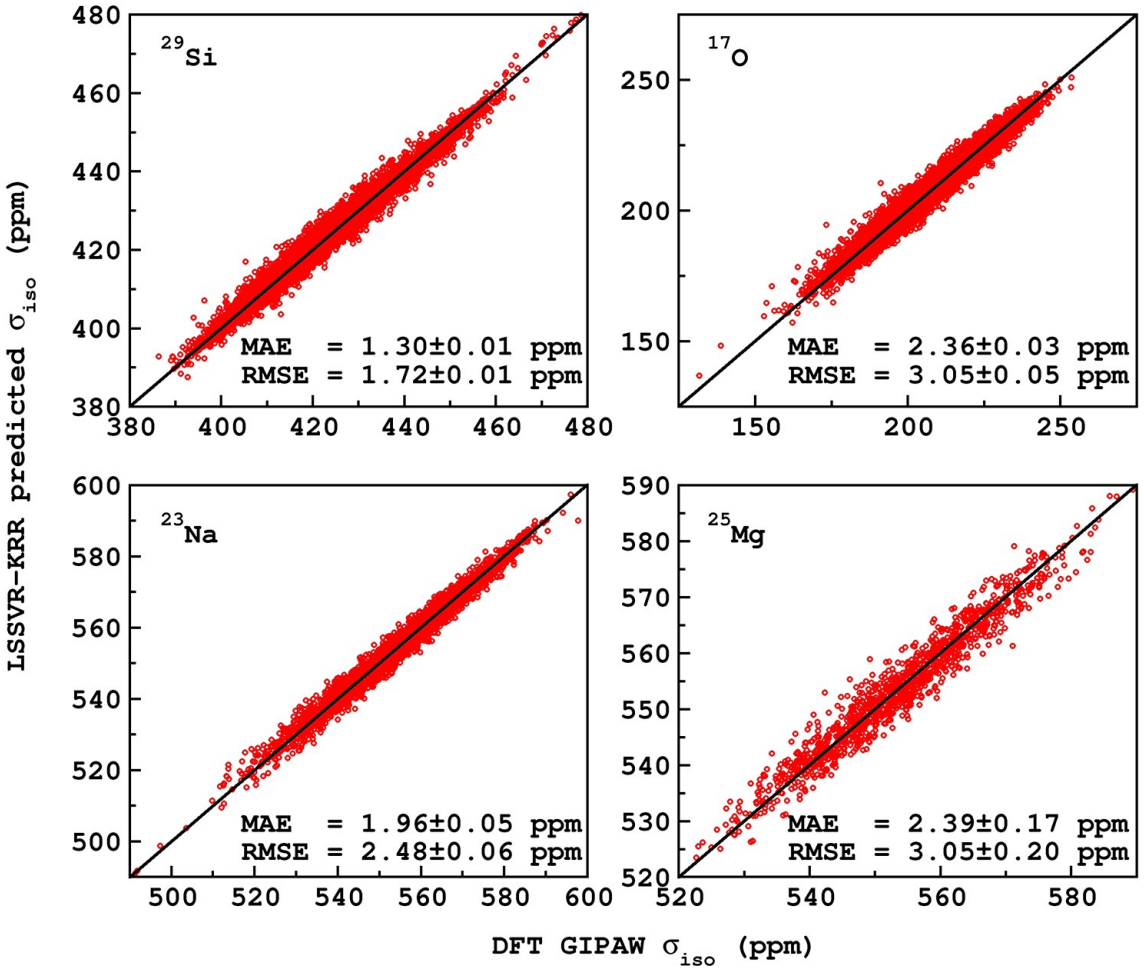



**Figure 3.** Scatter plot reporting the LSSVR-KRR predicted versus DFT $\sigma_{iso}$ of the studied nuclei (DS_tot dataset).


© 2024 The Authors. ChemPhysChem published by Wiley-VCH GmbH

14397641, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cphc.202300782 by University Modena, Wiley Online Library on [14/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

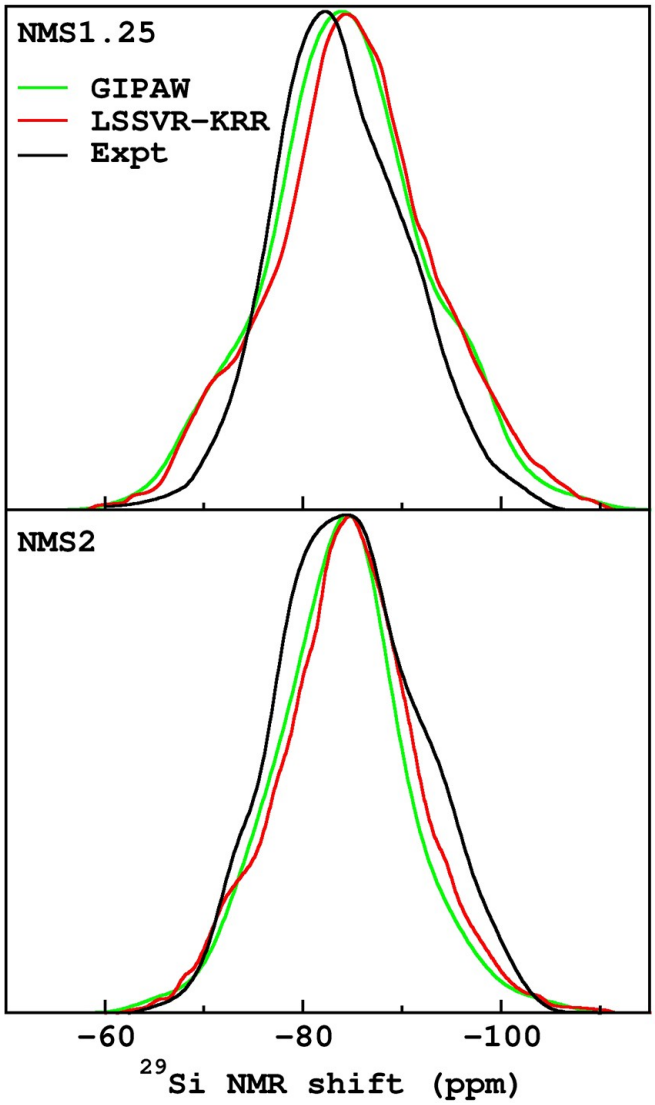



**Figure 4.** Comparison of experimental $^{29}$Si MAS NMR spectra of NMS1.25 and NMS2 glasses with DFT GIPAW calculations and LSSVR-KRR prediction on 800 atoms structures relaxed at the DFT-0 K level.

The spectra predicted with the LSSVR-KRR model nearly perfectly match the one obtained using the GIPAW calculations, and only small discrepancies are observed. The cosine similarity parameter between the GIPAW and LSSVR-KRR spectra, defined as $(A \bullet B)/(\|A\| \bullet \|B\|)$ where A and B represent vectors of the GIPAW and LSSVR-KRR spectral values at the same shifts, have been calculated and resulted 0.996 and 0.994 for NMS1.25 and NMS2, respectively. This proves that the model can be trustfully used to simulate spectra of larger structures with GIPAW accuracy. This also validates the choice of optimal small models (400 atoms) for building the database.

### 2.2. Impact of the Interatomic Potentials

To understand the impact of the potentials, we trained one LSSVR-KRR model on one dataset (DS_MD-BMP, DS_MD-PMMCS, or DS_MD-CS) and tested it on the (two) other ones. A more complete discussion of the structural properties predicted with the three potentials can be found in Section S7 of the ESI.

The scatter plots ($\sigma_{ISO}^{DFT}$ *vs* $\sigma_{ISO}^{ML}$) of these sub-models are reported in Figure 5.

It is evident that a model trained on a single potential does not reliably reproduce data for structures obtained with another. This is clearly due to the differences in Si–O distances and Si–O–Si bond angles (and other structural features) and $\sigma_{iso}$ distributions between the MD structures, as reported in Figure 6 (using NMS0.33 as a representative example). This shows the limitations in the transferability from one potential to another of LSSVR-KRR (and in general of ML) regressors even if it must be admitted that results are still qualitatively acceptable (for a coarse approach).

This is particularly true for CS potential, for which the prediction of CS test data gives a very narrow scatter plot with the lowest MAE (1.08 ppm) and RMSE (1.39 ppm) values but the prediction of BMP (MAE = 2.94, RMSE = 12.42 ppm) and PMMCS (MAE = 2.19, RMSE = 8.42 ppm) ones is very poor.

The models trained on PMMCS and BMP datasets (DS_MD-PMMCS and DS_MD-BMP) give slightly worse predictions on their own test sets with respect to CS (DS_MD-CS), but they show better, even if still poor, transferability to other potentials. This is probably due to the broader Si–O–Si BAD and Si[Qn] distribution obtained with the PMMCS and Si–O RDF obtained with the BMP potentials (see Figures S7 and S8 of the ESI), which leads to a better sampling of the feature space of the other potentials.

Figure 6 shows the distribution of the Si–O–Si angle with respect to the Si–O distance (panel a) and $^{29}$Si isotropic magnetic shielding (panel b).

From Figure 6, it is evident that none of the potentials completely encompasses all the structural feature's distributions of the other ones, so they can only be used to predict data originated with the same potential, as confirmed, in Figure 7, by the $^{29}$Si MAS NMR spectra of NMS0.33 glass obtained with the three potentials from DFT calculations.

Each spectrum has a region not covered by the others, making the prediction with models trained on different potentials not reliable, as part of the input is out of the training domain.

Training a model with the data from all the potentials (only MD structures not DFT relaxed) leads to an MAE of 1.4 ppm and RMSE of 1.9 ppm which is, except for BMP, worse than the model trained and tested on data from the same potential but always better than the case of models tested on different ones.

In general, it is evident that the inclusion in the database of different kinds of structures that allow the exploration of larger configurational space gives higher transferability of the LSSVR-KRR model. This comes at the price of lower accuracy for specific structures, for which the best model is the one trained only on data consistent with them.


© 2024 The Authors. ChemPhysChem published by Wiley-VCH GmbH

14397641, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cphc.202300782 by University Modena, Wiley Online Library on [14/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

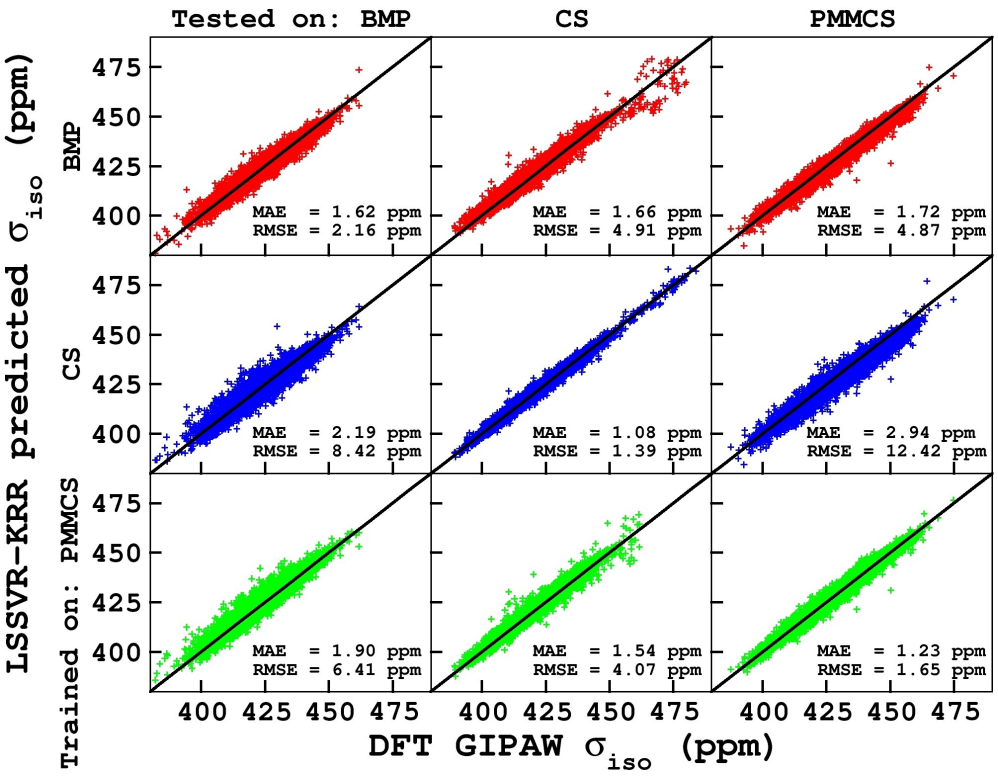



**Figure 5.** Scatter plots showing LSSVR-KRR predicted versus DFT $^{29}$Si $\sigma_{iso}$ of sub-models trained and tested on data from different potentials.

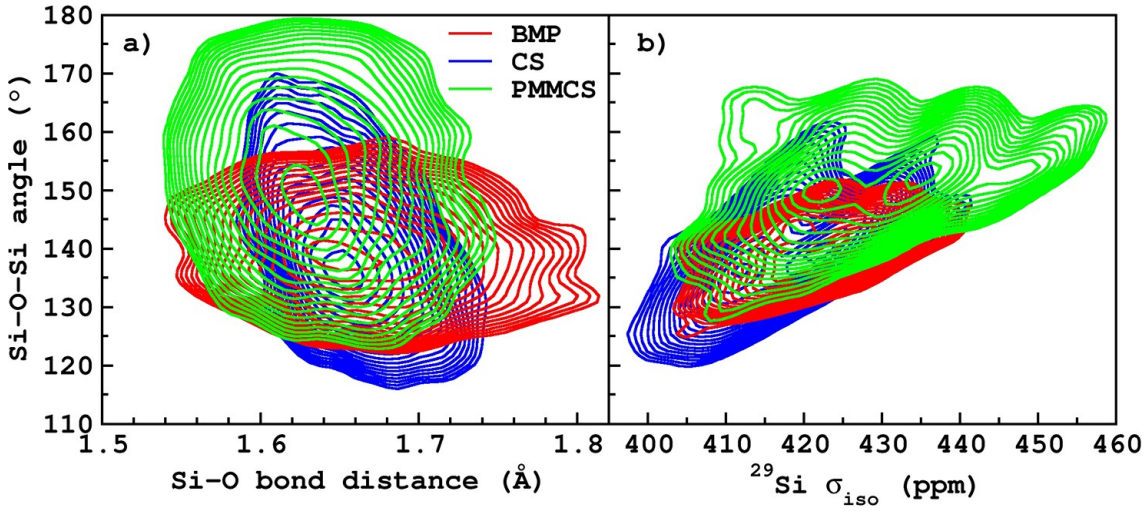



**Figure 6.** Contour plot showing the distribution of the Si–O–Si angle with respect to the Si–O distance (a) and isotropic magnetic shielding (b) obtained with the studied potentials.

### 2.3. Effect of the Temperature

Following the approach used for the interatomic potentials, models trained at each temperature were tested on their own test sets and on the other temperatures (gathering all FF data at a single temperature)

Figure 8 shows that models trained with low-temperature structures can only be used to predict low-temperature data as the error at high temperatures is very large (MAE = 3.5 ppm, RMSE = 20.18 ppm for T = 1000 K). Increasing the temperature leads, obviously, to the exploration of larger feature spaces that are not included in the low-temperature data. Instead, the models trained at high-T predict the shieldings of the structures extracted at low temperature with the same accuracy as the ones at high-T (MAE = 1.39 and RMSE = 3.12 ppm when testing on 0 K dataset and training with 1000 K dataset whereas MAE = 2.33 ppm and RMSE = 3.00 ppm when testing on 1000 K dataset and training on the 1000 K dataset).


© 2024 The Authors. ChemPhysChem published by Wiley-VCH GmbH

14397641, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cphc.202300782 by University Modena, Wiley Online Library on [14/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

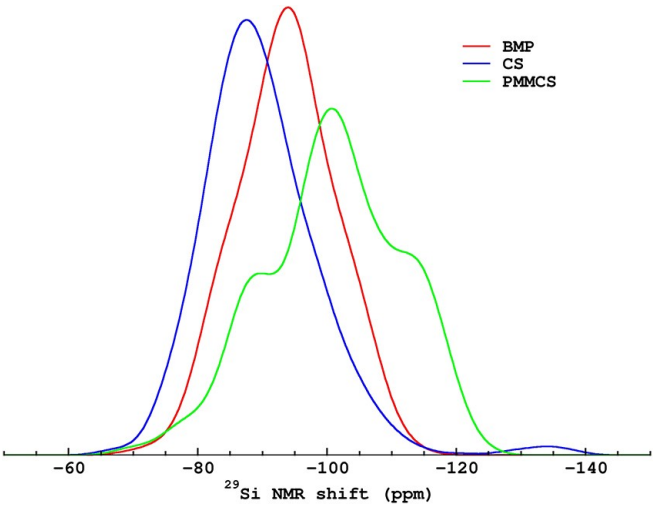



**Figure 7.** $^{29}$Si MAS NMR spectra of NMS0.33 glass at 300 K obtained with the three considered potentials.

This seems to demonstrate that training a model only on high T data can be enough to predict low T ones because the structural feature distributions explored at high temperatures encompass completely the low temperatures ones. This is indeed shown in Figure 9 that reports the Si–O–Si angle distribution with respect to the Si–O distance and $^{29}$Si isotropic magnetic shielding obtained at different temperatures. A detailed analysis of the Si[Q$^n$], Si–O–Si bond angle distribution, and Si–O partial radial distribution function is reported in Figures S9 and S10 of the ESI.

It can be noted that, while the angle vs distance distributions (panel a) obtained at high temperatures completely include the ones obtained at lower temperatures, the same cannot be stated for angle vs $\sigma_{iso}$ where a small region of low-temperature data is not included in the high temperature one. It is possible that other parameters not represented here affect the $\sigma_{iso}$ value explaining the non-overlapping region.

Figure 8 shows that even if the prediction of $^{29}$Si $\sigma_{iso}$ at 0 K with a model trained at 1000 K is of good quality, it provides MAE and RMSE values higher than the ones obtained with the 0 K model. The high RMSE indicates that higher temperature yields to a model more prone to generate outliers. This less accurate prediction is probably due to the fact that the low-energy structures are less sampled at high-T and thus less populated in the 1000 K dataset with respect to the 0 K dataset since the number of points in the different datasets are the same.

In conclusion, the inclusion of high-temperature data in the database seems important to access a broader configurational

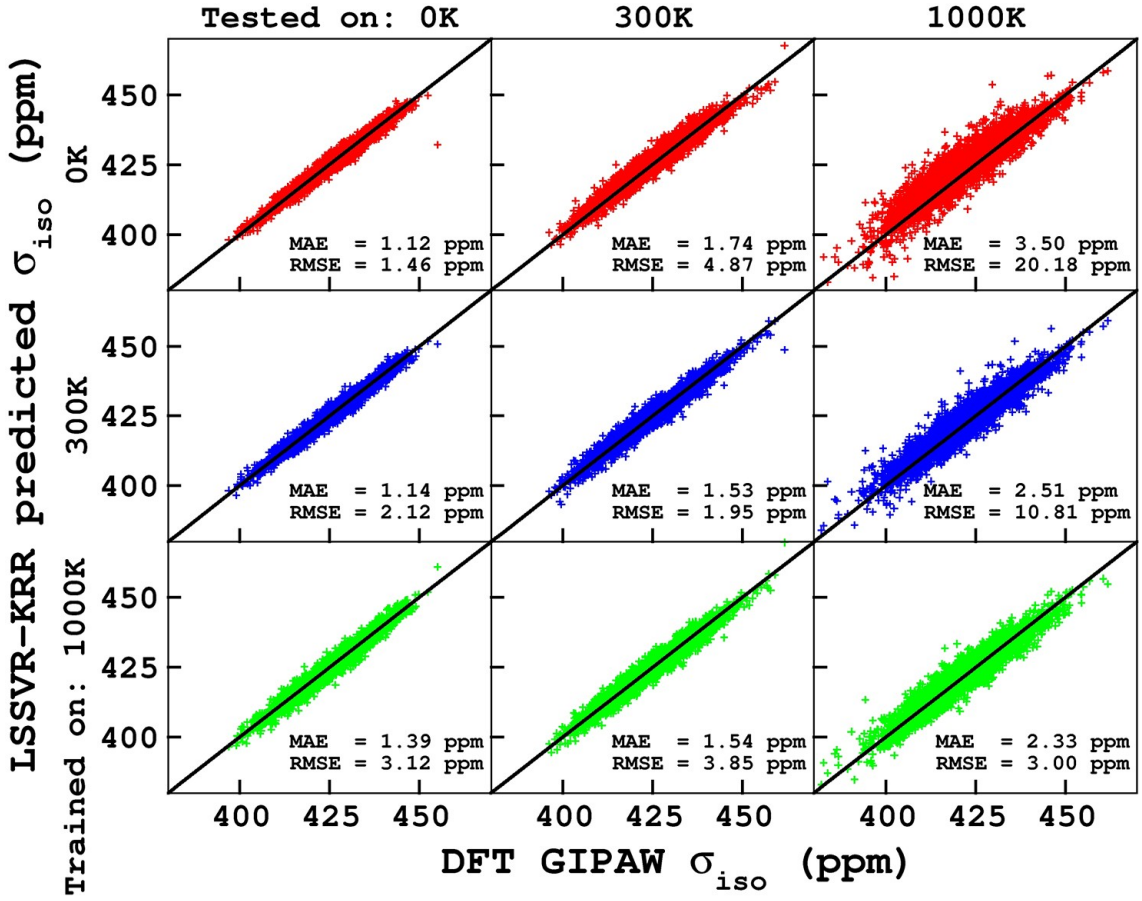



**Figure 8.** Scatter plots showing LSSVR-KRR predicted versus DFT $^{29}$Si $\sigma_{iso}$ of sub-models trained and tested on data from different temperatures.


© 2024 The Authors. ChemPhysChem published by Wiley-VCH GmbH

14397641, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cphc.202300782 by University Modena, Wiley Online Library on [14/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

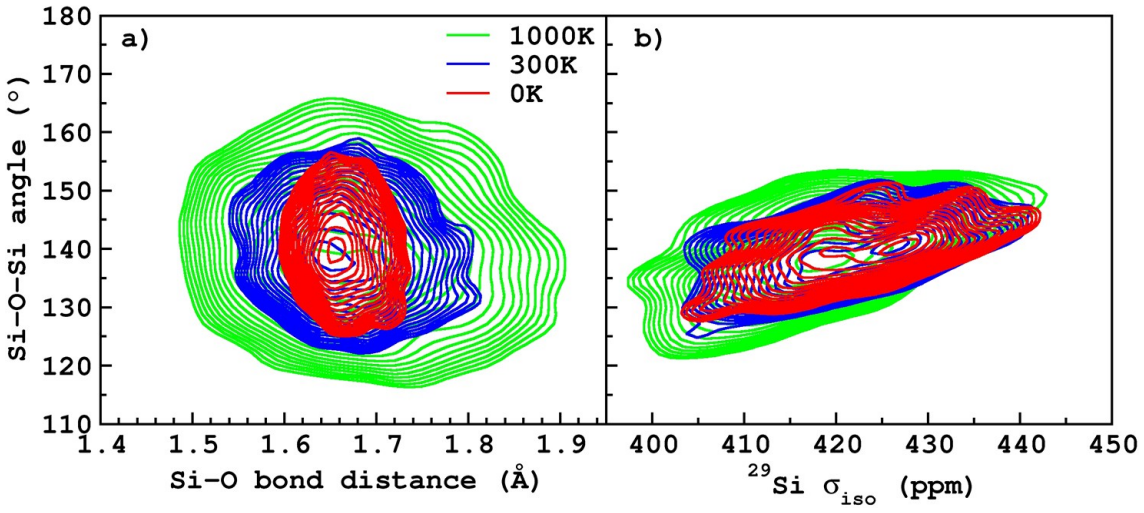



**Figure 9.** Contour plot showing the distribution of the Si–O–Si angle with respect to the Si–O distance (a) and isotropic magnetic shielding (b) obtained at the considered temperatures.

space but probably more low-energy structures should be included if the target is to predict shielding at low temperature.

### 2.4. Other Sub-Dataset Models

To investigate the transferability of a model to composition not included of the training set and the effect of DFT relaxation, LSSVR-KRR models were trained on the DS-MD/DFT and DS-NMSX subdatasets and applied to the complementary part of the total database, adopting the same approach used for interatomic potentials and temperature.

A good predictive power is found for all excluded compositions (MAE = 1.25–1.34 ppm, RMSE = 2.83–3.19 ppm) when training on the DS-NMSX sub-datasets.

The DS-DFT database explores a narrower configurational space and a model trained on it shows very accurate prediction (MAE = 0.93 ppm, RMSE = 1.19 ppm) for its own test set but fails when applied to MD structures (MAE = 3.76 ppm, RMSE = 28.14 ppm). In contrast, training on DS-MD ensures good predictions for both the DFT and MD datasets.

A detailed analysis is reported in Section S8 of the ESI.

### 2.5. Prediction of NMR Shifts for Large Models

The LSSVR-KRR model trained on the DS-tot dataset was used to compute the 29Si $\sigma_{iso}$ of glass structural models with up to 20000 atoms and to generate the corresponding NMR spectra (Equation (1)).

Figure 10 reports the 29Si MAS NMR spectra of NMS1.25 and NMS2 glasses obtained from the LSSVR-KRR prediction on structures containing 400, 800, 4000, and 20000 atoms, obtained with the BMP potential, compared with experiments.[8,52] The prediction was based on the 300 K MD simulation, averaging over 300 structures (corresponding to 150 ps) extracted from the trajectory. This allowed us to average possible thermal fluctuations[21] that could be present at 300 K and to investigate the evolution of the 29Si $\sigma_{iso}$ of the atoms during the trajectory as reported in Figure 11.

The quality of the simulated spectra, in terms of both peak position and shape, improves when the number of atoms increases in the simulation boxes and seems to converge for more than 4000 atoms, in particular for NMS2 glass. Since the GIPAW-DFT approach provides errors on the isotropic chemical shifts of 29Si in silicate crystals[refs] and the ML model reproduce the DFT shieldings with errors of the same magnitude we think that the discrepancies between the simulated and experimental spectra are probably due to the classical interatomic potential and the fast quench-rate used in MD simulations. Usually, to limit this discrepancy, a DFT 0 K optimization is applied to the structures used for NMR simulation which strongly improves the position of the spectra (see also Figure 4) maintaining the connectivity originated from the MD but optimizing the bond distances and angles at the DFT level. This was not applicable to this case because the number of atoms used are out of the DFT possibilities.

Using the LSSVR-KRR regressor, and in the perspective of studying thermal fluctuations, the time-evolution of the silicon-29 were calculated at 300 K as shown in Figure 11. The $\sigma_{iso}$ value calculated for each silicon atom fluctuates in a range of almost 10 ppm during the trajectory. None of the atoms gives outliers or a systematic drift of the $\sigma_{iso}$ value during the simulation allowing to trustfully improve the prediction by averaging the signals given by atoms during the MD evolution.

It is interesting to remark the computational efficiency of the LSSVR-KRR method with respect to the GIPAW-DFT calculations. In fact, the computation of the NMR data with DFT for one structure comprising 400 and 800 atoms required 1 and 4 hours using 48 CPUs, respectively. The computation of the isotropic chemical shifts of a trajectory including 300 structures for the 400 atoms systems with the trained ML model took


© 2024 The Authors. ChemPhysChem published by Wiley-VCH GmbH

14397641, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cphc.202300782 by University Modena, Wiley Online Library on [14/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

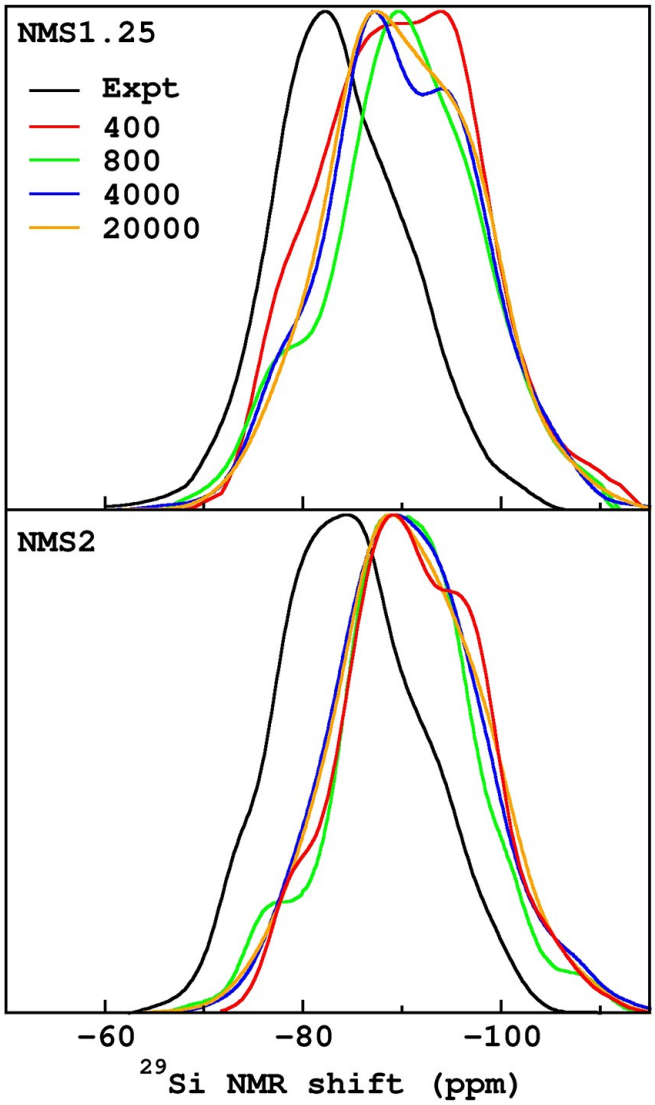



**Figure 10.** Comparison between experimental[8,52] 29Si MAS NMR spectra of NMS1.25 and NMS2 glasses and the ones obtained using LSSVR-KRR model on structures containing number of atoms from 400 to 20000.

10 minutes (0.2 s per structure) on 1 CPU. For the 800, 4000 and 20000 atoms systems the computational time were 9, 14 and 27 seconds per structure, respectively.

The LSSVR-KRR model can predict NMR parameters from simulation boxes that contain a number of atoms out of reach for DFT calculations with very low computational cost and high accuracy. This allows limiting the expensive DFT-GIPAW calculations to simulation boxes containing few atoms and to extend the NMR simulations to larger systems using ML models. Furthermore, the LSSVR-KRR prediction can be highly parallelized in a very efficient way as each calculation is independent, giving excellent scalability.

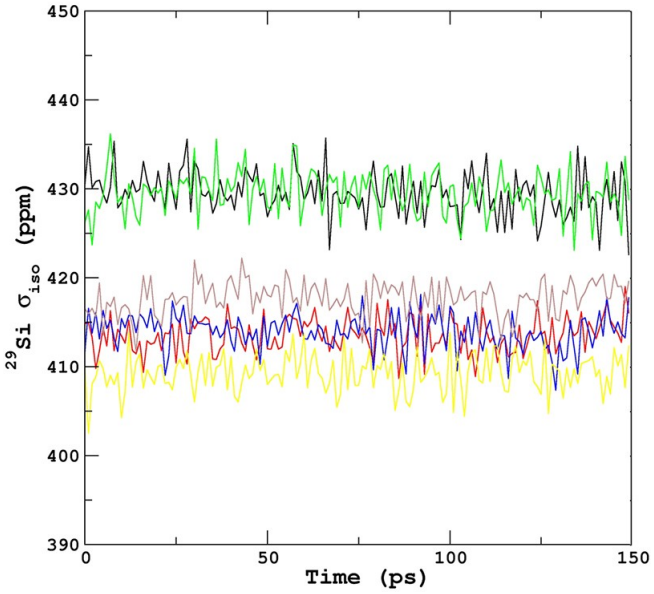



**Figure 11.** Evolution of the $\sigma_{iso}$ value of five representative silicon atoms in the 300 K trajectory of NMS1.25 glass (800 atoms).

## 3. Conclusions

A machine learning model based on Least-Square Support Vector Regression–Kernel Ridge Regression (LSSVR-KRR), was trained on NMR isotropic magnetic shielding ($^{17}$O, $^{23}$Na, $^{25}$ Mg, and $^{29}$Si) for a series of (Na, Mg)-silicate glasses simulated with Molecular Dynamics (MD) with different potentials and at various temperatures. The NMR data were calculated by applying the DFT-GIPAW approach using simulation boxes containing 400 atoms. An incomplete Cholesky decomposition (ICD) was found to perform very well and efficiently to reduce the training set to a small number of representative data (the landmark points) used for the Nyström approximation of the kernel matrix, underlying the LSSVR technique.

The Smooth Overlap of Atomic Positions (SOAP) descriptor, using Bessel spherical functions of the first kind as radial basis functions, was used. The LSSVR-KRR and SOAP parameters were systematically optimized using a 5-fold cross-validation approach.

The trained ML model shows an excellent predicting power, with a Mean Absolute Error (MAE) of 1.30. 2.4, 2.0, and 2.4 ppm for $^{29}$Si, $^{17}$O, $^{23}$Na, and $^{25}$ Mg, respectively and a Root Mean Square Error (RMSE) of 1.7, 3.1, 2.5, and 3.1 ppm, in the same order.

A detailed study of the database composition was performed, showing an overview of the relevant structural data predicted in different conditions. This was done by training LSSVR-KRR models with sub-datasets comprising data obtained at each temperature, with each potential, before or after 0 K DFT relaxation, and leaving out one composition at a time. All these partial models were then used to predict the NMR parameters on the remaining structures.


© 2024 The Authors. ChemPhysChem published by Wiley-VCH GmbH

14397641, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cphc.202300782 by University Modena, Wiley Online Library on [14/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

We showed that models trained on data from a single potential give good predictions on test sets from the same potential but transfer poorly to geometries obtained with other ones. This is due to the different structural features sampled by each potential. In fact, none of the considered potentials predict distributions of structural data (Bond angle distribution, Partial Si–O radial distribution functions, Si[$Q^n$] speciation …) that completely overlap and include the ones predicted with the others, and this results in different domains of $\sigma_{iso}$. By contrast, a model trained on data from all the potentials gives slightly lower accuracy with respect to models trained on specific data but exhibits good transferability.

We found it important to include high-temperature data in the training dataset to improve the model as the configurational space explored is expanded. Also in this case, training a model only on low-temperature data gives very good results on the same temperature predictions, but it fails at higher temperatures.

We find that leaving one glass composition out of the training model does not significantly affect the quality of the prediction on the glass itself proving the good transferability of the LSSVR-KRR model to compositions that are out (but close) of the training domain.

The model trained on the total dataset (all potentials, all temperatures) was applied to larger structures (800, 4000, and 20000 atoms) than the ones used for training (400 atoms). To validate the approach, a comparison was made between 29Si MAS NMR spectra of an 800-atom structure obtained with DFT-GIPAW calculations and predicted with LSSVR-KRR showing excellent agreement.

The application to large models with small computational costs allows the simulation of NMR spectra for several geometries from a dynamic's trajectory, averaging thermal fluctuations. The quality of the spectra improves by increasing the size of the simulation boxes, even if a convergence is observed when going to very large dimensions.

## Supporting Information Summary

The ESI contains data figures and tables not reported in the main text, the description of the force-fields used, more computational details and the description of the LSSVR-KRR algorithm used.

## Acknowledgements

This work was granted access to the HPC resources of TGCC under the allocation DARI-A0090906303 (2020) and DARI-A0110906303 (2021) attributed by GENCI (Grand Equipement National de Calcul Intensif). AP and MB acknowledge financial support from PNRR MUR project ECS_00000033_ECOSISTER. AP thanks NVIDIA for granting the project 'Neural Networks for Atomistic Simulations of Oxide Based Materials' through the donation of 1 NVIDIA A100 for PCIe within the NVIDIA Academic Hardware Grant Program. Open Access publishing facilitated by Università degli Studi di Modena e Reggio Emilia, as part of the Wiley - CRUI-CARE agreement.

## Conflict of Interests

The authors declare no conflict of interest.

## Data Availability Statement

The datasets used to train the ML model have been stored in the https://zenodo.org/repository. DOI 10.5281/zenodo.11121462

[1] N. Bisbrouck, M. Bertani, F. Angeli, T. Charpentier, D. de Ligny, J.-M. Delaye, S. Gin, M. Micoulaut, *J. Am. Ceram. Soc.* **2021**, *104*, 4518–4536.
[2] F. Angeli, T. Charpentier, D. De Ligny, C. Cailleteau, *J. Am. Ceram. Soc.* **2010**, *93*, 2693–2704.
[3] F. Angeli, T. Charpentier, M. Gaillard, P. Jollivet, *J. Non-Cryst. Solids* **2008**, *354*, 3713–3722.
[4] S. Kroeker, J. F. Stebbins, *Am. Mineral.* **2000**, *85*, 1459–1464.
[5] S. K. Lee, J. F. Stebbins, *Geochim. Cosmochim. Acta* **2009**, *73*, 1109–1119.
[6] S. K. Lee, H.-I. Kim, E. J. Kim, K. Y. Mun, S. Ryu, *J. Phys. Chem. C* **2016**, *120*, 737–749.
[7] H. Maekawa, T. Maekawa, K. Kawamura, T. Yokokawa, *J. Non-Cryst. Solids* **1991**, *127*, 53–64.
[8] S. Y. Park, S. K. Lee, *Geochim. Cosmochim. Acta* **2018**, *238*, 563–579.
[9] H. Eckert, *Int. J. Appl. Glass Sci.* **2018**, *9*, 167–187.
[10] R. Youngman, *Materials (Basel)* **2018**, *11*, 476.
[11] N. Stone-Weiss, H. Bradtmüller, M. Fortino, M. Bertani, R. E. Youngman, A. Pedone, H. Eckert, A. Goel, *J. Phys. Chem. C* **2020**, *124*, 17655–17674.
[12] T. Charpentier, M. C. Menziani, A. Pedone, *RSC Adv.* **2013**, *3*, 10550–10578.
[13] C. J. Pickard, F. Mauri, *Phys. Rev. B* **2001**, *63*, 245101.
[14] M. Profeta, F. Mauri, C. J. Pickard, *J. Am. Chem. Soc.* **2003**, *125*, 541–548.
[15] T. Charpentier, S. Ispas, M. Profeta, F. Mauri, C. J. Pickard, *J. Phys. Chem. B* **2004**, *108*, 4147–4161.
[16] J. Cuny, S. Messaoudi, V. Alonzo, E. Furet, J.-F. Halet, E. Le Fur, S. E. Ashbrook, C. J. Pickard, R. Gautier, L. Le Polles, *J. Comput. Chem.* **2008**, *29*, 2279–2287.
[17] T. Charpentier, *Solid State Nucl. Magn. Reson.* **2011**, *40*, 1–20.
[18] C. Bonhomme, C. Gervais, F. Babonneau, C. Coelho, F. Pourpoint, T. Azaïs, S. E. Ashbrook, J. M. Griffin, J. R. Yates, F. Mauri, C. J. Pickard, *Chem. Rev.* **2012**, *112*, 5733–5779.
[19] S. Sneddon, D. M. Dawson, C. J. Pickard, S. E. Ashbrook, *Phys. Chem. Chem. Phys.* **2014**, *16*, 2660–2673.
[20] A. Pedone, *Int. J. Quantum Chem.* **2016**, *116*, 1520–1531.
[21] B. Hehlen, B. Rufflé, in *Encyclopedia of Glass Science, Technology, History, and Culture* (Eds.: P. Richet, R. Conradt, A. Takada, J. Dyon), Wiley, **2021**, 287–300.
[22] G. R. Schleder, A. C. M. Padilha, C. M. Acosta, M. Costa, A. Fazzio, *J. Phys. Mater.* **2019**, *2*, 032001.
[23] E. Jonas, S. Kuhn, N. Schlörer, *Magn. Reson. Chem.* **2022**, *60*, 1021–1031,.
[24] M. Ceriotti, M. J. Willatt, G. Csányi, in *Handbook of Materials Modeling : Methods: Theory and Modeling* (Eds: W. Andreoni, S. Yip), Springer International Publishing, Cham **2018**, 1–27.
[25] F. M. Paruzzo, A. Hofstetter, F. Musil, S. De, M. Ceriotti, L. Emsley, *Nat Commun* **2018**, *9*, 4501.
[26] J. Cuny, Y. Xie, C. J. Pickard, A. A. Hassanali, *J. Chem. Theory Comput.* **2016**, *12*, 765–773.
[27] Z. Chaker, M. Salanne, J.-M. Delaye, T. Charpentier, *Phys. Chem. Chem. Phys.* **2019**, *21*, 21709–21725.
[28] R. Gaumard, D. Dragún, J. N. Pedroza-Montero, B. Alonso, H. Guesmi, I. Malkin Ondík, T. Mineva, *Computation* **2022**, *10*, 74.
[29] T. Ohkubo, A. Takei, Y. Tachi, Y. Fukatsu, K. Deguchi, S. Ohki, T. Shimizu, *J. Phys. Chem. A* **2023**, *127*, 973–986.
[30] F. Chauchard, R. Cogdill, S. Roussel, J. M. Roger, V. Bellon-Maurel, *Chemometr. Intell. Lab. Syst.* **2004**, *71*, 141–150.


© 2024 The Authors. ChemPhysChem published by Wiley-VCH GmbH

14397641, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cphc.202300782 by University Modena, Wiley Online Library on [14/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

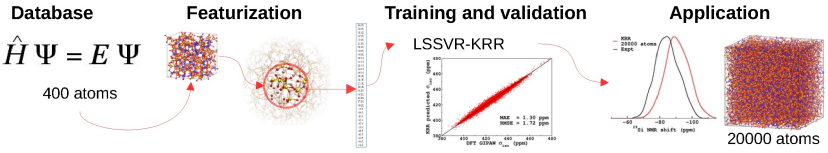
[31] U. Thissen, M. Pepers, B. Üstün, W. J. Melssen, L. M. C. Buydens, *Chemometr. Intell. Lab. Syst.* **2004**, *73*, 169–179.
[32] R. M. Balabin, E. I. Lomakina, *Phys. Chem. Chem. Phys.* **2011**, *13*, 11710.
[33] P. Drineas, M. W. Mahoney, *J. Mach. Lear. Res.* **2005**, *23*, 2153–2175.
[34] C. Williams, M. Seeger, 2001, 13, Proceedings NeurIPS.
[35] B. B. Karki, L. P. Stixrude, *Science* **2010**, *328*, 740–742.
[36] L. Stixrude, B. Karki, *Science* **2005**, *310*, 297–299.
[37] D. L. Morse, J. W. Evenson, *Int. J. Appl. Glass Sci.* **2016**, *7*, 409–412.
[38] J. M. Oliveira, R. N. Correia, M. H. Fernandes, J. Rocha, *J. Non-Cryst. Solids* **2000**, *265*, 221–229.
[39] T. Kokubo, H. Kushitani, C. Ohtsuki, S. Sakka, T. Yamamuro, *J Mater Sci: Mater Med* **1992**, *3*, 79–83.
[40] M. Diba, F. Tapia, A. R. Boccaccini, L. A. Strobel, *Int. J. Appl. Glass Sci.* **2012**, *3*, 221–253.
[41] D. Bellucci, E. Veronesi, M. Dominici, V. Cannillo, *Mater. Sci. Eng., C* **2020**, *110*, 110699.
[42] J. C. Mauro, A. Tandia, K. D. Vargheese, Y. Z. Mauro, M. M. Smedskjaer, *Chem. Mater.* **2016**, *28*, 4267–4277.
[43] T. K. Bechgaard, G. Scannell, L. Huang, R. E. Youngman, J. C. Mauro, M. M. Smedskjaer, *J. Non-Cryst. Solids* **2017**, *470*, 145–151.
[44] H. Bradtmüller, T. Uesbeck, H. Eckert, T. Murata, S. Nakane, H. Yamazaki, *J. Phys. Chem. C* **2019**, *123*, 14941–14954.
[45] M. Wang, M. M. Smedskjaer, J. C. Mauro, G. Sant, M. Bauchy, *Phys. Rev. Applied* **2017**, *8*, 054040.
[46] M. Lograd, H. Eckert, T. Murata, S. Nakane, H. Yamazaki, *J. Am. Ceram. Soc.* **2021**, *104*, 2250–2267.
[47] R. Guo, C. T. Brigden, S. Gin, S. W. Swanton, I. Farnan, *J. Non-Cryst. Solids* **2018**, *497*, 82–92.
[48] M. Bertani, N. Bisbrouck, J.-M. Delaye, F. Angeli, A. Pedone, T. Charpentier, *J. Am. Ceram. Soc.* n.d., n/a, DOI: 10.1111/jace.19157.
[49] N. Bisbrouck, M. Micoulaut, J.-M. Delaye, M. Bertani, T. Charpentier, S. Gin, F. Angeli, *J. Phys. Chem. B* **2021**, *125*, 11761–11776.
[50] R. L. McGreevy, *J. Phys.: Condens. Matter* **2001**, *13*, R877–R913.
[51] T. Charpentier, E. Chesneau, L. Cormier, G. Tricot, in *14th International Conference on the Structure of Non-Crystalline Materials, Kobe, Japan* **2019**.
[52] A. M. B. Silva, C. M. Queiroz, S. Agathopoulos, R. N. Correia, M. H. V. Fernandes, J. M. Oliveira, *J. Mol. Struct.* **2011**, *986*, 16–21.
[53] A. Pedone, G. Malavasi, A. N. Cormack, U. Segre, M. C. Menziani, *Chem. Mater.* **2007**, *19*, 3144–3154.
[54] M. F. Guest, A. M. Elena, A. B. G. Chalk, *Mol. Simul.* **2021**, *47*, 194–227.
[55] A. Pedone, G. Malavasi, M. C. Menziani, A. N. Cormack, U. Segre, *J. Phys. Chem. B* **2006**, *110*, 11780–11795.
[56] M. Bertani, M. C. Menziani, A. Pedone, *Phys. Rev. Mater.* **2021**, *5*, 045602.
[57] M. Bertani, A. Pallini, M. Cocchi, M. C. Menziani, A. Pedone, *J. Am. Ceram. Soc.* **n/a**, *n.d.*, DOI 10.1111/jace.18681.
[58] A. Tilocca, N. H. de Leeuw, A. N. Cormack, *Phys. Rev. B* **2006**, *73*, 104209.
[59] A. Pedone, G. Malavasi, M. C. Menziani, *J. Phys. Chem. C* **2009**, *113*, 15723–15730.
[60] B. G. Dick, A. W. Overhauser, *Phys. Rev.* **1958**, *112*, 90–103.
[61] A. Pedone, M. Bertani, L. Brugnoli, A. Pallini, *J. Non-Cryst. Solids: X* **2022**, *15*, 100115.
[62] G. Kresse, J. Furthmüller, *Phys. Rev. B* **1996**, *54*, 11169–11186.
[63] J. P. Perdew, K. Burke, M. Ernzerhof, *Phys. Rev. Lett.* **1996**, *77*, 3865–3868.
[64] J. R. Yates, C. J. Pickard, F. Mauri, *Phys. Rev. B* **2007**, *76*, 024401.
[65] T. Charpentier, P. Kroll, F. Mauri, *J. Phys. Chem. C* **2009**, *113*, 7917–7929.
[66] E. Gambuzzi, A. Pedone, M. C. Menziani, F. Angeli, D. Caurant, T. Charpentier, *Geochim. Cosmochim. Acta* **2014**, *125*, 170–185.
[67] A. P. Bartók, R. Kondor, G. Csányi, *Phys. Rev. B* **2013**, *87*, 184115.
[68] J. Behler, M. Parrinello, *Phys. Rev. Lett.* **2007**, *98*, 146401.
[69] G. Montavon, K. Hansen, S. Fazli, M. Rupp, F. Biegler, A. Ziehe, A. Tkatchenko, **2012**, *25*, Proceeding NeurIPS .
[70] A. Seko, A. Togo, I. Tanaka, *in* Nanoinformatics (Ed: I. Tanaka), Springer, Singapore **2018**, 3–23.
[71] W. Li, Y. Ando, *Phys. Chem. Chem. Phys.* **2018**, *20*, 30006–30020.
[72] B. Onat, C. Ortner, J. R. Kermode, *J. Chem. Phys.* **2020**, *153*, 144106.
[73] M. F. Langer, A. Goeßmann, M. Rupp, *npj Comput. Mater.* **2022**, *8*, 1–14.
[74] V. L. Deringer, A. P. Bartók, N. Bernstein, D. M. Wilkins, M. Ceriotti, G. Csányi, *Chem. Rev.* **2021**, *121*, 10073–10141.
[75] F. Musil, A. Grisafi, A. P. Bartók, C. Ortner, G. Csányi, M. Ceriotti, *Chem. Rev.* **2021**, *121*, 9759–9815.
[76] M. A. Caro, *Phys. Rev. B* **2019**, *100*, 024112.
[77] R. Jinnouchi, F. Karsai, G. Kresse, *Phys. Rev. B* **2019**, *100*, 014105.
[78] E. Kocer, J. K. Mason, H. Erturk, *AIP Adv.* **2020**, *10*, 015021.
[79] C. M. Bishop, *Pattern Recognition and Machine Learning*, Springer, New York **2006**.
[80] V. Vapnik, *The Nature of Statistical Learning Theory*, Springer Science & Business Media, New York, NY, USA **1999**.
[81] C. E. Rasmussen, C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, Cambridge, MA, USA **2005**.
[82] W. Pronobis, K.-R. Müller, *in Machine Learning Meets Quantum Physics. Lecture Notes in Physics* (Eds: K. T. Schütt, S. Chmiela, O. A. von Lilienfeld, A. Tkatchenko, K. Tsuda, K.-R. Müller), Springer International Publishing, Cham **2020**, Vol. 968.
[83] C. J. C. Burges, B. Schölkopf, A. J. Smola, *Advances in Kernel Methods: Support Vector Learning*, MIT Press, Cambridge, MA, USA **1998**.
[84] K.-R. Muller, S. Mika, G. Ratsch, K. Tsuda, B. Scholkopf, *IEEE Trans. Neural Netw.* **2001**, *12*, 181–201.
[85] C.-J. Lin, J. J. Moré, *SIAM J. Sci. Comput.* **1999**, *21*, 24–45.
[86] P. M. Williams, *Neural Comput.* **1996**, *8*, 843–854.
[87] E. Anderson, Z. Bai, C. Bischof, S. Blackford, J. Demmel, J. Dongarra, J. Du Croz, A. Greenbaum, S. Hammarling, A. McKenney, D. Sorensen, *LAPACK Users' Guide–Third Edition*, Society For Industrial And Applied Mathematics, Philadelphia, PA **1999**.
[88] M. Ahmed, R. Seraj, S. M. S. Islam, *Electronics* **2020**, *9*, 1295.
[89] J. Jiang, C. Song, H. Zhao, C. Wu, Y. Liang, in *2008 IEEE International Conference on Granular Computing*, IEEE, Hangzhou, **2008**, 340–345.
[90] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, B. Kozinsky, *Nat Commun* **2022**, *13*, 2453.

Manuscript received: October 20, 2023
Revised manuscript received: May 6, 2024
Accepted manuscript online: July 25, 2024
Version of record online: ■■, ■■


© 2024 The Authors. ChemPhysChem published by Wiley-VCH GmbH

14397641, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cphc.202300782 by University Modena, Wiley Online Library on [14/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License

# RESEARCH ARTICLE




A Kernel Ridge Regression (KRR) model using a Least Square Support Vector Regression (LSSVR) approach is used for the accurate prediction of NMR isotropic magnetic shielding ($\sigma_{iso}$) of active nuclei ($^{17}O$, $^{23}Na$, $^{25}Mg$, and $^{29}Si$) in a series of (Mg, Na)–silicate glasses. The trained model is able of fast prediction for structures containing up to 20000 atoms.

*M. Bertani, A. Pedone, F. Faglioni, T. Charpentier**

1 – 14

**Accelerating NMR Shielding Calculations Through Machine Learning Methods: Application to Magnesium Sodium Silicate Glasses**

14397641, 0, Downloaded from https://chemistry-europe.onlinelibrary.wiley.com/doi/10.1002/cphc.202300782 by University Modena, Wiley Online Library on [14/10/2024]. See the Terms and Conditions (https://onlinelibrary.wiley.com/terms-and-conditions) on Wiley Online Library for rules of use; OA articles are governed by the applicable Creative Commons License