Catalysis Science & Technology

ROYAL SOCIETY OF CHEMISTRY

PAPER

View Article Online
View Journal | View Issue

Cite this: *Catal. Sci. Technol.*, 2025, 15, 878

The speciation of phosphates adsorbed on γ-alumina revealed by ^{31}P NMR, AIMD and machine learning†

Adrian Hühn,[a] Tao Jiang,[a] Manuel Corral Valero, [b] Mickaël Rivallan, [b] Anne Lesage, [c] Carine Michel *[a] and Pascal Raybaud *[ab]

Received 26th September 2024,
Accepted 19th December 2024

DOI: 10.1039/d4cy01152j

rsc.li/catalysis

The chemical nature of adsorbed inorganic additives such as phophates used in the preparation of heterogeneous catalysts is suspected to impact their resulting activity. Predominant phosphate species located on the surfaces of the γ-alumina catalytic support are identified by using one-dimensional ^{31}P NMR spectra as the only experimental input. The detailed insight is made possible by combining machine learning (ML) ^{31}P chemical shift prediction and *ab initio* molecular dynamics (AIMD) to sample conformers of 10 representative possible structures and generate theoretical spectra, which were then used to decompose mathematically the broad experimental peak. At low P concentration, several types of monomeric species are found to co-exist on the γ-alumina (110) facets. Increasing the P concentration yields a marked increase in one monomeric species and one dimeric species both located on the (110) facets, whereas phosphates are mainly absent from the (100) facet. The NMR spectra broadening is interpreted by two levels of structural disorders: the various types of P species and the conformational distribution of each species. We finally propose some implications for the catalytic properties.

Introduction

Investigating phosphate speciation present on γ-alumina (γ-Al_2O_3) powders impregnated with H_3PO_4 is widely relevant for application fields as diverse as heterogeneous catalysis,[1,2] environmental science,[3] biology[4] and pharmaceutical formulations.[5] Considering more particularly γ-Al_2O_3 supported heterogenous catalysts, several studies have reported the impact of phosphates additives on the performances of MoS_2 based catalysts for hydrotreatment.[6,7] The use of phosphorus inorganic compounds during preparation was also shown to improve the thermal stability, of the metallic active phase such as cobalt in Fischer–Tropsch synthesis[8] or palladium in CO oxidation[9] and of the γ-Al_2O_3 support in biomass conversion processes.[10] In these cases, the phosphorus species are strongly suspected to act as a chemical binder between the alumina surface and the metallic active phases at the various stages of the catalyst life cycle: preparation, activation, reaction, regeneration or recycling. For this reason, unravelling the atomic scale nature and location of phosphorus species on the alumina surface is crucial for catalysis.

Solid-state nuclear magnetic resonance (NMR) spectroscopy is a unique analytical technique to unravel the atomic-scale structure of solid substrates, ranging from polymers to glasses, heterogenous catalysts, pharmaceutical drugs and biological assemblies.[5,11,12] The interpretation of NMR spectra is increasingly supported by chemical shift (CS) prediction using density functional theory (DFT) models.[13] However, the computational burden associated with those predictions often limits the exploration of the chemical space. Machine learning (ML) approaches have been recently proposed to accelerate CS prediction, both on molecular solids and inorganic materials.[14–16] This approach proves to be especially powerful when applied to forecast the NMR chemical shifts of amorphous systems.[17] However, in cases where extensive experimental databases are lacking, and the spectra are dominated by significant line broadening, the structural characterization of disordered systems using ML techniques remains highly challenging.

The direct interpretation of one-dimensional (1D) ^{31}P solid-state NMR spectra is limited regarding the identification and quantification of the various phosphate species adsorbed on oxide surfaces.[18–23] Indeed, the NMR spectra typically display a single broad line spanning 20–30 ppm in width, which shifts towards more negative values

[a] *CNRS, Laboratoire de Chimie UMR 5182, ENS de Lyon, 46 allée d'Italie, Lyon F-69342, France. E-mail: carine.michel@ens-lyon.fr*
[b] *IFP Énergies nouvelles, Rond-point de l'échangeur de Solaize, BP 3, 69360 Solaize, France. E-mail: pascal.raybaud@ifpen.fr*
[c] *Université de Lyon, CNRS, ENS Lyon, Université Lyon 1, Centre de RMN à Hauts Champs de Lyon, UMR 5082, 5 rue de la Doua, 69100 Villeurbanne, France*

† Electronic supplementary information (ESI) available. See DOI: https://doi.org/10.1039/d4cy01152j

when increasing the P concentration.[12,18–23] This signal results from the contributions of several non-equivalent phosphate species (monomer, dimer, oligomer) with different local environments, adsorbed on various surface sites. By combining static DFT calculations with two-dimensional (2D) Dynamic Nuclear Polarization Surface Enhanced NMR Spectroscopy (DNP SENS) ^{31}P–^{27}Al correlation experiments,[12] we previously identified 10 compatible structures of phosphates species adsorbed on γ-alumina sites. However, quantifying their respective contribution was not possible. Even with DFT calculated CS at hand,[24] the NMR spectra could not be decomposed due to the significant broadening of the NMR lines and the numerous possible species, with only subtle differences in their environment (varying H-bonds, P–O bonds and O–P–O angles).

To overcome the hurdle of structural disorder effects, *ab initio* molecular dynamics (AIMD) can be used nowadays to sample conformers at a given temperature. Since calculating the ^{31}P CS of each sampled structure is too costly, an alternative is to establish generalized structure-chemical shift relationships. For instance, in the case of crystalline materials (such as calcium phosphates[25] and aluminium phosphates[26]) multivariate regressions correlating simple structural parameters (P–O bonds and Al–O–P angles) with calculated ^{31}P chemical shift have been found applicable to a wide range of AlPO materials, where Al sites exhibits a close and well-defined local environment. However, they cannot be easily extrapolated to phosphate species found in a disordered environment, such as those adsorbed on oxide surfaces.

An other possibility is to build a ML model combining structural descriptors[27] with regression models,[28] trained on an extensive set of DFT chemical shifts. Some examples of DFT-ML-predicted CS of various nuclei have been reported so far with reasonable accuracy.[14,15,29] Recently, Cuny *et al.* used artificial neural networks trained on DFT calculations combined with AIMD to sample structures and predict ^{29}Si and ^{17}O NMR spectra of various materials (including silica glasses).[14]

In the present work, our aim is to reach a quantitative speciation analysis of the P species adsorbed on γ-Al_2O_3 based on the deconvolution of 1D ^{31}P NMR spectra measured at various P concentrations (0.4 to 4.1 P nm^{-2}, see Fig. SI.2†) for samples prepared as described in.[12] For that, we will use DFT-ML predicted ^{31}P CS combined with AIMD simulations to determine ^{31}P CS histograms at ambient temperature and to account for the broadening and the chemical shift evolution of the NMR signal as a function of P concentrations.

Results and discussion

Structural DFT database of adsorbed phosphates

We previously identified 10 possible structures reported in Fig. 1b). They are the most stable out of a large dataset of 1322 structures of adsorbed phosphate species on γ-Al_2O_3 identified by DFT calculations in our previous work.[12] This database is as diverse as possible, including three γ-Al_2O_3 surfaces with varying hydroxylation states ((100), n(110), R(110)), four adsorption modes of phosphate monomers: physisorbed (v_0), monodentate (v_1), bidentate (v_2) or

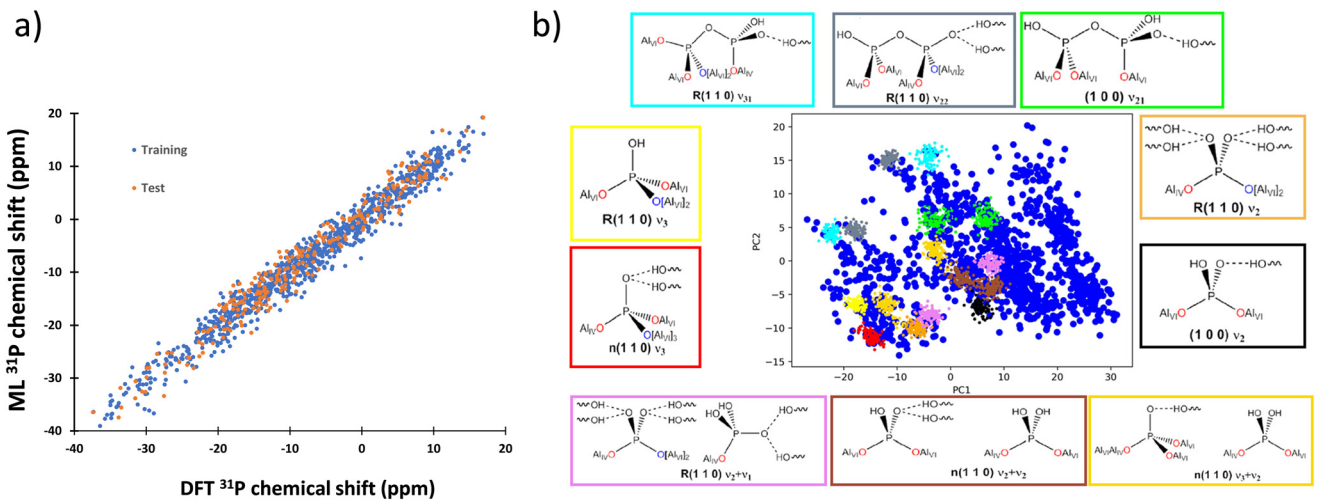

Fig. 1 a) Parity plots obtained after the initial regression model (with 3100 descriptors) for ^{31}P chemical shift: training set (blue dots), test set (orange dots). b) Score plot of the principal component analysis of the LMBTR descriptors. Blue dots: Dataset of 1322 structures optimized at the PBE + D3 level (used to compute the PCA vectors and to train and validate the ML model). Colored dots: Projections on the PCA space of structures from the AIMD trajectories of the 10 relevant P species adsorbed on γ-Al_2O_3, using one color per structure. The structures are schematically represented. Al_{IV} corresponds to 4-coordinated aluminum atoms, Al_{VI} corresponds to 6-coordinated aluminum atoms. OH groups with wiggled bonds are hydroxyl groups found on the γ-Al_2O_3 surface. Three hydrated γ-Al_2O_3 surfaces were considered: (100), non-reconstructed and reconstructed (110),[30,31] named n(110) and R(110) respectively. v_x indicates the number of bonds between the phosphate oxygens and the alumina. v_x + v_y is used for co-adsorption of two phosphate monomers while v_{xy} is used for dimeric phosphate. More details can be found in ref. 12.

tridentate (ν_3), co-adsorbed monomers ($\nu_x + \nu_y$) and phosphate dimers (ν_{xy}), binding to four fold or six fold coordinated Al sites. However, the predicted chemical shift of those 10 species is not enough to deconvolute unambiguously the 1D CP MAS NMR spectra due to signal broadening. To limit the number of parameters to be adjusted during the deconvolution, but still embrace the possible chemical diversity, we first need to predict the ^{31}P CS histograms at 300 K for the 10 adsorbed P species that were previously identified as relevant.[12]

Machine learning model for ^{31}P chemical shift prediction

We first trained a ML model to identify the correlation between the structure optimized at the PBE-D3 level[32,33] and the ^{31}P CS obtained using GIPAW[24] calculations, exploiting our dataset of 1322 structures.[12] The chemical environment of P was described using local many-body tensor representation (LMBTR) descriptors with a reduced size to avoid overfitting issues (Table SI.1†).[34] The correlation between the structure and the ^{31}P CS was obtained using a ridge regularized least-square regression algorithm with regularization parameter set to 0.02 (see methods). The root mean square error (RMSE) of our ML model (after removal of the lowest rank coefficients) is 2.2 ppm with respect to DFT calculated ^{31}P CS as illustrated in Fig. 1a), which is very close to the reported error for the DFT CS values (2 ppm; see methods and Table SI.2†). According to this comparison, the values obtained from both approaches in our relevant structures are fairly close, and close to the average error represented by RMSE. Thus, we can reasonably argue that the expected overall accuracy in our approach corresponds to the highest statistical error from both ML and DFT computations, that is, 2.2 ppm. This accuracy is reasonably good for such disordered surfaces and independent of the random selection during training. Moreover, since the differences in the CS among the theoretical systems used for spectra deconvolution is usually higher that this error estimate, we do not expect this source of error to have an impact on our conclusions. Then, we sampled the structural distribution of each 10 adsorbed P species[12] using AIMD trajectories of 35 ps at 300 K after 11 ps of thermalization (see Methods). A Principal Component Analysis (PCA) of the LMBTR descriptors shows that the structures sampled using AIMD overlaps with the database structures generated at T = 0 K, ensuring a good transferability from the database to the target configuration space (Fig. 1b).

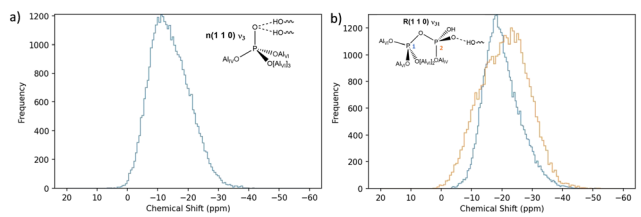

Fig. 2 Sample histograms of ^{31}P CS calculated using our ML model on structures sampled from AIMD for a) monomer n(110) ν_3, b) dimer R(110) ν_{31}.

Histogram of ^{31}P chemical shifts at ambient temperature

Applying this ML model on the structures sampled by AIMD, we obtained the ^{31}P CS histogram at 300 K associated with each configuration of phosphates adsorbed on alumina. Fig. 2 shows the ^{31}P CS histograms obtained for two relevant species: the phosphate monomer n(110) ν_3 and the dimer R(110) ν_{31} (for all the other P species see SI.3†). Those histograms are broadened and exhibit two peaks when two phosphate moieties are present (dimers and co-adsorption of two monomers). Besides, they are slightly skewed to the negative chemical shifts as also observed for the experimental spectra. This slight asymmetry can only be captured by predicting the variations of the chemical shift over a large set of structures sampled by AIMD.

The mean CS obtained at 300 K is close to that computed at 0 K with the same ML model, although it has a general tendency to be shifted to more negative values by less than 3 ppm (with the exception of some species). The associated standard deviation, quantifying the theoretical spectra broadening during AIMD is comprised between 5.2 and 8.1 ppm depending on the P species (Table SI.2†). The origin of this broadening is rather complex and can be induced by several parameters: one of them is the structural disorder[35] which may itself recover several effects. On the one hand, it contains the contribution of various individual species: here, the various phosphate species adsorbed on different alumina sites such as R(110), n(110) or (100) and modes ν_x, ν_{xy}. On the other hand, it contains the contribution of several conformational structures (due to the distortion of angles, bonds…) of a given species which co-exist on the alumina surface. In the latter case, the distribution of these conformers closely depends on the temperature which helps to overcome weak energy barriers between each conformer. The AIMD simulation at the given temperature allows a rather large sampling of these conformers. In particular, we observe that the CS distribution is broader when the phosphate moiety is bonded to alumina through only one P–O–Al bond (ν_1) as in (100) ν_{21}, R(110) ν_{31} and R(110) $\nu_2 + \nu_1$. This greater flexibility of the monodentate can be tracked back in the AIMD trajectories monitoring the root mean square fluctuation of the PO_4 entities (0.14 for n(110) ν_3 and the ν_3 of the dimer R(110)-ν_{31}, 0.26 for the ν_1 of the dimer R(110) ν_{31}). An earlier attempt to quantify the effect of structural deformation was proposed by DFT computation of CS along several points of some arbitrary chosen low vibrational modes in a molecular crystal.[36] The present AIMD-ML approach goes beyond by considering all possible modes and their coupling under explicit thermal conditions.

Decomposition of the experimental NMR spectra

The experimental broadening (Fig. SI.2†) is by far larger than the one simulated for one single species (containing either one or two P atoms), thus more than one phosphate species must be invoked on γ-Al_2O_3. To identify them, we

used the simulated ^{31}P CS histograms to decompose the experimental ^{31}P CP MAS spectra measured for 5 concentrations (from 0.4 to 4.1 P nm^{-2}, see SI.2†). First, the total area of each simulated histogram of monomers was normalized to 1. For systems containing more than one P atom (dimers and co-adsorption cases), the two histograms were summed up together. Multivariate curve resolution alternating least squares method (MCR-ALS) was used to decompose the experimental NMR spectra (SI.4†). We retained the best solution minimizing the RMSE value (Table SI.3†). At this stage, we would like to stress that is difficult to assess the uniqueness of this solution, since there is no possibility to check our numerical approach against results from well-known samples of phosphate species adsorbed on γ-Al_2O_3. Nevertheless, firstly, according to our results (described in what follows), monomer species are not sufficient to recover the experimental signal (*i.e.*: we need to consider co-adsorbates and dimers). Secondly, the CS ranges of monomer and dimer species are well differentiated (see Table SI.2†). One might only question the species R(110) ν_2 and R(110) ν_3 which have rather close histograms (Fig. SI.3† and 3) and so their relative ratio is rather difficult to predict. Finally, the trends observed for the solution with minimized RMSE value are chemically sound as analyzed in the following.

The overall shape of the 5 spectra (broadening and skew to the negative values) is well rendered, while the CS of maximum position evolves as a function of P coverage (Fig. SI.6†). At the lowest P concentration (0.4 P nm^{-2} in Fig. 3a), various monomeric species are required to match the experimental spectrum, whereas for the highest concentration (4.1 P nm^{-2} in Fig. 3b) one monomeric and one dimeric species are mainly contributing.

Evolution of phosphate species as a function of concentration

Thanks to the previous decomposition, a more quantitative trends of surface chemical speciation can be provided as shown in Fig. 4, which reveals a chemically sound evolution of their surface concentration as function of the P

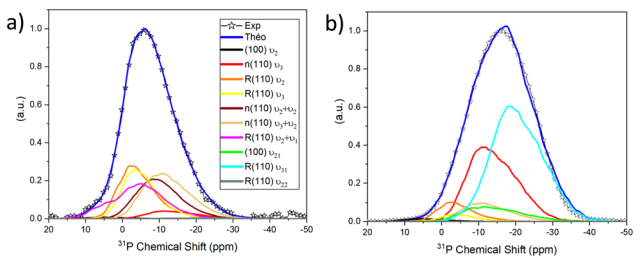

Fig. 3 Decomposition of the experimental NMR spectra by using the theoretical histograms of 4 main species obtained with the AIMD-ML approach for two relevant P concentrations: a) 0.5 P nm^{-2}, b) 4.1 P nm^{-2}. For improving the fit with experimental spectra and enhancing the clarity, the simulated histograms have been smoothed, by starting from raw histograms like those reported in Fig. 2 and SI.3.†

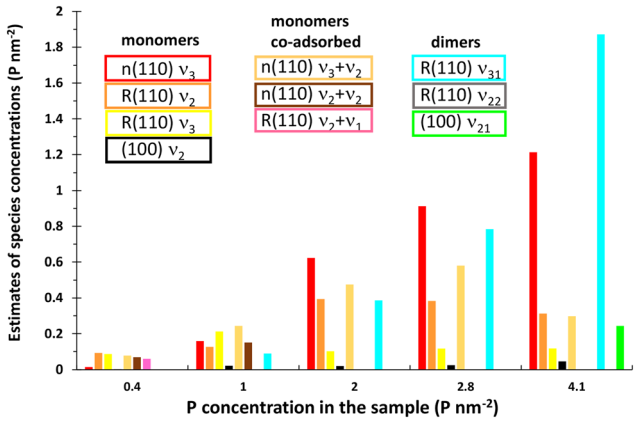

Fig. 4 Quantitative estimates of the surface concentration for each phosphate species adsorbed on the three γ-alumina facets (in P nm^{-2}) and for each total P surface concentration. Their relative contributions were calculated from the deconvolution of the ^{31}P NMR lines as shown in Fig. 3 and SI.6† and scaled from the knowledge of the total P surface concentration (in P nm^{-2}).

concentration. Since all phosphorus sites are surrounded by multiple hydroxyl groups in the present conditions of mild drying (hydroxylation of alumina surfaces remains high), we believe that the ^{31}P CP signals provide reliable estimates of the relative contributions of each species. In other words, we expect that the efficiency of CP transfer is consistent for all surface species. For the two samples with the lower P concentrations (0.4 and 1 P nm^{-2}), almost all monomeric species (including co-adsorbed ones and excepting the (100) ν_2) are present with similar proportions (around 15–25% each). For the three samples with P concentrations of 2, 2.8 and 4.1 P nm^{-2}, 4 species are predominant including 3 monomers (n(110) ν_3, R(110) ν_2 and n(110) ν_3 + ν_2) and one dimer, R(110) ν_{31}, being located on the two (110) facets. For 4.1 P nm^{-2}, n(110) ν_3 and R(110) ν_{31}, whose proportions are about 30% and 46%, respectively, become the most predominant. By contrast, the monomeric species (100) ν_2 remains negligible for all P concentrations, whereas the (100) ν_{21} appears only at the highest P concentration (4.1 P nm^{-2}). This implies that phosphates are preferentially adsorbed on the (110) rather than on the (100) facet. Indeed, on γ-Al_2O_3 crystallites, the (110) facets are expected to be predominant and exhibit a wider diversity of Al and hydroxyl sites.[31,37] In particular, the tetrahedral Al sites present on this facet are involved in the phosphate-alumina bonding of the four predominant species, which is fully consistent with our previous experimental 2D DNP enhanced through-bond ^{31}P–^{27}Al INEPT correlation spectra,[12] but can be here extracted out of 1D CP MAS NMR spectra.

The n(110) ν_3 monomer, either isolated or co-adsorbed with n(110) ν_2, dominate for P coverages equal to 2 and 2.8 P nm^{-2}. These n(110) ν_3 species continuously increases for samples with 0.4 up to 4.1 P nm^{-2}, whereas the R(110) ν_2 species reaches a plateau before declining for 4.1 P nm^{-2}. This decrease is coherently associated to the formation of the dimeric R(110) ν_{31} species located on the same facet which continuously increases from 0 to 4.1 P nm^{-2} and becomes predominant for 4.1 P nm^{-2}. When increasing the P

concentration, the experimental NMR spectrum shifts to negative values. In Fig. SI.2,† a shift of about −10 ppm is observed between 1.0 and 4.1 P nm^{-2}. This trend is obviously related to the apparition of a dimeric species and demonstrates quantitatively the higher prevalence of dimers at higher P concentration. We also notice that another dimeric species (100) v_{21} appears at 4.1 P nm^{-2} resulting from the possible saturation of the (110) facets.

To assess the impact of the AIMD-ML approach to determine the spectra broadening, we achieve a MCR-ALS decomposition of the experimental by using a simple gaussian component centered on the 0 K DFT CS value with a fixed full width at half maximum (FWHM) of 6 ppm for each species (Fig. SI.7†). We firstly notice that RMSE values degrade with respect to the AIMD-ML approach (Table SI.4†). Moreover, if we except the dimeric R(110) v_{31} species, the other predominant species differ significantly from the previous ones (Fig. SI.8†). The origins of these differences must be found in the values of the chemical shifts determined at 300 K, as well as in the asymmetrical components calculated from AIMD-ML. Those asymmetrical components give more constraints for fitting the experimental signal which reveal different species. This justifies the key interest of using the AIMD-ML approach.

Possible impacts for catalysis

The use of computational chemistry to apprehend phenomenon linked to the catalyst preparation is a rather challenging task.[38] Since phosphates additives are generally introduced during the preparation steps and remain present all along the catalyst life cycle, it is relevant to discuss how the previous results could qualitatively influence catalytic properties.

In our previous DFT-NMR work,[12] we already proposed some possible implications of the nature of the phosphate species adsorbed on the γ-alumina (110) and (100) surfaces for the supported catalysts. We underlined the fact that once Al–O–P bonds are formed, this may directly impact the reactivity of alumina surface. One the one hand, as it has been analyzed by infra-red spectroscopy,[39] the presence of such adsorbed phosphates will modify the surface acidity in terms of number and strength of Al Lewis sites and OH Brønsted sites. Moreover, we expect that adsorbed phosphates will also interact with the metallic sites (M) of the active phase through Al–O–P–O–M bonds. This interaction could play a role at the various stage of the catalysis preparation: during the impregnation/activation/reduction steps and in reaction conditions. As discussed in previous experimental works,[7,8] this interaction will modify the physico-chemical properties of the metallic active phase, its activation mechanism and its resulting dispersion. In the present work, thanks to the ML model and the NMR spectra decomposition, we reveal that the phosphate species are predominantly located on the (110) surfaces in close interaction with Al_{IV} sites. This trend is qualitatively in line with some previous experimental investigations showing that the various orientations of α-alumina single crystals,[2] may tune the interaction strength of phosphate: the α-alumina surface orientation containing Al_{IV} site being identified also as one inducing the highest dispersion of phosphate species.

According to previous work, those Al_{IV} sites might be at the origin of the destabilization of the γ-alumina surface in hydrothermal conditions.[40] If phosphates are now interacting with those sites, this location may thus protect the surface by preventing Al_{IV} sites from the hydrolysis attack of water molecules. Similar proposals have been previously made for justifying the improved thermal stability of palladium oxidation catalysts.[9] Recent NMR-DFT works have highlighted that such Al_{IV} sites could also be located on the edges of γ-alumina nanoparticles.[41] Hence, if some phosphates species are anchored on these Al_{IV} sites located on edges, they might also prevent the γ-alumina edges from being also attacked in hydrothermal conditions. Finally, the distribution of phosphate species is expected to be sensitive to a change of the alumina nanoparticles' morphologies as it is often determined at the preparation step.

Conclusions

In summary, by using a DFT data set of more than 1300 structures, we built a ML model to predict ^{31}P CS of phosphate monomeric and dimeric species adsorbed on γ-alumina surfaces with a reasonable accuracy of 2.2 ppm. Combining this ML model with AIMD simulations, we simulated frequency histograms of the observed CS values for 10 relevant phosphate species. Using these theoretical frequency histograms for the mathematical decomposition of the one-dimensional ^{31}P CP MAS NMR spectra allows isolating and quantifying each species unlike conventional MCR-ALS approaches using experimental references. In particular, 4 species (3 monomeric phosphates and 1 dimer) located on the (110) facets (either reconstructed or not), account for the population of surface species observed on the γ-alumina powders with P concentrations from 2 to 4.1 P nm^{-2}. The (100) surface plays a negligible role. This analysis highlights also a chemically relevant evolution of the monomeric *vs.* dimeric species as a function of P concentration. From a more fundamental aspect, this study reveals that the NMR spectra broadening originates from two main types of structural disorder: on the one hand, the contribution of the 3 different phosphate species adsorbed on the alumina surface and, on the other hand, the distribution of local conformations associated to each individual species. The latter can only be captured by AIMD at finite temperature.

We hope that this quantified analysis of the nature and location of the adsorbed phosphates species may help to better understand the behavior of the alumina surface of heterogeneous catalysts synthesized in presence of phosphorus additives. Firstly, it would be interesting to apply this workflow, combining experiments with first principles calculations and machine learning approaches, to surface science studies on well-defined samples.[2] This would improve the reliability of the model and a quantification of

thermal effects on the CS. This approach could also be further applied to study subtle effects of alumina crystallites' morphology (including edges[41]) on the nature and location of phosphates species. This AIMD-ML methodology could probably be generalized to determine other NMR parameters (such as quadrupolar constants or anisotropic shift) and to decompose 1D and 2D NMR of other nuclei such as ^{1}H, ^{17}O, ^{27}Al present on the alumina surface or any other systems. Lastly, a similar approach can be valuable for other spectroscopic analyses such as infrared or X-ray absorption spectroscopies.

Methods

DFT calculations

AIMD simulations were performed with the VASP simulation package, in the canonical <NVT> ensemble, at 300 K, using an Anderson thermostat with 0.1 collision probability and 1 fs timestep. The dynamics of these systems were equilibrated during 11 ps and results were computed from 35 ps trajectories. Electron wavefunctions were calculated with the PBE exchange–correlation functional[42] and D3 Grimme corrections.[33] The cut off was set to 300 eV, the *k*-grid to (2 × 2 × 1) - which corresponds to a reciprocal space sampling of 0.04 $Å^{-1}$, and the energy convergence criterion for self-consistent field calculations was set to 10^{-5} eV per cycle. For the chemical shift DFT calculations at 0 K and all other DFT parameters and γ-alumina (γ-Al_2O_3) surface models related to the screening of the adsorption structures, the reader may refer to our previous work.[12] The accuracy of chemical shift calculations at 0 K was also assessed in the same previous work, where we compared our theoretical results with reference solid aluminophosphates materials and found out that the expected error in our DFT CS calculations is within the order of 2 ppm.

Machine learning model

The data set is split in test and training subsets with a ratio of 0.2, the test subset is left untouched and only used for reporting the final performance of the model. A five-fold cross-validation is applied on the 80% training set for hyperparameter tuning. The final machine learning (ML) model is used to generate the chemical shifts histograms along AIMD trajectories by computing their value at each AIMD time step.

The ML model was trained using a ridge regularized least-square regression and local many-body tensor representation (LMBTR) descriptors. The initial set of LMBTR descriptors contained vectors of 3100 dimensions. We reduced overfitting by reducing the size of these descriptors (see Fig. SI.1† showing that the difference between the training set RMSE and the test set RMSE gets smaller when using less descriptors). Firstly, we removed components of null variance (992 descriptors) and then removed descriptors with the lowest contribution to the model, keeping 500 of the initial set of 3100 descriptors. Detailed parameters are provided in Table SI.1.† These calculations were performed with an in-house python code and the scikit-learn and Dscribe libraries.[34,43]

Data availability

The data supporting this article have been included as part of the ESI.† The 10 upmost relevant structures used for the start of the AIMD simulations (Fig. 1) are available in ref. 12.

Author contributions

A. H. contributed to data curation, formal analysis, investigation, methodology, validation, visualization, writing original draft. T. J. contributed to data curation, formal analysis, methodology and visualization. M. C. V. contributed to data curation, formal analysis, investigation, methodology, validation, visualization, writing original draft and supervision. M. R. contributed to data curation, formal analysis, investigation, methodology, validation, visualization, and review. A. L. contributed to formal analysis, validation, writing review. C. M. and P. R. contributed to formal analysis, investigation, methodology, visualization, writing original draft and review, supervision, funding acquisition and project administration.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

This work is part of the "RatiOnAl Design for CATalysis" (ROAD4CAT) industrial chair, project IDEXLYON funded by the French National Research Agency (ANR-16-IDEX-0005) and the Commissariat-General for Investment (CGI) within the framework of Investissements d'Avenir program ("Investment for the future"). The authors thank the SYSPROD project and AXELERA Pôle de Compétitivité for financial support (PSMN Data Center). Calculations were performed using HPC resources (Jean Zay and Occigen) from GENCI-CINES (Grant A0020806134) and ENER 440 from IFP Energies nouvelles.

References

1 (*a*) P. Euzen, P. Raybaud, X. Krokidis, H. Toulhoat, J. Le Loarer, J. L. Le Loarer and C. Froidefond, in *Handbook of Porous Solids*, ed. F. Schüth, K. S. W. Sing and J. Weitkamp, Wiley-VCH Verlag GmbH, Weinheim, 2002, pp. 1591–1677; (*b*) H. Toulhoat and P. Raybaud, *Catalysis by Transition Metal Sulphides. From Molecular Theory to Industrial Application*, Editions Technip, Paris, 2013.

2 R. Garcia de Castro, J. Bertrand, B. Rigaud, E. Devers, M. Digne, A.-F. Lamic-Humblot, G. Pirngruber and X. Carrier, *Chem. – Eur. J.*, 2020, **26**, 14623.

3 W. Li, X. Feng, Y. Yan, D. L. Sparks and B. L. Phillips, *Environ. Sci. Technol.*, 2013, **47**, 8308.

4 T. Georgelin, M. Jaber, H. Bazzi and J.-F. Lambert, *Origins Life Evol. Biospheres*, 2013, **43**, 429.
5 J. Viger-Gravel, F. M. Paruzzo, C. Cazaux, R. Jabbour, A. Leleu, F. Canini, P. Florian, F. Ronzon, D. Gajan and A. Lesage, *Chem. – Eur. J.*, 2020, **26**, 8976.
6 (*a*) S. Eijsbouts, J. N. M. Van Gestel, J. A. R. Van Veen, V. H. J. De Beer and R. Prins, *J. Catal.*, 1991, **131**, 412; (*b*) L. van Haandel, G. M. Bremmer, E. Hensen and T. Weber, *J. Catal.*, 2017, **351**, 95; (*c*) A. Vikár, H. E. Solt, G. Novodárszki, M. R. Mihályi, R. Barthos, A. Domján, J. Hancsók, J. Valyon and F. Lónyi, *J. Catal.*, 2021, **404**, 67.
7 O. Poulet, R. Hubaut, S. Kasztelan and J. Grimblot, *Bull. Soc. Chim. Belg.*, 1991, **100**, 857.
8 M. H. Woo, J. M. Cho, K.-W. Jun, Y. J. Lee and J. W. Bae, *ChemCatChem*, 2015, **7**, 1460.
9 J. Dong, J. Wang, J. Wang, M. Yang, W. Li and M. Shen, *Catal. Sci. Technol.*, 2017, **7**, 5038.
10 T. van Cleve, D. Underhill, M. Veiga Rodrigues, C. Sievers and J. W. Medlin, *Langmuir*, 2018, **34**, 3619.
11 B. Reif, S. E. Ashbrook, L. Emsley and M. Hong, *Nat. Rev. Methods Primers*, 2021, **1**, 2.
12 A. Hühn, D. Wisser, M. Corral Valero, T. Roy, M. Rivallan, L. Catita, A. Lesage, C. Michel and P. Raybaud, *ACS Catal.*, 2021, **11**, 11278.
13 W. Zhang, S. Xu, X. Han and X. Bao, *Chem. Soc. Rev.*, 2012, **41**, 192.
14 J. Cuny, Y. Xie, C. J. Pickard and A. A. Hassanali, *J. Chem. Theory Comput.*, 2016, **12**, 765.
15 Z. Chaker, M. Salanne, J.-M. Delaye and T. Charpentier, *Phys. Chem. Chem. Phys.*, 2019, **21**, 21709.
16 (*a*) R. Gaumard, D. Dragún, J. N. Pedroza-Montero, B. Alonso, H. Guesmi, I. Malkin Ondík and T. Mineva, *Computation*, 2022, **10**, 74; (*b*) J. B. Kleine Büning and S. Grimme, *J. Chem. Theory Comput.*, 2023, **19**, 3601; (*c*) P. A. Unzueta, C. S. Greenwell and G. J. O. Beran, *J. Chem. Theory Comput.*, 2021, **17**, 826.
17 M. Cordova, M. Balodis, A. Hofstetter, F. Paruzzo, S. O. Nilsson Lill, E. S. E. Eriksson, P. Berruyer, B. Simões de Almeida, M. J. Quayle, S. T. Norberg, A. Svensk Ankarberg, S. Schantz and L. Emsley, *Nat. Commun.*, 2021, **12**, 2964.
18 B. B. Johnson, A. V. Ivanov, O. N. Antzutkin and W. Forsling, *Langmuir*, 2002, **18**, 1104.
19 Y. Kim and R. J. Kirkpatrick, *Eur. J. Soil Sci.*, 2004, **55**, 243.
20 W. Li, A.-M. Pierre-Louis, K. D. Kwon, J. D. Kubicki, D. R. Strongin and B. L. Phillips, *Geochim. Cosmochim. Acta*, 2013, **107**, 252.
21 E. Decanio, *J. Catal.*, 1991, **132**, 498.
22 E. R. H. van Eck, A. P. M. Kentgens, H. Kraus and R. Prins, *J. Phys. Chem.*, 1995, **99**, 16080.
23 W. Li, J. Feng, K. D. Kwon, J. D. Kubicki and B. L. Phillips, *Langmuir*, 2010, **26**, 4753.
24 J. R. Yates, C. J. Pickard and F. Mauri, *Phys. Rev. B*, 2007, 76.
25 F. Pourpoint, A. Kolassiba, C. Gervais, T. Azaïs, L. Bonhomme-Coury, C. Bonhomme and F. Mauri, *Chem. Mater.*, 2007, **19**, 6367.
26 (*a*) D. M. Dawson and S. E. Ashbrook, *J. Phys. Chem. C*, 2014, **118**, 23285; (*b*) D. M. Dawson, J. M. Griffin, V. R. Seymour, P. S. Wheatley, M. Amri, T. Kurkiewicz, N. Guillou, S. Wimperis, R. I. Walton and S. E. Ashbrook, *J. Phys. Chem. C*, 2017, **121**, 1781; (*c*) D. M. Dawson, V. R. Seymour and S. E. Ashbrook, *J. Phys. Chem. C*, 2017, **121**, 28065.
27 (*a*) A. P. Bartók, R. Kondor and G. Csányi, *Phys. Rev. B*, 2013, 87; (*b*) S. De, A. P. Bartók, G. Csányi and M. Ceriotti, *Phys. Chem. Chem. Phys.*, 2016, **18**, 13754.
28 C. E. Rasmussen and C. K. I. Williams, *Gaussian process for machine learning*, The MIT Press, London, England, 2006.
29 T. Ohkubo, A. Takei, Y. Tachi, Y. Fukatsu, K. Deguchi, S. Ohki and T. Shimizu, *J. Phys. Chem. A*, 2023, **127**, 973.
30 R. Wischert, P. Laurent, C. Copéret, F. Delbecq and P. Sautet, *J. Am. Chem. Soc.*, 2012, **134**, 14430.
31 M. Digne, P. Sautet, P. Raybaud, P. Euzen and H. Toulhoat, *J. Catal.*, 2002, **211**, 1.
32 P. E. Blöchl, *Phys. Rev. B*, 1994, **50**, 17953.
33 S. Grimme, J. Antony, S. Ehrlich and H. Krieg, *J. Chem. Phys.*, 2010, **132**, 154104.
34 L. Himanen, M. O. Jäger, E. V. Morooka, F. Federici Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, *Comput. Phys. Commun.*, 2020, **247**, 106949.
35 B. Simões de Almeida, D. Torodii, P. Moutzouri and L. Emsley, *J. Magn. Reson.*, 2023, **355**, 107557.
36 S. Cadars, A. Lesage, C. J. Pickard, P. Sautet and L. Emsley, *J. Phys. Chem. A*, 2009, **113**, 902.
37 (*a*) T. Pigeon, C. Chizallet and P. Raybaud, *J. Catal.*, 2022, **405**, 140; (*b*) M. Digne, P. Sautet, P. Raybaud, P. Euzen and H. Toulhoat, *J. Catal.*, 2004, **226**, 54.
38 M. Corral Valero and P. Raybaud, *J. Catal.*, 2020, **391**, 539.
39 C. Morterra, G. Magnacca and P. P. Demaestri, *J. Catal.*, 1995, **152**, 384.
40 R. Réocreux, É. Girel, P. Clabaut, A. Tuel, M. Besson, A. Chaumonnot, A. Cabiac, P. Sautet and C. Michel, *Nat. Commun.*, 2019, **10**, 3139.
41 A. T. F. Batista, T. Pigeon, J. Meyet, D. Wisser, M. Rivallan, D. Gajan, L. Catita, F. Diehl, A.-S. Gay, C. Chizallet, A. Lesage and P. Raybaud, *ACS Catal.*, 2023, **13**, 6536.
42 J. P. Perdew, K. Burke and M. Ernzerhof, *Phys. Rev. Lett.*, 1996, **77**, 3865.
43 F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and É. Duchesnay, *J. Mach. Learn. Res.*, 2011, **12**, 2825.

