computation

MDPI

Article

# Regression Machine Learning Models Used to Predict DFT-Computed NMR Parameters of Zeolites

**Robin Gaumard [1], Dominik Dragún [2], Jesús N. Pedroza-Montero [1], Bruno Alonso [1], Hazar Guesmi [1], Irina Malkin Ondík [2,3] and Tzonka Mineva [1,*]**

[1] ICGM, CNRS, ENSCM, Universite de Montpellier, 34296 Montpellier, France; robin.gaumard@enscm.fr (R.G.); jesus-nain.pedroza-montero@umontpellier.fr (J.N.P.-M.); bruno.alonso@enscm.fr (B.A.); hazar.guesmi@enscm.fr (H.G.)
[2] FIIT STU in Bratislava, Ilkovičova 2, 84216 Bratislava, Slovakia; domco.dragun@gmail.com (D.D.); malkin.ondik@gmail.com (I.M.O.)
[3] MicroStep-MIS, Spol. S.R.O., Čavojského 1, 84104 Bratislava, Slovakia
* Correspondence: tzonka.mineva@enscm.fr

**Abstract:** Machine learning approaches can drastically decrease the computational time for the predictions of spectroscopic properties in materials, while preserving the quality of the computational approaches. We studied the performance of kernel-ridge regression (KRR) and gradient boosting regressor (GBR) models trained on the isotropic shielding values, computed with density-functional theory (DFT), in a series of different known zeolites containing out-of-frame metal cations or fluorine anion and organic structure-directing cations. The smooth overlap of atomic position descriptors were computed from the DFT-optimised Cartesian coordinates of each atoms in the zeolite crystal cells. The use of these descriptors as inputs in both machine learning regression methods led to the prediction of the DFT isotropic shielding values with mean errors within 0.6 ppm. The results showed that the GBR model scales better than the KRR model.

**Keywords:** NMR; machine learning; zeolites

check for updates

**Citation:** Gaumard, R.; Dragún, D.; Pedroza-Montero, J.N.; Alonso, B.; Guesmi, H.; Malkin Ondík, I.; Mineva, T. Regression Machine Learning Models Used to Predict DFT-Computed NMR Parameters of Zeolites. *Computation* **2022**, *10*, 74. https://doi.org/10.3390/computation10050074

Academic Editor: Henry Chermette

Received: 30 March 2022
Accepted: 10 May 2022
Published: 13 May 2022

**Publisher's Note:** MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

## 1. Introduction

Machine learning (ML) coupled with density functional theory (DFT) calculations has been rapidly emerging for predictions of nuclear magnetic resonance (NMR) isotropic shielding values [1–9]. The role of the experimental NMR investigations to recognise the local atomic environment in chemical and biological systems has been established for decades. Theoretical DFT calculations, using either the gauge-invariant atomic orbital (GIAO) or gauge invariant-projector augmented wave (GIPAW), have been widely employed to improve the NMR signal assignments and/or identify the local structural environment and molecular interactions of the targeted nucleus [10,11]. The interest in the last few years in developing and applying ML models for the prediction of NMR parameters thus originates in the importance of the rapid achievement of accurate theoretical NMR parameters.

Hitherto, several ML models [12] have been built and applied for predicting NMR isotropic shielding ($\sigma_{iso}$) or, respectively, the chemical shift ($\delta = \sigma_{ref} - \sigma_{iso}$) of $^{1}$H, $^{13}$C, $^{13}$O, and $^{13}$N nuclei in small organic, aromatic molecules or molecular crystals [2,6,13–20]. These ML models comprise deep neural networks (DNNs) [15], convolutional neural networks (CNNs) [16], the IMPRESSION model based on kernel-ridge regression (KRR) [6,19,20], linear-ridge regression [2], gradient boosting regression (GBR) [21,22], graph neural networks (GNNs) [23,24], and the Δ-ML method [7]. Chemical shifts of proteins have been predicted using random forest regression (RFR) [13,14,17,18]. Despite the strong decrease of the computational time to train the model and predict the NMR parameters, in comparison to the GIAO and GIPAW calculations, most of the ML models yielded somewhat less accurate results in comparison to the experimental data than the DFT $\sigma_{iso}$ with PBE exchange–correlation functionals [7]. Significantly lesser is the amount of works devoted to NMR property calculations in silicates [1–3]. The ML precision in predicting $^{29}$Si and $^{17}$O chemical shifts in these amorphous solids is found more accurate than in the organic compounds. For example, the ML-predicted deviation from DFT-GIPAW calculations is obtained to be only 0.7 ppm for $^{29}$Si and 1.5 ppm for $^{17}$O in $SiO_2$ glasses [2]. The supervised feed-forward neural network representation yielded mean absolute errors (MAEs) of $\delta_{iso} < 1$ ppm for $^{29}$Si in ZSM-11 and a-cristobalite [1]. The same NN model also performed very well for the $^{17}$O quadrupolar coupling constant predictions, giving MAEs ($C_q(^{17}O)$) of 0.07 MHz in cristobalite and 0.06 MHz in ZSM-11 zeolite.

One of the most significant tasks to take into account in the ML applications is the choice of the descriptors, representing the local chemical environment of each atom in the system. This choice is not trivial because it greatly depends on the shape of the molecular system (simple organic molecules or crystalline materials) and on the considered data set [17]. The most widely used descriptor for predicting the NMR properties in organic molecules and materials is the smooth overlap of the atomic positions (SOAP) descriptor. This descriptor can also be used as a kernel when it is coupled with the kernel-ridge regression methods. Indeed the SOAP descriptor has been already found very efficient to describe the local chemical environment of a large range of chemical compounds, and in particular, it allows obtaining the accurate prediction of NMR properties [1,2]. Furthermore, the symmetry functions are widely used for describing the chemical environment in the neural network representation [1]. Molecular descriptors and fragment descriptors [25] led to predicting with a great accuracy the J-coupling constants in small organic molecules. The ML combination with DFT is therefore a promising tool, and further validations are of high interest.

In this work, we apply two simple state-of-the-art regression ML methods, namely KRR and GBR, to predict $\sigma_{iso}$ in a set of crystalline zeolite structures, selected from the International Zeolite Association's (IZA) structure database [26]. The zeolites are the crystalline alumino-silicate porous materials with waste industrial applications as catalysts or molecular sieves. The three-dimensional zeolite structure is composed by tetrahedron units with Si atoms in the centre and four oxygen atoms at the vertices, which can organise in a variety of porous frameworks, with pores of sizes varying between 2 and 10 nm [27,28]. ML methods coupled to DFT computations have already emerged for predicting mechanical properties [29], nitrogen adsorption [30], molar volumes, and cohesive energies [31] in zeolites. The success rate of these ML applications to zeolites vary according to the predicted properties and the proposed ML approach [32]. Among the spectroscopy techniques, used to study zeolite structures and chemical compositions, most of the NMR techniques can today be routinely applied to the as-synthesised zeolitic materials. We therefore found it of interest to examine and report in this work the performance of simple ML methods trained on the computed DFT $\sigma_{iso}$ values in a series of known zeolite structures.

## 2. Methods and Computational Details

### *2.1. Kernel-Ridge Regression*

The first ML approach used by us is KRR [6], which consists of a combination of the ridge regression and the kernel method. The KRR model is suitable for complex continuous data, which cannot be described by a linear regression. Unlike the linear regression, the kernel-ridge regression method offers larger flexibility by transforming the input with a regression function.

Below, we briefly illustrate the KRR scheme. In the case of the ML linear regression algorithm, the goal is to minimise a function Ω called the quadratic cost [33], which is defined as

$$\Omega(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\left(Y_i - \mathbf{w}^T\mathbf{X}_i\right)^2, \tag{1}$$

where $\mathbf{X}_i$ represents the vector of the input data, $Y_i$ are the scalar output data, $N$ corresponds to the dimension of the input data, and the vector $\mathbf{w}$ is the vector of weights that will be optimised during the training process. In the case of the ridge regression algorithm, an additional term is implemented to the previous quadratic cost in order to prevent overfitting problems during the training stage by regularising its value. Hence, the form of the quadratic cost becomes

$$\Omega(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N}\left(Y_i - \mathbf{w}^T\mathbf{X}_i\right)^2 + \frac{1}{2}\lambda||\mathbf{w}||_2^2, \tag{2}$$

where $\lambda$ is a positive parameter that controls the value of the vector norm $\mathbf{w}$. This step is called L2-regularisation because of the use of the L2-norm of the vector $\mathbf{w}$. In order to determine the parameter $\lambda$, a cross-validation algorithm is widely used [34]. Thus, by minimising the function $\Omega(\mathbf{w})$, it leads to a simple linear problem to be solved for the set of weights as follows:

$$\sum_{i=1}^{N}\left(Y_i - \mathbf{w}^T\mathbf{X}_i\right)\mathbf{X}_i = \lambda\mathbf{X}_i. \tag{3}$$

These optimised weights are thus obtained as

$$\mathbf{w} = \left(\lambda\mathbf{I} + \sum_{i=1}^{N}\mathbf{X}_i\mathbf{X}_i^T\right)^{-1}\left(\sum_{j=1}^{N}Y_j\mathbf{X}_j\right), \tag{4}$$

where $\mathbf{I}$ is the identity matrix.

This linear regression method is limited to problems that can be described as a linear function; thus, to overcome this limit, a non-linear kernel function is introduced in order to measure the similarity between two samples of a high-dimensional space. The most widely used kernel function is the Gaussian kernel function. In the KRR method, the vector of the input data, $\mathbf{X}_i$, is substituted by the non-linear kernel function $\varphi(\mathbf{X}_i)$. Therefore, we can rewrite the expression of the optimised weight parameters as a function of $\varphi(\mathbf{X}_i)$:

$$\begin{aligned}\mathbf{w} &= \left(\lambda\mathbf{I} + \sum_{i=1}^{N}\varphi(\mathbf{X}_i)\varphi(\mathbf{X}_i^T)\right)^{-1}\left(\sum_{j=1}^{N}Y_j\varphi(\mathbf{X}_j)\right)\\ &= \left(\lambda\mathbf{I} + \varphi(\mathbf{X}_i)\varphi(\mathbf{X}_i^T)\right)^{-1}\varphi(\mathbf{X}_i)Y_i\\ &= \varphi(\mathbf{X}_i)\left(\varphi(\mathbf{X}_i^T)\varphi(\mathbf{X}_i) + \lambda\mathbf{I}\right)^{-1}Y_i.\end{aligned} \tag{5}$$

By defining the coefficient $\alpha_i = \left(\varphi(\mathbf{X}_i^T)\varphi(\mathbf{X}_i) + \lambda\mathbf{I}\right)^{-1}Y_i$, the optimised weights are simply expressed as

$$\mathbf{w} = \sum_{i=1}^{N}\alpha_i\varphi(\mathbf{X}_i). \tag{6}$$

Therefore, during the training phase of the kernel-ridge regression, the aim is to calculate $\alpha_i$, which are subsequently used to predict the output values. For the KRR model, we used the code from the open repository [35]. The similarity of the input vectors is determined based on the user-defined similarity function, e.g., kernel, in our case, the difference between the SOAP vectors.

### 2.2. Gradient Boosting Regression

The gradient boosting regression is a powerful regression firstly introduced by Freund and Schapire [36,37] through an adaptive boosting algorithm. At the beginning, this regression method was used for classification problems [38] and later on adapted for regression problems [21,22]. The aim of the gradient boosting regression method is to find a function $f^*$ that minimises the loss function $\Theta$ [39] defined as

$$f^*(\mathbf{X}, Y) = \arg\min E_{\mathbf{X},Y}[\Theta(Y, f(\mathbf{X}))], \tag{7}$$

where $\mathbf{X}$ is the vector of input variables, $Y$ is the output variable, and $E_{\mathbf{X},Y}$ represents the floor function applied for the vector $\mathbf{X}$ and the variable $Y$. In the boosted model, the function $f(\mathbf{X})$ is defined as a weighted linear combination of base learners by the following formula:

$$f(\mathbf{X}) = \sum_{i=1}^{N} \alpha_i h_i(\mathbf{X}, \beta_i), \tag{8}$$

where $\alpha_i$ are the real coefficients of the linear combination and $\beta_i$ are the parameters of the base learners $h_i(\mathbf{X}, \beta_i)$. The minimisation of the loss function $\Theta$ is carried out via an optimisation of the function $f$ using the recursive relation

$$f_{m+1}(\mathbf{X}) = f_m(\mathbf{X}) + \arg\min \sum_{i=1}^{N} \Theta(Y_i, f_m(\mathbf{X}_i + h_{m+1}(\mathbf{X}_i))). \tag{9}$$

The here-used gradient boosting regression method enables us to create strongly learning trees from poorly learning trees [40]. This approach utilises boosting so that the trees are created sequentially, as opposed to random forests, where the trees are generated in parallel. Each new tree is created with an effort to reduce the prediction error learning from the errors of the previous tree. The goal is to achieve the lowest possible error while keeping the predicted values as accurate as possible. We used the Anaconda distribution for Python 3.8.5, utilising the scikit-learn program package [40,41] with the GBR model, where the *random_state* hyperparameter was set to 0 and the rest of the hyperparameters were set to the default values.

### *2.3. SOAP Descriptors*

Two data sets in comma-separated values (CSV) format were prepared using the DFT-optimised Cartesian coordinates of the zeolites and the isotropic shielding value in ppm for each atom in the zeolites. The first data set contains the Cartesian coordinates ($x$, $y$, and $z$) of each zeolite, the calculated $\sigma_{iso}$, the name of the chemical element, and the name of the zeolite (taken from the IZA). The second CSV file contains 3 × 3 tensors and the name of the corresponding zeolite. We used the DScribe package [42] to convert our data to smooth overlap of atomic positions (SOAP) descriptor vectors. Individual structures were represented as Atoms class objects from the ASE package [43,44] with the use of 3 × 3 tensors. We began by creating a DScribe.SOAP object, for which the parameters such as the number of basis functions, range, level l, and a list of all elements in our data were set. Subsequently, the DScribe.SOAP.create function was used to create a SOAP vector for each atom. The complete data set was split into a training and test set in a ratio of 8:2.

### *2.4. DFT Computational Details*

A periodic DFT-based approach was used to carry out a full geometrical optimisation (atomic positions and unit-cell parameters) of all the structures in the data set. The geometrical optimisations were carried out with the Crystal17 program, based on atom-centred Gaussian orbitals [45]. All-electron basis functions of double-$\zeta$ quality were used as follows: 6-31d1 for O, N, C, and H [46]; 85-11G* for Al [47] and Pople's basis set (6–21G) with polarisation for Si. The generalised gradient-corrected PBE approximation was used as the exchange correlation (XC) functional, augmented by the empirical London dispersion (D3) term with the Becke–Johnson damping function [48]. The optimised structural parameters of zeolites with OSDA, obtained with the Crystal code and all-electron databases, were found by us to agree well with the experimental bond distances and bond angles [49–51].

For this reason, we applied the same computational protocol for the optimisation of the zeolite structures in this work.

Single-point energy calculations were carried out for these optimised geometries in order to compute the isotropic shielding values of all the atoms in the zeolites. For this, we used the open-source code QUANTUM ESPRESSO [52–54] with the GIPAW method in combination with the ultrasoft pseudopotentials with GIPAW reconstruction [55,56], from the USSP pseudopotential database [57]. The wave-function and charge density energy cut-offs were set to 60 Ry and 720 Ry, respectively. A Monkhorst–Pack grid of k-points [58] corresponding to a maximum spacing of 0.06 $\text{Å}^{-1}$ in the reciprocal space was used. The self-consistent field (SCF) energy convergence tolerance was set to $10^{-10}$.

## 3. Results and Discussion

### *3.1. Data Set and DFT Isotropic Shielding*

To ensure a heterogeneity of the atomic environments giving rise to different DFT $\sigma_{iso}$ values, we considered zeolites containing Al, Na, and Li cations, as well as MFI-type zeolites, containing the organic structure-directing agents (OSDAs), which are the tetrapropylammonium (TPA) and tripropylethylammonium (TPEA) cations. Among the MFI-OSDA types of structures, we considered five pure silica structures (silicalite-1), labelled as MFI-TPA and MFI-TPEA in Table S1 in the Supporting Information (SI) Section. The four MFI-ETPA structures present the location of the TPEA ethyl chain either in the direct or zig-zag channels. In silicalite-1 zeolites, the fluorine anion is the charge-compensating ion. The remaining MFI-OSDA zeolites are those with the TPA cation and one $Al^{3+}$, which substitutes at each of the 24 non-equivalent Si-sites of the asymmetric unit. The initial structures of the pure inorganic zeolites were the crystallographic information files (CIFs) that were collected from the IZA database. We built the MFI-OSDA structures from the available crystallographic data for TPA (ETPA) [59] and ZSM-5-TPA [60] zeolites. These structures were optimised in our previous studies [61,62] using the same level of DFT theory. We thus constructed a more heterogeneous data set that contains Si, Al, N, C, H, Li, and F atomic environments. The geometries and DFT $\sigma_{iso}$ values of all atoms were used in the ML training and prediction calculations.

To access roughly the quality of the DFT $\sigma_{iso}$ results, we correlated them with the experimental chemical shifts of $^{29}$Si, which are available in the IZA database. The zeolites for which $\delta^{29}Si$ were collected are labelled by an asterisk in Table S1.

The linear fitting between the DFT and experimental data, illustrated in Figure 1, demonstrates that the PBE-D3 results followed reasonably well the overall experimental trend for the selected zeolites. It is worth noting that the experimental NMR data were recorded under different experimental conditions [26] and often for non-ideal zeolite structures that might contain defects, such as silanols, water, hydroxides, and in- or out-framework cations. Taking into account these factors and the linear fitting $R^2$ coefficient of 0.9995, as well as the root-mean-squared error (RMSE) of 2.44 ppm of the DFT values with respect to the fitted values against the experimental isotropic shieldings, we concluded a rather good correlation between the computed and experimental results.

The distributions of the calculated DFT isotropic shieldings of $^{29}$Si, $^{17}$O, $^{27}$Al, $^{13}$C, and $^{1}$H are reported in Figure 2. The majority of the Si atoms have DFT $\sigma_{iso}$($^{29}$Si) in the range 422–426 ppm, as follows from the maximum number of the chemical environments in this interval. Nevertheless, the predominant number of $\sigma_{iso}$($^{29}$Si) is obtained in a 400–440 ppm interval, and there are few Si-sites, for which $\sigma_{iso}$($^{29}$Si) < 350 ppm. The other nuclei, largely presented in the zeolites, are $^{17}$O and $^{1}$H. The peakin the oxygen atoms' distribution indicates that the largest number of oxygen sites has $\sigma_{iso}$($^{17}$O) values at around 196 ppm. The $\sigma_{iso}$($^{17}$O) values span a large interval between 150 and 250 ppm with several outliers outside this region. The isotropic shieldings of hydrogen sites are between 22 and 29 ppm, and the distribution of $^{13}$C is characterised by two distinguished peaksat around 100 and 150 ppm. The hydrogen and carbon sites belong to OSDAs in MFI and ZSM-5 zeolite types.

In the studied structures, the number of $Al^{3+}$ cations is significantly smaller, and their $\sigma_{iso}(^{27}Al)$ are spread in the 450–550 ppm interval.

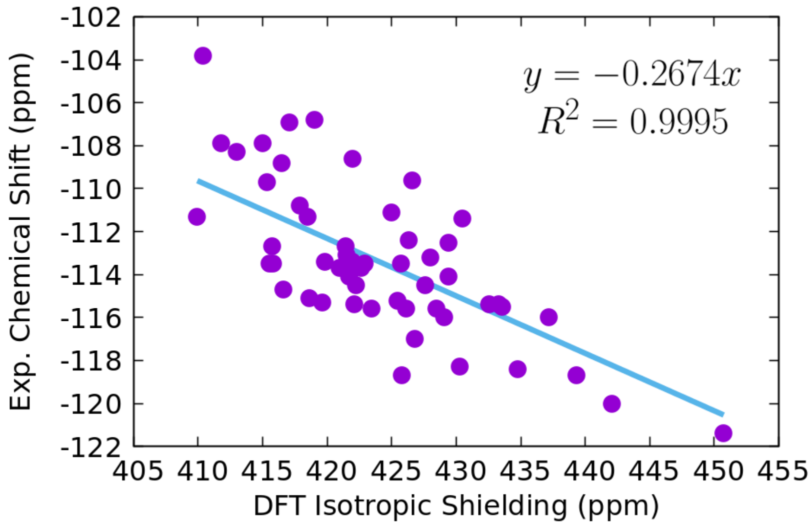

**Figure 1.** Comparison between DFT isotropic shielding and experimental chemical shift (both in ppm) alongside the linear fitting of the data (blue line).

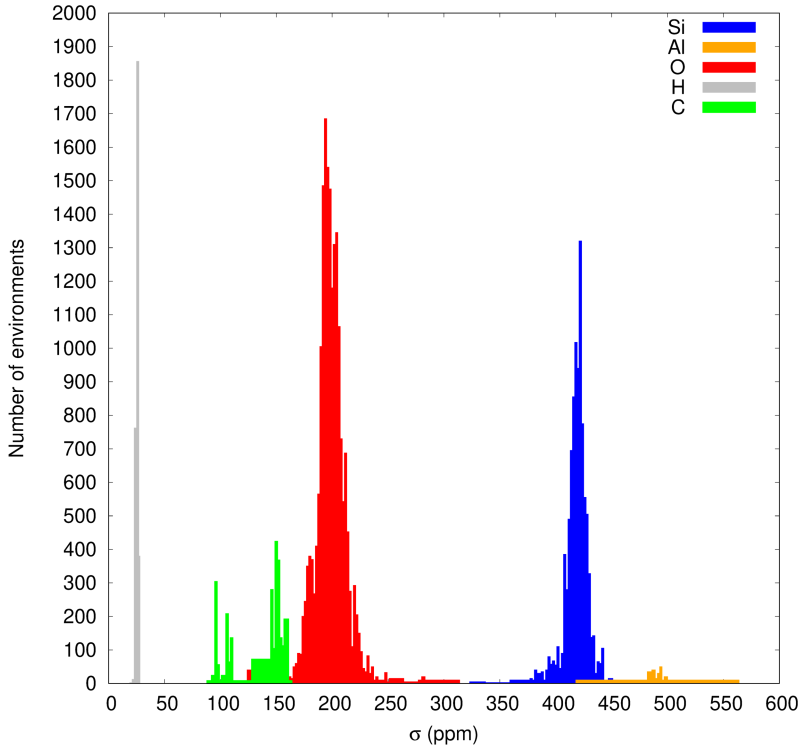

**Figure 2.** Distribution of the number of oxygen (O), silicon (Si), aluminium (Al), carbon (C), and hydrogen (H) atomic environments in zeolites, according to their isotropic shielding. The histograms are obtained with an interval of 2.0 ppm for C, O, Al, Si, and 1.0 ppm for H in the count of the number of atomic environment.

### *3.2. KRR and GBR Models to Predict NMR Isotropic Shielding*

In this section, we discuss the performance of the KRR and GBR models. The zeolite data set, discussed above, was split into training (first 80%) and validating sets (last 20%) of zeolites. As discussed in the Methods section, we used the training set to build the SOAP descriptors. First, we considered all atoms in the zeolite structures that represent a total of 14,513 atomic environments in the KRR and GBR models. In the second part, only the silicon atoms, with their Cartesian coordinates and $\sigma_{iso}(^{29}\text{Si})$ values, were collected in a smaller data set. The choice of Si atoms is because $\sigma_{iso}(^{29}\text{Si})$ experimental data are most often considered as fingerprints of the local structure around Si-sites and can account for the presence of silanols, oxygen, or silicon vacancies or other defect types. The number of Si-atomic environments in this smaller data set was reduced to 3756, and among them, 3004 were used as training data and 752 as validation data.

The distribution of the differences between $\sigma_{iso}$, computed with DFT and those predicted from the KRR and GBR models, is presented in Figures 3 and 4, respectively, whereas the correlations between the predicted vs. the DFT values are reported in Figures S1 and S2. In the KRR model, the regularisation hyperparameter $\alpha$ was set to 0.1. Here, only one outlier value is identified in the results from the KRR application. The predicted outlier $\sigma_{iso}$ = 487.7 ppm is down-shifted by about 26 ppm with respect to the "true" DFT $\sigma_{iso}$ = 513.82 ppm. This outlier is in the silimanite structure with the Cartesian coordinates equal to 2.67, 1.46, and 1.10 Å. The application of both the KRR and GBR models on the smaller set containing only the Si atomic environments and their $\sigma_{iso}(^{29}\text{Si})$ values in the interval 380–450 ppm yielded again an excellent correlation between the predicted vs. DFT computed data, as follows from the plot in Figures 5 and S3 (KRR) and Figures 6 and S4 (GBR). We obtained only one remarkable outlier $\sigma_{iso}(^{29}\text{Si})$ value when using the KRR model. This outlier is now in the ITW zeolite. Its predicted $\sigma_{iso}(^{29}\text{Si})$ value of 431.51 ppm is up-shifted with respect to the computed with DFT $\sigma_{iso}(^{29}\text{Si})$ = 419.99 ppm. The coordinates of the outlier Si atom are: x = −1.31, −1.24, −2.72 Å. No outliers were identified when applying the GBR model.

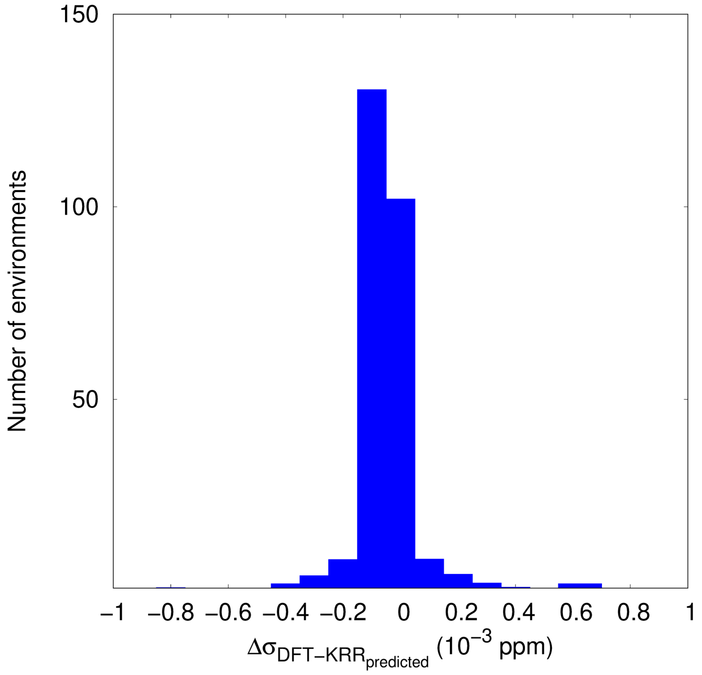

**Figure 3.** Distribution of the differences between the isotropic shielding values computed with DFT and those predicted with the ML-KRR method. All atomic environments are considered. The histograms are obtained with an interval of 0.1 ppm in the count of the number of atomic environments.

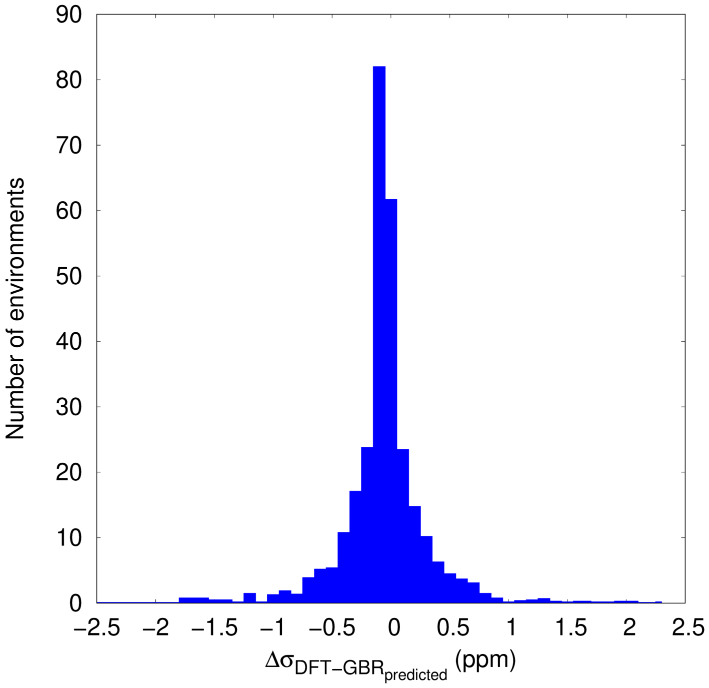

**Figure 4.** Distribution of the differences between the isotropic shielding values computed with DFT and those predicted with the ML-GBR method. All atomic environments are considered. The histograms are obtained with an interval of 0.1 ppm in the count of the number of atomic environments.

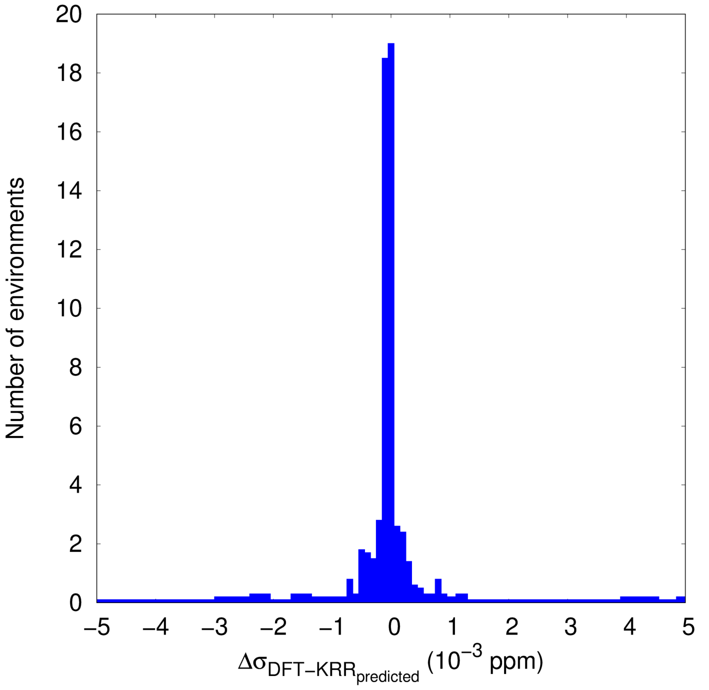

**Figure 5.** Distribution of the differences between the isotropic shielding values computed with DFT and those predicted with the ML-KRR method. Only silicon atomic environments are considered. The histograms are obtained with an interval of 0.1 ppm in the count of the number of atomic environments.

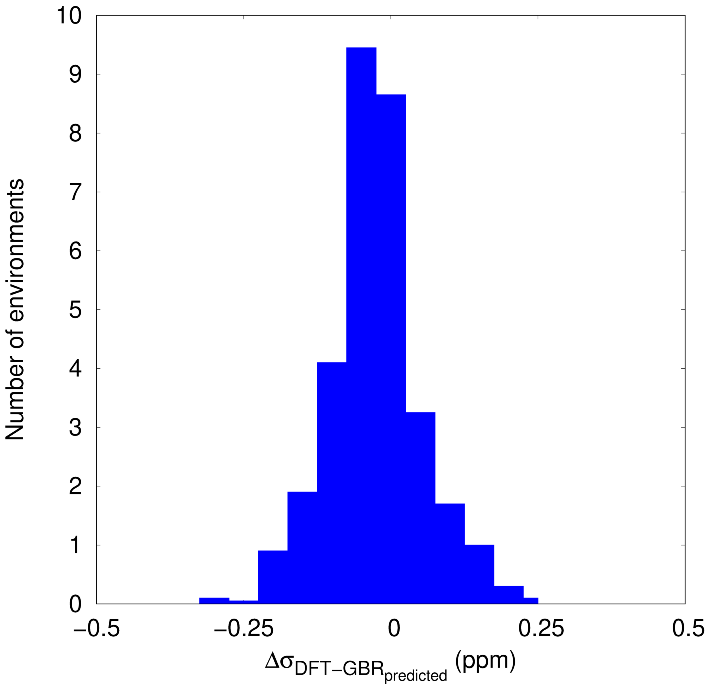

**Figure 6.** Distribution of the differences between the isotropic shielding values computed with DFT and those predicted with the ML-GBR method. Only silicon atomic environments are considered. The histograms are obtained with an interval of 0.05 ppm in the count of the number of atomic environments.

The mean-squared error (MSE), root-mean-squared error (RMSE), standard deviation error (STD), and the mean absolute error (MAE), as well as the $R^2$ coefficients are compared in Table 1. These results confirm the very good performance of both ML models leading to the $R^2$ coefficients of 0.997 (KRR, all atoms in the data set) and 0.999 for the other three sets of predictions. The MAE, RMSE, and MSE results are <0.6 ppm. The most notable differences between the performance of KRR and GBR models are the training and prediction time, also reported in Table 1. The GBR model appears to be faster by two orders of magnitude than the KRR model. Therefore, we concluded that the GBR model scales better than the KRR model.

**Table 1.** Training and prediction time, mean absolute error (MAE), root-mean-squared error (RMSE), mean-squared error (MSE), absolute and square standard deviation errors (STD AE and STD SE), and the $R^2$ coefficient of the KRR and GBR model predictions together with the average of both predicted values (AVG). The data shown are only for the Si atoms (Si) and all atoms (All) in the zeolite.

| | **Machine Learning Models** | | | | | |
|---|---|---|---|---|---|---|
| **Parameters** | **KRR (All)** | **GBR (All)** | **AVG (All)** | **KRR (Si)** | **GBR (Si)** | **AVG (Si)** |
| Training time (s) | 3796.4 | 49.0 | - | 136.7 | 12.2 | - |
| Prediction time (s) | 1900.8 | 0.6 | - | 74.2 | 0.02 | - |
| MAE (ppm) | 0.023 | 0.226 | 0.116 | 0.037 | 0.057 | 0.046 |
| STD AE (ppm) | 0.524 | 0.538 | 0.236 | 0.490 | 0.054 | 0.246 |
| MSE (ppm) | 0.275 | 0.341 | 0.069 | 0.241 | 0.006 | 0.062 |
| STD SE (ppm) | 12.669 | 8.158 | 1.285 | 5.304 | 0.011 | 1.341 |
| RMSE (ppm) | 0.524 | 0.584 | 0.262 | 0.491 | 0.008 | 0.250 |
| $R^2$ | 0.999 | 0.999 | - | 0.999 | 0.997 | - |

A combination of the KRR and GBR models might remove outliers and reduce the errors. A simple estimation of the combination between both regression approaches was carried out by assuming equal weight coefficients (0.5), that is taking the mean of the predicted isotropic shielding values by the KRR and GBR ML models. The resulting distributions of the differences between the mean of the predicted and DFT values, as well as the correlation plots between the predicted vs. DFT isotropic shielding data are plotted in Figures S5–S8. The average of both ML models approached the quality of the GBR predictions. The outlier, identified in the reduced set of silicon atoms, equals 425.78 ppm; thus, it is predicted to be closer to DFT $\sigma_{iso}(^{29}\text{Si} = 419.99$ ppm). The combination of both regression methods led to a significant decrease of the STD errors, MSE, and RMSE (Table 1) in comparison to the respective errors found from the application of each ML model. It therefore follows that the combination of regression methods might be a useful approach toward the removal of errors of a single regression model.

Discussing the quality of the predicted $\sigma_{iso}$ results with respect to those computed with DFT is not trivial in the case of the zeolite structures. As noted above, the rigorous comparison of the computed $\sigma_{iso}(^{29}\text{Si})$ and the experimental chemical shift data, collected from the IZA database (see Figure 1), is not straightforward. Despite this fact, considering that the RMS error of the linear fit of DFT $\sigma_{iso}(^{29}\text{Si})$ vs. the experimental $\delta_{iso}(^{29}\text{Si})$ results (Figure 1) amounts to 2.44 ppm, we concluded that the predicted values with RMSE in the range 0.008–0.5 ppm do not worsen the quality of the DFT method used by us. This suggests a very promising application of both the KRR and GBR models, not only to predict the $\sigma_{iso}$ of $^{29}$Si, but also for the other nuclei in the the zeolite data set, because outliers were not identified among these nuclei. However, we note the limited number or heterogeneity of C, H, F, and Li atomic environments. It is therefore not surprising that outliers were not established among those atoms. On the other hand, the number of oxygen environments is four-times the number of Si environments in the all-atom data set. The excellent correlation between the predicted vs. DFT-computed values can be therefore also concluded for $\sigma_{iso}(^{17}\text{O})$. The combination of SOAP descriptors with simple ML regression models appears to lead to a promising predictive capability of NMR isotropic shielding of $^{29}$Si and $^{17}$O in the zeolites, which is in line with previous work using SOAP descriptors and regression methods for predictions of NMR parameters in the organic solids [4] and silicates [1,2].

## 4. Conclusions

In this paper, we studied the capability of two simple machine learning regression models, KRR and GBR, to predict the $\sigma_{iso}$ values in a series of known zeolites. The DFT calculations with periodic boundary conditions were carried out to fully optimise the crystallographic zeolite structures, collected from the IZA database and the MFI-OSDA types of zeolites, and to compute the $\sigma_{iso}$ values for each atom in the data set. In addition to the inorganic zeolite framework, composed by Si, O, and Al atoms, the data set contains various out-frame cations, such as $\text{Li}^+$, $\text{F}^-$, and TPA and TPEA molecular cations.

The quality of the DFT $\sigma_{iso}(^{29}\text{Si})$ was found to be reasonably good compared to the available experimental $\delta_{iso}(^{29}\text{Si})$ in the IZA database. The SOAP descriptors, obtained from the optimised Cartesian coordinates of each atom in the DFT-based data set, were used as inputs in both machine learning regression models. Both the KRR and GBR approaches predicted isotropic shieldings with mean errors smaller than 1 ppm. The comparison between the training and predictions time gave a preference to the GBR, found to scale better than the KRR model. These results are promising for more extensive ML applications based on simple regression in combination with DFT calculations in order to accelerate the calculations of NMR parameters in various zeolitic materials.

**Supplementary Materials:** The following Supporting Information can be downloaded at: https://www.mdpi.com/article/10.3390/computation10050074/s1, Table S1, Figures S1–S8.

**Author Contributions:** Conceptualisation: T.M. and I.M.O.; construction zeolite database: R.G., H.G., B.A. and T.M.; ML software: D.D. and I.M.O.; data curation: J.N.P.-M., T.M. and I.M.O.; writing—original draft preparation: R.G., J.N.P.-M. and T.M.; supervision: T.M. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The optimised atomic Cartesian coordinates and isotropic shielding values are available at: https://mycore.core-cloud.net/index.php/s/FH6EGSBVBwwZZ8R (accessed on 29 March 2022).

**Acknowledgments:** The authors acknowledge the access to the HPC resources of CCRT/CINES/IDRIS, which was granted under the allocation A0090807369 and A0110807369 by GENCI.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Cuny, J.; Xie, Y.; Pickard, C.J.; Hassanali, A.A. Ab Initio Quality NMR Parameters in Solid-State Materials Using a High-Dimensional Neural-Network Representation. *J. Chem. Theory Comput.* **2016**, *12*, 765–773. [CrossRef]
2. Chaker, Z.; Salanne, M.; Delaye, J.M.; Charpentier, T. NMR shifts in aluminosilicate glasses via machine learning. *Phys. Chem. Chem. Phys.* **2019**, *21*, 21709–21725. [CrossRef] [PubMed]
3. Liu, S.; Li, J.; Bennett, K.C.; Ganoe, B.; Stauch, T.; Head-Gordon, M.; Hexemer, A.; Ushizima, D.; Head-Gordon, T. Multiresolution 3D-DenseNet for Chemical Shift Prediction in NMR Crystallography. *J. Phys. Chem. Lett.* **2019**, *10*, 4558–4565. [CrossRef] [PubMed]
4. Paruzzo, F.M.; Hofstetter, A.; Musil, F.; De, S.; Ceriotti, M.; Emsley, L. Chemical shifts in molecular solids by machine learning. *Nat. Commun.* **2018**, *9*, 4501. [CrossRef] [PubMed]
5. Rupp, M.; Ramakrishnan, R.; von Lilienfeld, O.A. Machine Learning for Quantum Mechanical Properties of Atoms in Molecules. *J. Phys. Chem. Lett.* **2015**, *6*, 3309–3313. [CrossRef]
6. Gerrard, W.; Bratholm, L.A.; Packer, M.J.; Mulholland, A.J.; Glowacki, D.R.; Butts, C.P. IMPRESSION—Prediction of NMR parameters for 3-dimensional chemical structures using machine learning with near quantum chemical accuracy. *Chem. Sci.* **2020**, *11*, 508–515. [CrossRef]
7. Unzueta, P.A.; Greenwell, C.S.; Beran, G.J.O. Predicting Density Functional Theory-Quality Nuclear Magnetic Resonance Chemical Shifts via Δ-Machine Learning. *J. Chem. Theory Comput.* **2021**, *17*, 826–840. [CrossRef]
8. Cordova, M.; Balodis, M.; Hofstetter, A.; Paruzzo, F.; Nilsson Lill, S.O.; Eriksson, E.S.E.; Berruyer, P.; Simões de Almeida, B.; Quayle, M.J.; Norberg, S.T.; et al. Structure determination of an amorphous drug through large-scale NMR predictions. *Nat. Commun.* **2021**, *12*, 2964. [CrossRef]
9. Aguilera-Segura, S.M.; Dragún, D.; Gaumard, R.; Di Renzo, F.; Malkin Ondík, I.; Mineva, T. Thermal fluctuations and conformational effects on NMR parameters in $\beta$-O-4 lignin dimer from QM/MM and machine learning approaches. *Phys. Chem. Chem. Phys.* **2022**, *24*, 8820–8831. [CrossRef]
10. 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*, 1–20. [CrossRef]
11. Dib, E.; Mineva, T.; Alonso, B. Chapter Three—Recent Advances in $^{14}$N Solid-State NMR *Annu. Rep. NMR Spectrosc.* **2016**, *87*, 175–235. [CrossRef]
12. Jonas, E.; Kuhn, S.; Schlörer, N. Prediction of chemical shift in NMR: A review. *Magn. Reson. Chem.* 2021, *in press*. [CrossRef] [PubMed]
13. Arun, K.; Langmead, C.J. Structure based chemical shift prediction usgin random forests non-linear regression. In Proceedings of the 4th Asia-Pacific Bioinformatics Conference, Taipei, Taiwan, 13–16 February 2006; Series on Advances in Bioinformatics and Computational Biology; World Scientific Publishing Co.: Singapore, 2005; Volume 3, pp. 317–326. [CrossRef]
14. Han, B.; Liu, Y.; Ginzinger, S.W.; Wishart, D.S. SHIFTX2: Significantly improved protein chemical shift prediction. *J. Biomol. NMR* **2011**, *50*, 43. [CrossRef] [PubMed]
15. Gao, P.; Zhang, J.; Peng, Q.; Zhang, J.; Glezakou, V.A. General Protocol for the Accurate Prediction of Molecular $^{13}$C/$^{1}$H NMR Chemical Shifts via Machine Learning Augmented DFT. *J. Chem. Inf. Model.* **2020**, *60*, 3746–3754. [CrossRef]
16. Gao, P.; Zhang, J.; Sun, Y.; Yu, J. Toward Accurate Predictions of Atomic Properties via Quantum Mechanics Descriptors Augmented Graph Convolutional Neural Network: Application of This Novel Approach in NMR Chemical Shifts Predictions. *J. Phys. Chem. Lett.* **2020**, *11*, 9812–9818. [CrossRef]
17. Haghighatlari, M.; Li, J.; Heidar-Zadeh, F.; Liu, Y.; Guan, X.; Head-Gordon, T. Learning to Make Chemical Predictions: The Interplay of Feature Representation, Data, and Machine Learning Methods. *Chem* **2020**, *6*, 1527–1542. [CrossRef]
18. Li, J.; Bennett, K.C.; Liu, Y.; Martin, M.V.; Head-Gordon, T. Accurate prediction of chemical shifts for aqueous protein structure on "Real World" data. *Chem. Sci.* **2020**, *11*, 3180–3191. [CrossRef]
19. Gupta, A.; Chakraborty, S.; Ramakrishnan, R. Revving up $^{13}$C NMR shielding predictions across chemical space: Benchmarks for atoms-in-molecules kernel machine learning with new data for 134 kilo molecules. *Mach. Learn. Sci. Technol.* **2021**, *2*, 035010. [CrossRef]
20. Gerrard, W.; Yiu, C.; Butts, C.P. Prediction of $^{15}$N chemical shifts by machine learning. *Magn. Reson. Chem.* 2021, *in press*. [CrossRef]

21. Beygelzimer, A.; Hazan, E.; Kale, S.; Luo, H. Online gradient boosting. In Proceedings of the Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015, Montreal, QC, Canada, 7–12 December 2015; pp. 2458–2466.
22. Biau, G.; Cadre, B.; Rouvière, L. Accelerated gradient boosting. *Mach. Learn.* **2019**, *108*, 971–992. [CrossRef]
23. Guan, Y.; Shree Sowndarya, S.V.; Gallegos, L.C.; John, P.C.S.; Paton, R.S. Real-time prediction of $^{1}$H and $^{13}$C chemical shifts with DFT accuracy using a 3D graph neural network. *Chem. Sci.* **2021**, *12*, 12012–12026. [CrossRef] [PubMed]
24. Jonas, E.; Kuhn, S. Rapid prediction of NMR spectral properties with quantified uncertainty. *J. Cheminform.* **2019**, *11*, 50. [CrossRef] [PubMed]
25. Ito, K.; Xu, X.; Kikuchi, J. Improved Prediction of Carbonless NMR Spectra by the Machine Learning of Theoretical and Fragment Descriptors for Environmental Mixture Analysis. *Anal. Chem.* **2021**, *93*, 6901–6906. [CrossRef] [PubMed]
26. Baerlocher, C.; McCusker, L. Database of Zeolite Structures. Available online: http://www.iza-structure.org/databases/ (accessed on 28 March 2022).
27. Hölderich, W.; Hesse, M.; Näumann, F. Zeolites: Catalysts for Organic Syntheses. *Angew. Chem. Int. Ed. Engl.* **1988**, *27*, 226–246. [CrossRef]
28. Cejka, J.; van Bekkum, H.; Corma, A.; Schueth, F. Introduction to Zeolite Molecular Sieves. In *Introduction to Zeolite Molecular Sieves*, 3rd ed.; John Wiley and Sons: Hoboken, NJ, USA, 2007; p. 1094.
29. Evans, J.D.; Coudert, F.X. Predicting the Mechanical Properties of Zeolite Frameworks by Machine Learning. *Chem. Mater.* **2017**, *29*, 7833–7839. [CrossRef]
30. Gu, Y.; Liu, Z.; Yu, C.; Gu, X.; Xu, L.; Gao, Y.; Ma, J. Zeolite Adsorption Isotherms Predicted by Pore Channel and Local Environmental Descriptors: Feature Learning on DFT Binding Strength. *J. Phys. Chem. C* **2020**, *124*, 9314–9328. [CrossRef]
31. Helfrecht, B.A.; Semino, R.; Pireddu, G.; Auerbach, S.M.; Ceriotti, M. A new kind of atlas of zeolite building blocks. *J. Chem. Phys.* **2019**, *151*, 154112. [CrossRef]
32. Kwak, S.J.; Kim, H.S.; Park, N.; Park, M.J.; Lee, W.B. Recent progress on Al distribution over zeolite frameworks: Linking theories and experiments. *Korean J. Chem. Eng.* **2021**, *38*, 1117–1128. [CrossRef]
33. Welling, M. Kernel ridge Regression. Available online: https://web2.qatar.cmu.edu/~gdicaro/10315-Fall19/additional/welling-notes-on-kernel-ridge.pdf (accessed on 28 March 2022).
34. An, S.; Liu, W.; Venkatesh, S. Fast cross-validation algorithms for least squares support vector machine and kernel ridge regression. *Pattern Recognit.* **2007**, *40*, 2154–2162. [CrossRef]
35. ML-CSC-tutorial. Available online: https://github.com/fullmetalfelix/ML-CSC-tutorial (accessed on 24 March 2022).
36. Freund, Y.; Schapire, R.E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *J. Comput. Syst. Sci.* **1997**, *55*, 119–139. [CrossRef]
37. Freund, Y.; Schapire, R.E. A Short Introduction to Boosting. In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI '99), Stockholm, Sweden, 31 July–6 August 1999; Morgan Kaufmann Publishers Inc.: San Francisco, CA, USA, 1999; pp. 1401–1406.
38. Natekin, A.; Knoll, A. Gradient boosting machines, a tutorial. *Front. Neurorobot.* **2013**, *7*. [CrossRef] [PubMed]
39. He, Z.; Lin, D.; Lau, T.; Wu, M. Gradient Boosting Machine: A Survey. *arXiv* **2019**, arXiv:1908.06951.
40. Hastie, T.; Tibshirani, R.; Friedman, J. Boosting and Additive Trees. In *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009; pp. 337–387. [CrossRef]
41. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
42. Himanen, L.; Jäger, M.O.; Morooka, E.V.; Federici Canova, F.; Ranawat, Y.S.; Gao, D.Z.; Rinke, P.; Foster, A.S. DScribe: Library of descriptors for machine learning in materials science. *Comput. Phys. Commun.* **2020**, *247*, 106949. [CrossRef]
43. Larsen, A.H.; Mortensen, J.J.; Blomqvist, J.; Castelli, I.E.; Christensen, R.; Dułak, M.; Friis, J.; Groves, M.N.; Hammer, B.; Hargus, C.; et al. The atomic simulation environment—A Python library for working with atoms. *J. Phys. Condens. Matter* **2017**, *29*, 273002. [CrossRef]
44. Bahn, S.R.; Jacobsen, K.W. An object-oriented scripting interface to a legacy electronic structure code. *Comput. Sci. Eng.* **2002**, *4*, 56–66. [CrossRef]
45. Dovesi, R.; Erba, A.; Orlando, R.; Zicovich-Wilson, C.M.; Civalleri, B.; Maschio, L.; Rérat, M.; Casassa, S.; Baima, J.; Salustro, S.; et al. Quantum-mechanical condensed matter simulations with CRYSTAL. *WIRES Comput. Mol. Sci.* **2018**, *8*, e1360. [CrossRef]
46. Gatti, C.; Saunders, V.R.; Roetti, C. Crystal field effects on thetopological properties of the electron density in molecular crystals: The case of urea. *J. Chem. Phys.* **1994**, *101*, 10686–10696. [CrossRef]
47. Catti, M.; Valerio, G.; Dovesi, R.; Causà, M. Quantum-mechanical calculation of the solid-state equilibrium MgO+$\alpha$-$Al_2O_3 \rightleftarrows MgAl_2O_4$ (spinel) versus pressure. *Phys. Rev. B* **1994**, *49*, 14179–14187. [CrossRef]
48. Grimme, S.; Antony, J.; Ehrlich, S.; Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H-Pu. *J. Chem. Phys.* **2010**, *132*, 154104. [CrossRef]
49. Dib, E.; Mineva, T.; Gaveau, P.; Véron, E.; Sarou-Kanian, V.; Fayon, F.; Alonso, B. Probing Disorder in Al-ZSM-5 Zeolites by 14N NMR Spectroscopy. *J. Phys. Chem. C* **2017**, *121*, 15831–15841. [CrossRef]

50. Mineva, T.; Dib, E.; Gaje, A.; Petitjean, H.; Bantignies, J.L.; Alonso, B. Zeolite Structure Direction: Identification, Strength and Involvement of Weak CHO Hydrogen Bonds. *Chem. Phys. Chem.* **2020**, *21*, 149–153. [CrossRef] [PubMed]
51. Al-Nahari, S.; Ata, K.; Mineva, T.; Alonso, B. Ubiquitous Presence of Intermolecular CHO Hydrogen Bonds in As-synthesized Host-Guest Zeolite Materials. *ChemistrySelect* **2021**, *6*, 9728–9734. [CrossRef]
52. Giannozzi, P.; Baroni, S.; Bonini, N.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Chiarotti, G.L.; Cococcioni, M.; Dabo, I.; et al. QUANTUM ESPRESSO: A modular and open-source software project for quantum simulations of materials. *J. Phys. Condens. Matter* **2009**, *21*, 395502. [CrossRef]
53. Giannozzi, P.; Andreussi, O.; Brumme, T.; Bunau, O.; Nardelli, M.B.; Calandra, M.; Car, R.; Cavazzoni, C.; Ceresoli, D.; Cococcioni, M.; et al. Advanced capabilities for materials modelling with Quantum ESPRESSO. *J. Phys. Condens. Matter* **2017**, *29*, 465901. [CrossRef]
54. Quantum Espresso. Available online: https://www.quantum-espresso.org (accessed on 28 March 2022).
55. Pickard, C.J.; Mauri, F. All-electron magnetic response with pseudopotentials: NMR chemical shifts. *Phys. Rev. B* **2001**, *63*, 245101. [CrossRef]
56. Yates, J.R.; Pickard, C.J.; Mauri, F. Calculation of NMR chemical shifts for extended systems using ultrasoft pseudopotentials. *Phys. Rev. B* **2007**, *76*, 024401. [CrossRef]
57. Quantum Espresso Pseudopotentials. Available online: http://www.quantum-espresso.org/pseudopotentials (accessed on 28 March 2022).
58. Monkhorst, H.J.; Pack, J.D. Special points for Brillouin-zone integrations. *Phys. Rev. B* **1976**, *13*, 5188–5192. [CrossRef]
59. Chao, K.J.; Lin, J.C.; Wang, Y.; Lee, G. Single crystal structure refinement of TPA ZSM-5 zeolite. *Zeolites* **1986**, *6*, 35–38. [CrossRef]
60. Yokomori, Y.; Idaka, S. The structure of TPA-ZSM-5 with Si/Al = 23. *Microporous Mesoporous Mater.* **1999**, *28*, 405–413. [CrossRef]
61. Dib, E.; Mineva, T.; Veron, E.; Sarou-Kanian, V.; Fayon, F.; Alonso, B. ZSM-5 Zeolite: Complete Al Bond Connectivity and Implications on Structure Formation from Solid-State NMR and Quantum Chemistry Calculations. *J. Phys. Chem. Lett.* **2018**, *9*, 19–24. [CrossRef] [PubMed]
62. Fabbiani, M.; Al-Nahari, S.; Piveteau, L.; Dib, E.; Veremeienko, V.; Gaje, A.; Dumitrescu, D.G.; Gaveau, P.; Mineva, T.; Massiot, D.; et al. Host–Guest Silicalite-1 Zeolites: Correlated Disorder and Phase Transition Inhibition by a Small Guest Modification. *Chem. Mater.* **2022**, *34*, 366–387. [CrossRef]