Supplementary Material (SM)

NMR shifts in aluminosilicate glasses via Machine Learning

Authors:
Ziyad Chaker,*[a] Mathieu Salanne,[b] Jean-Marc Delaye,[c] and Thibault Charpentier*[a]
Affiliations:
[a] NIMBE, CEA, CNRS, Université Paris-Saclay, CEA-Saclay, F-91191 Gif-sur-Yvette cedex, FRANCE;
E-mails: zyad.chaker@cea.fr (ziyadchaker@gmail.com) and thibault.charpentier@cea.fr
[b] Maison de la Simulation, USR 3441, CEA, CNRS, INRIA, Université Paris-Sud, Université de Versailles, F-91191 Gif-sur-Yvette, France.
[c] CEA, DEN, Service d'études de vitrification et procédés hautes températures, 30207 Bagnols-sur-Cèze, France.
Corresponding Authors:
***Thibault Charpentier: thibault.charpentier@cea.fr**
***Ziyad Chaker: ziyadchaker@gmail.com (Contact for help on the codes and data usage)**

Data provided in electronic supplementary materials:

In the electronic form of supplementary materials, we provide all our input data in the form of ISOCAR files (DFT-GIPAW σ_{iso} data in the folder "**Diso_ISOCAR_DATA**") and the corresponding VASP POSCAR input structure files (folder "**Glasses_Structures_DATA**"). We also give the codes (folder "**Descriptors_Codes**") necessary to compute SOAP, BPSF and ARDF descriptors (direct input to the ML codes) since the full descriptors files' sizes are too large for uploading in the PCCP platform. The raw data files (folders "**Descriptors_BPSF_ARDF_SOAP_LRR_SiO2_Learning_Curves**", "**Algorithms_LRR_LKRR_GKRR_SOAP_SiO2_LearningCurves_Construction**" and "**Cross-validation_8_ML_Algorithms_RawData**") are provided for constructing the learning curves for the different descriptors (figure 1) and algorithms considered (figure 5) as well as for the algorithms comparison in Table S3 (raw data and python codes to generate the plots). We also provide all the raw data and a python code (folder "**SOAP_Descriptors_CharacterizationSurfaces**") for the production of all SOAP descriptors characterization figures (plots and surfaces of figures 3, 6, 9, S5, S8, S12). Finally, we provide the main ML code constructed and used in our work (MACLAREN.sh and its two corresponding python routines : MALL.py and ML_functions.py in the folder "**Main_MACLAREN_MLCODE**"). Since no documentation is available, yet, for these home-made codes, please contact the corresponding authors (Ziyad Chaker or Thibault Charpentier) of the article for further information on the usage of these codes (ziyadchaker@gmail.com, thibault.charpentier@cea.fr).

Notes:

The learning curves (figures 1 and 5) have been computed using the following combinations of structures provided in supplementary materials: 0KSiO_2-(1 to 10) for (x=10); 0KSiO_2-(1 to 10) and 0KSiO_2-n(1 to 10 for (x=20); 0KSiO_2-(1 to 10) and 0KSiO_2-n(1 to 10 for (x=20) and 300KSiO_2-(1 to 10 for (x=30); and so on until including the 97 SiO_2 structures in the training/validation set. The test set is composed, for all these computations, of the two structures: 2000KSiO_2-n9 and 2000KSiO_2-n10. Note that due to DFT-GIPAW calculations convergence issues, two structures from the 501 used in this work are redundant: NAS3-1 at 0K is the same than NAS3-2 at 0K and 30Na-1 at 1500K is the same than 30Na-2 at 1500K. These two systems can be safely ignored to reproduce the results of our work. Nevertheless one can also use them (as we have done) within the very same ML set (training or validation) with no impact on the results obtained. This will just result in increasing the weight of these structures during the training or validation process.

List of Tables

List of Figures

Table S1 Compositions and the cubic cell dimensions the glasses considered in this work: Vitreous silica $(SiO_2)_{120}$; Sodosilicates x={10 to 50}NS corresponding to $(Na_2O)_x(SiO_2)_{100-x}$; Sodo-aluminosilicates NAS3, NAS4 and NAS5 corresponding to $(Al_2O_3)_{25}(Na_2O)_{25}(SiO_2)_{50}$, $(Al_2O_3)_{17.5}(Na_2O)_{17.5}(SiO_2)_{70}$ and $(Al_2O_3)_{12.5}(Na_2O)_{12.5}(SiO_2)_{75}$, respectively. The column "Total" refers to the total number of structures considered for each glass composition

Glasses	Compositions				Number of atomic structures						Edge
	^{29}Si	^{17}O	^{23}Na	^{27}Al	0K	300K	1000K	1500K	2000K	Total	length (Å)
SiO_2	120	240	-	-	20	19	20	20	20	99	17.5896
10NS	90	190	20	-	10	10	10	10	10	50	16.3522
20NS	80	180	40	-	10	10	10	10	10	50	16.1512
30NS	70	170	60	-	10	10	10	10	10	50	15.9846
40NS	60	160	80	-	10	10	10	10	10	50	15.8610
50NS	50	150	100	-	10	10	10	10	10	50	15.8193
NAS3	50	200	50	50	10	10	10	10	10	50	16.7762
NAS4	70	210	35	35	8	10	10	10	10	48	16.9511
NAS5	75	200	25	25	10	10	10	10	10	50	16.6251
NAS3-L	100	400	100	100	-	2	-	-	-	2	21.1367
NAS4-L	100	400	100	100	-	2	-	-	-	2	21.3570

Table S2 Sizes of the vectors of SOAP descriptors for each oxide glass considered for different numbers n_{max} (= L_{max}) of basis functions used for the descriptor construction. With N_s, the number of species in the system, the SOAP vector of descriptors is calculated as: $(1/4)\cdot L_{max}\cdot(L_{max}+1)^2\cdot N_s\cdot(N_s+1)$

SOAP	Number of elements in the vector of SOAP descriptors		
parameter	SiO_2	Na_2O-SiO_2 (NS)	Al_2O_3-Na_2O-SiO_2 (NAS)
$L_{max} = 2$	27	54	90
$L_{max} = 3$	72	144	240
$L_{max} = 4$	150	300	500
$L_{max} = 5$	270	540	900
$L_{max} = 6$	441	882	1470
$L_{max} = 7$	672	1344	2240
$L_{max} = 8$	972	1944	3240
$L_{max} = 9$	1350	2700	4500

Table S3 ML errors ($\Delta\sigma$ - the FWHM of the ML absolute errors distribution, RMSE - root mean square error and MAE - mean absolute error) obtained for relaxed (0K) and room-temperature (300K) SiO_2 glasses (39 structures in the reference set) NMR σ_{iso} predictions. The results are shown for both nuclei for the different algorithms described in the computational part: linear ridge regression (LRR)[63], kernel ridge regression with a linear kernel (LKRR) and a Gaussian kernel (GKRR)[62], elastic net regression (ENR), random forest regression (RFR), Bayesian ridge regression (BRR), k-nearest neighbors (k-NN) and artificial neural networks (ANN)[63,66]. All calculations are performed with the SOAP descriptor. These estimation are averaged over a 10-fold CV for LRR, BRR, k-NN and ANN while a 3-fold CV is used for ENR, RFR, LKRR and GKRR. The values given between brackets correspond to the resulting CV standard deviations

	^{29}Si nucleus in SiO_2 system								
	Train set			Validation set			Test set		
ML algorithm	$\Delta\sigma$	RMSE	MAE	$\Delta\sigma$	RMSE	MAE	$\Delta\sigma$	RMSE	MAE
LRR	0.75(0.18)	0.75(0.18)	0.58(0.13)	0.88(0.27)	0.88(0.27)	0.63(0.14)	0.84(0.15)	0.84(0.15)	0.65(0.12)
LKRR	0.82(0.18)	0.82(0.18)	0.63(0.13)	0.90(0.16)	0.90(0.16)	0.67(0.12)	0.85(0.16)	0.85(0.16)	0.66(0.13)
GKRR	0.44(0.18)	0.44(0.18)	0.34(0.15)	0.91(0.34)	0.91(0.34)	0.58(0.20)	1.07(0.17)	1.07(0.17)	0.81(0.13)
ENR	1.01(0.06)	1.01(0.06)	0.77(0.04)	1.06(0.09)	1.06(0.09)	0.80(0.05)	1.01(0.04)	1.01(0.04)	0.79(0.04)
RFR	1.28(0.24)	1.28(0.24)	0.89(0.20)	3.01(0.88)	3.05(0.85)	2.25(0.75)	3.61(0.07)	3.64(0.05)	2.86(0.04)
BRR	2.41(0.12)	2.41(0.12)	1.81(0.05)	2.47(0.17)	2.49(0.19)	1.89(0.08)	2.35(0.03)	2.37(0.05)	1.83(0.03)
k-NN	0.00(0.00)	0.00(0.00)	0.00(0.00)	3.80(1.23)	3.92(1.27)	2.31(1.27)	5.30(0.12)	5.37(0.19)	4.29(0.17)
ANN	2.95(0.34)	2.95(0.34)	2.22(0.22)	2.99(0.33)	3.02(0.36)	2.29(0.25)	2.92(0.21)	2.94(0.22)	2.28(0.17)
	^{17}O nucleus in SiO_2 system								
	Train set			Validation set			Test set		
ML algorithm	$\Delta\sigma$	RMSE	MAE	$\Delta\sigma$	RMSE	MAE	$\Delta\sigma$	RMSE	MAE
LRR	1.38(0.09)	1.38(0.09)	1.06(0.06)	1.49(0.11)	1.49(0.11)	1.13(0.08)	1.58(0.09)	1.58(0.09)	1.19(0.06)
LKRR	1.39(0.02)	1.39(0.02)	1.07(0.01)	1.48(0.04)	1.49(0.04)	1.13(0.02)	1.50(0.01)	1.51(0.01)	1.15(0.01)
GKRR	0.72(0.26)	0.72(0.26)	0.55(0.20)	1.23(0.20)	1.23(0.20)	0.86(0.21)	1.64(0.16)	1.64(0.16)	1.23(0.11)
ENR	1.62(0.03)	1.62(0.03)	1.24(0.02)	1.70(0.04)	1.70(0.04)	1.30(0.02)	1.71(0.01)	1.71(0.01)	1.34(0.00)
RFR	1.52(0.16)	1.52(0.16)	1.08(0.18)	3.73(0.79)	3.73(0.79)	2.65(0.69)	4.65(0.07)	4.65(0.07)	3.55(0.06)
BRR	3.97(0.17)	3.97(0.17)	2.97(0.13)	4.09(0.27)	4.10(0.27)	3.09(0.20)	4.03(0.11)	4.03(0.11)	3.04(0.08)
k-NN	0.00(0.00)	0.00(0.00)	0.00(0.00)	5.12(1.37)	5.14(1.38)	2.97(1.50)	7.34(0.13)	7.35(0.13)	5.77(0.11)
ANN	3.03(0.11)	3.03(0.11)	2.29(0.07)	3.16(0.17)	3.16(0.17)	2.37(0.13)	3.25(0.17)	3.25(0.17)	2.37(0.07)

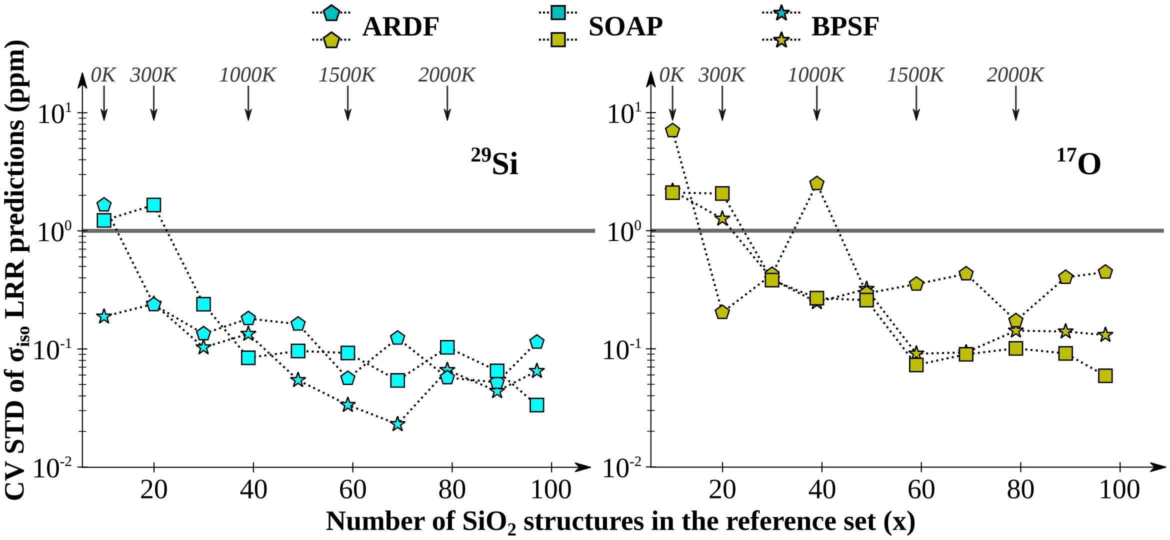

Fig. S1 LRR σ_{iso} predictions CV (cross-validation) standard deviations (STD), obtained from the 10-fold CV process applied to obtain the learning curves reported in figure 1, as a function of the training/validation set size. The horizontal grey line indicates the 1 ppm standard deviation value. The LRR error reported ($\Delta\sigma$) is the FWHM of the distribution of absolute LRR deviations from DFT-GIPAW calculated data. The vertical arrows indicate the temperatures of the systems appended to the training/validation set for increasing values of the reference set size (x). The test set is composed of two SiO_2 structures of the most challenging 2000K ones

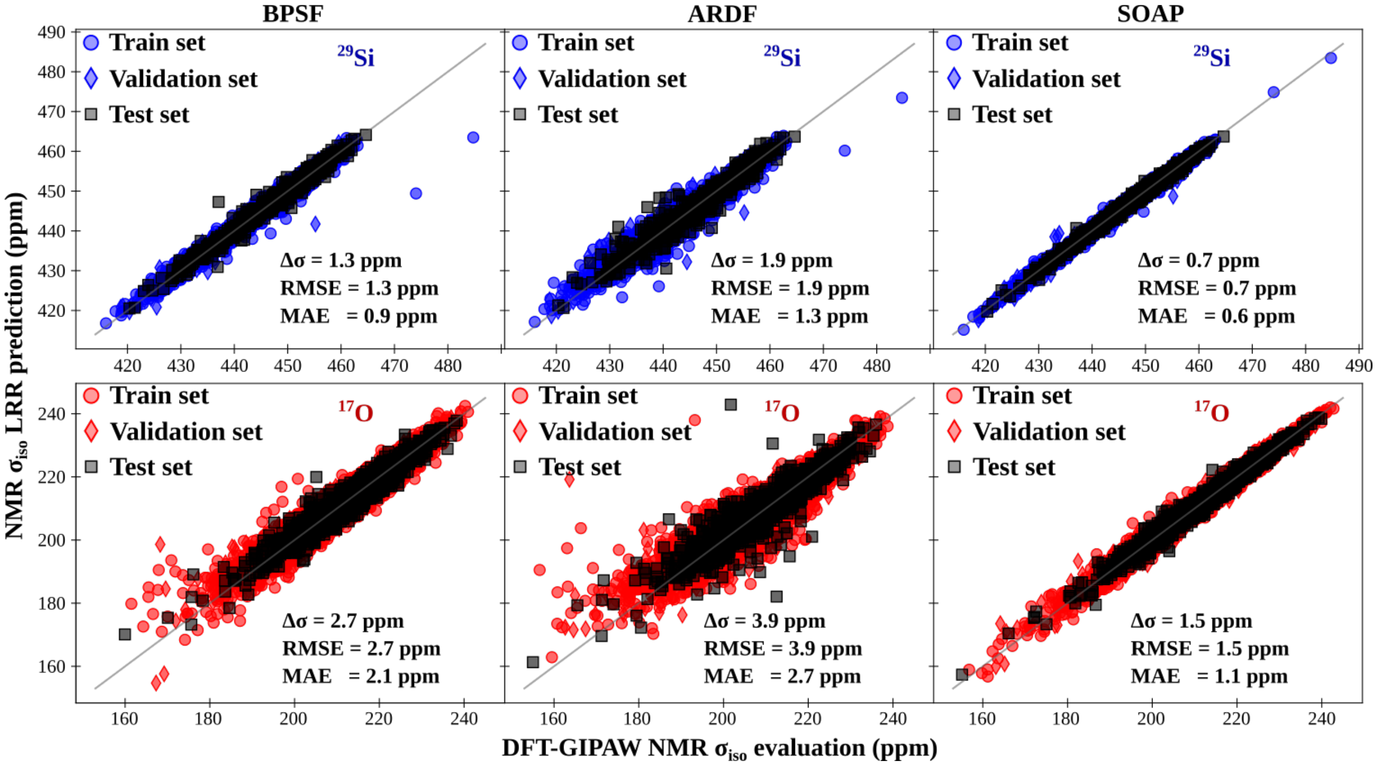

Fig. S2 LRR-NMR isotropic magnetic shielding predictions for 0K and 300K SiO_2 systems (39 structures) for each of the three descriptors considered and the two nuclei involved in these structures. The oblique grey line indicates the exact matching between LRR predictions and DFT estimations. The LRR error reported ($\Delta\sigma$) is the FWHM of the distribution of absolute LRR deviations from DFT-GIPAW calculated data. The corresponding root-mean square errors (RMSE) and mean absolute errors (MAE) are also reported in each case

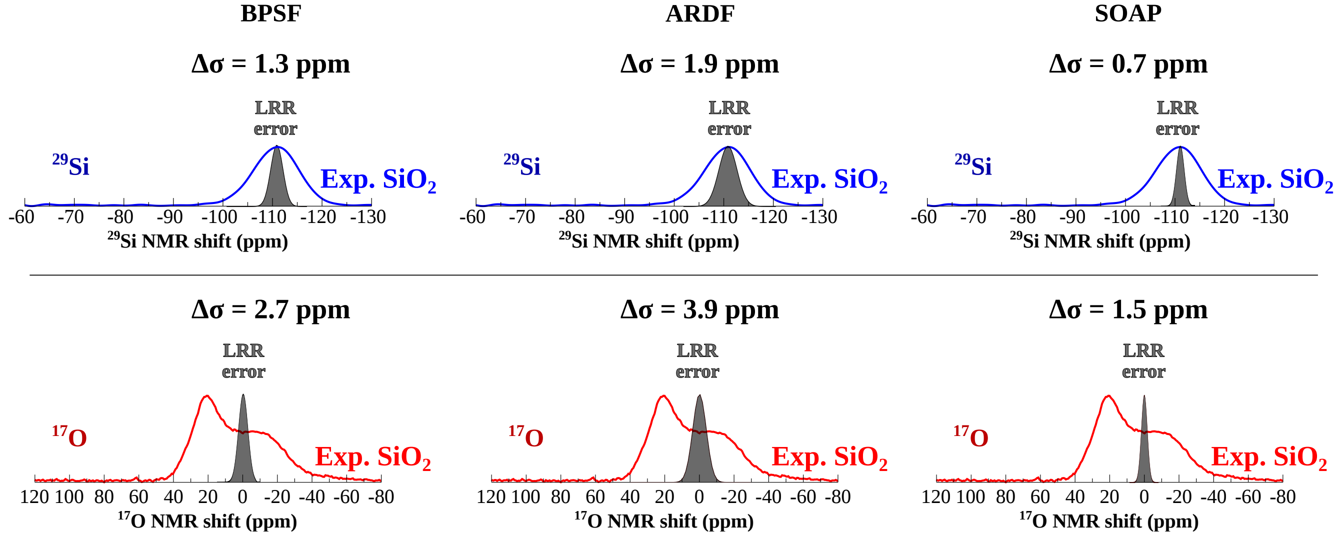

Fig. S3 LRR-SOAP test set error distributions (grey distributions) superimposed with the experimental NMR spectra (colored solid lines) of a typical SiO$_2$ glass. The LRR results are shown for the three different descriptors considered in this work. The LRR error reported ($\Delta\sigma$) is the FWHM of the distribution of absolute LRR deviations from DFT-GIPAW calculated data

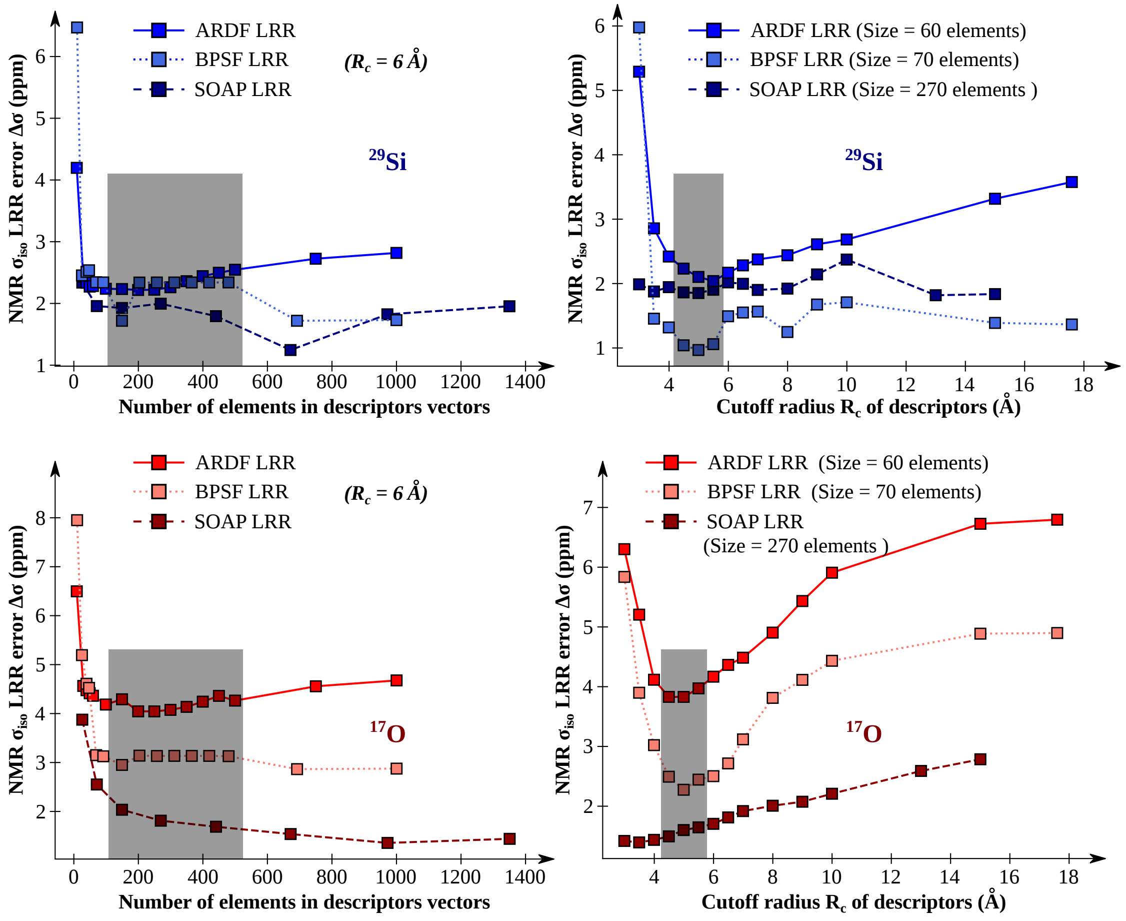

Fig. S4 LRR-NMR σ_{iso} predictions as a function of the absolute (number of elements) descriptor vectors sizes (left panel) and cutoff radius (right panel) for the ARDF, BPSF and SOAP descriptors considered in the case of SiO$_2$ system. The reference set includes only the relaxed SiO$_2$ systems (20 structures) and the LRR error reported ($\Delta\sigma$) is the FWHM of the distribution of absolute LRR deviations from DFT-GIPAW calculated data

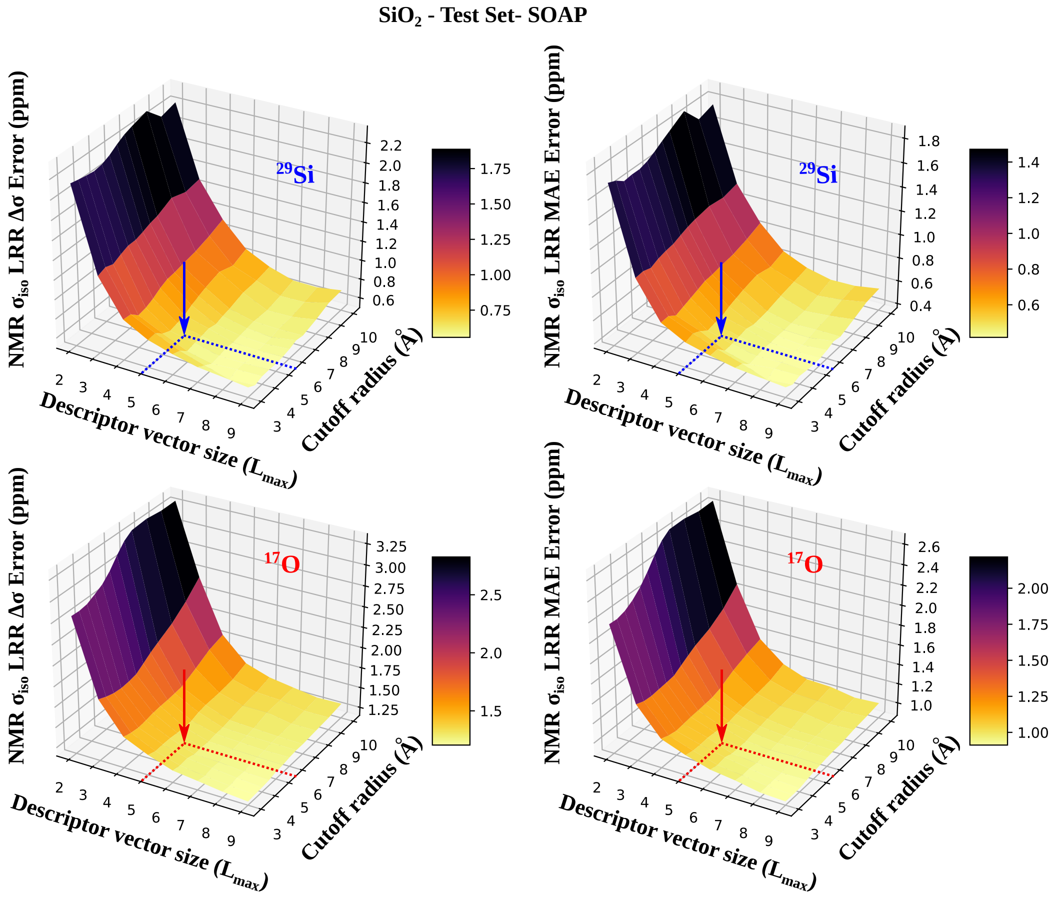

Fig. S5 LRR NMR σ_{iso} predictions errors (Left: $\Delta\sigma$; Right: mean absolute error) as function of the SOAP descriptors sizes (Lmax) and cutoff radius (R$_c$). The reference set considered (19 structures) is composed of all SiO$_2$ systems at room temperature (300K). $\Delta\sigma$ error is the FWHM of the distribution of absolute LRR deviations from DFT-GIPAW calculated data and MAE is the mean absolute error. The vertical arrows indicate the SOAP parameters choice used in most of our work (L$_{max}$ = 5 and R$_c$ = 5.5 Å). The resulting LRR errors $\Delta\sigma$, RMSE and MAE for the points indicated by the arrows are, respectively, 0.7, 0.7 and 0.5 ppm for ^{29}Si; 1.4, 1.4 and 1.1 ppm for ^{17}O

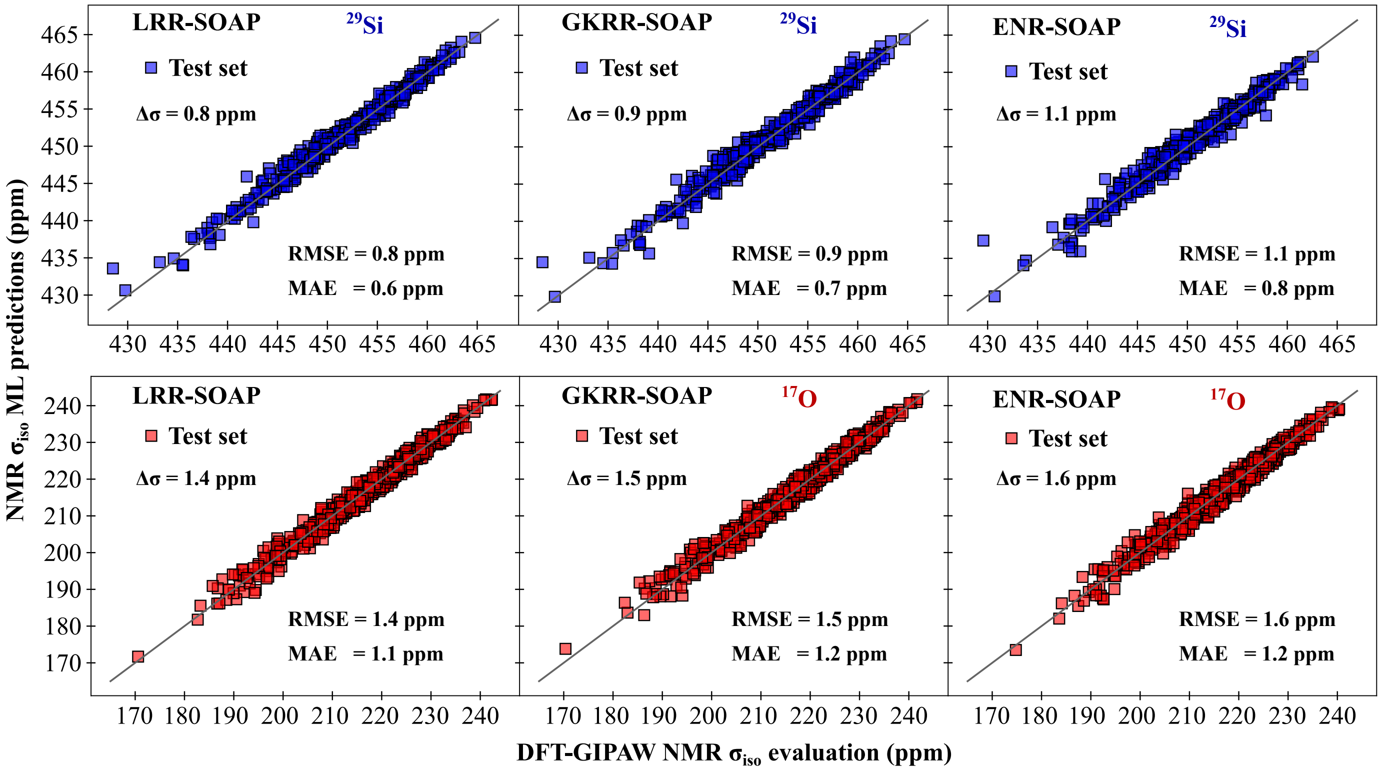

Fig. S6 ML-SOAP vs DFT NMR isotropic magnetic shieldings results for the 300K SiO_2 systems (19 structures) using LRR (left panels), GKRR (central panels) and ENR (right panels) for both species. The ML error reported (Δσ) is the FWHM of the distribution of absolute ML predictions deviations from DFT-GIPAW calculated data. The corresponding root-mean square errors (RMSE) and mean absolute errors (MAE) are also reported in each case

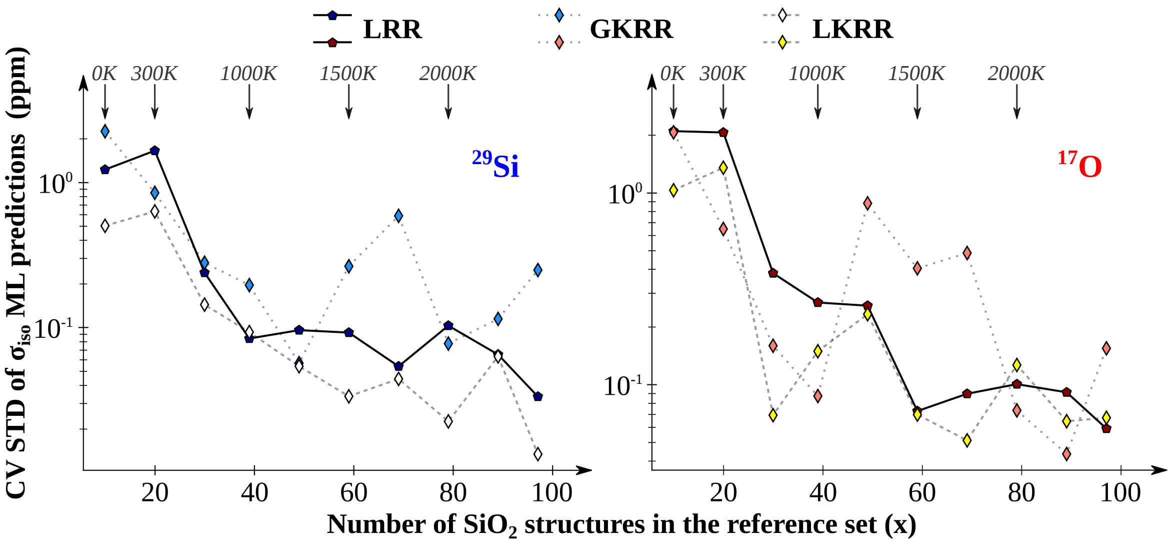

Fig. S7 The ML σ_{iso} predictions standard deviations, obtained from a 10-fold CV for LRR and 3-fold CV for GKRR and LKRR used to obtain the learning curves reported in figure 5, as function of the reference set size. The test ML error reported ($\Delta\sigma$) is the FWHM of the distribution of absolute LRR deviations from DFT-GIPAW calculated data. The vertical arrows indicate the temperatures of the systems appended to the training/validation set for increasing values of the reference set size (x). The test set is composed of two SiO_2 structures of the most challenging 2000K ones

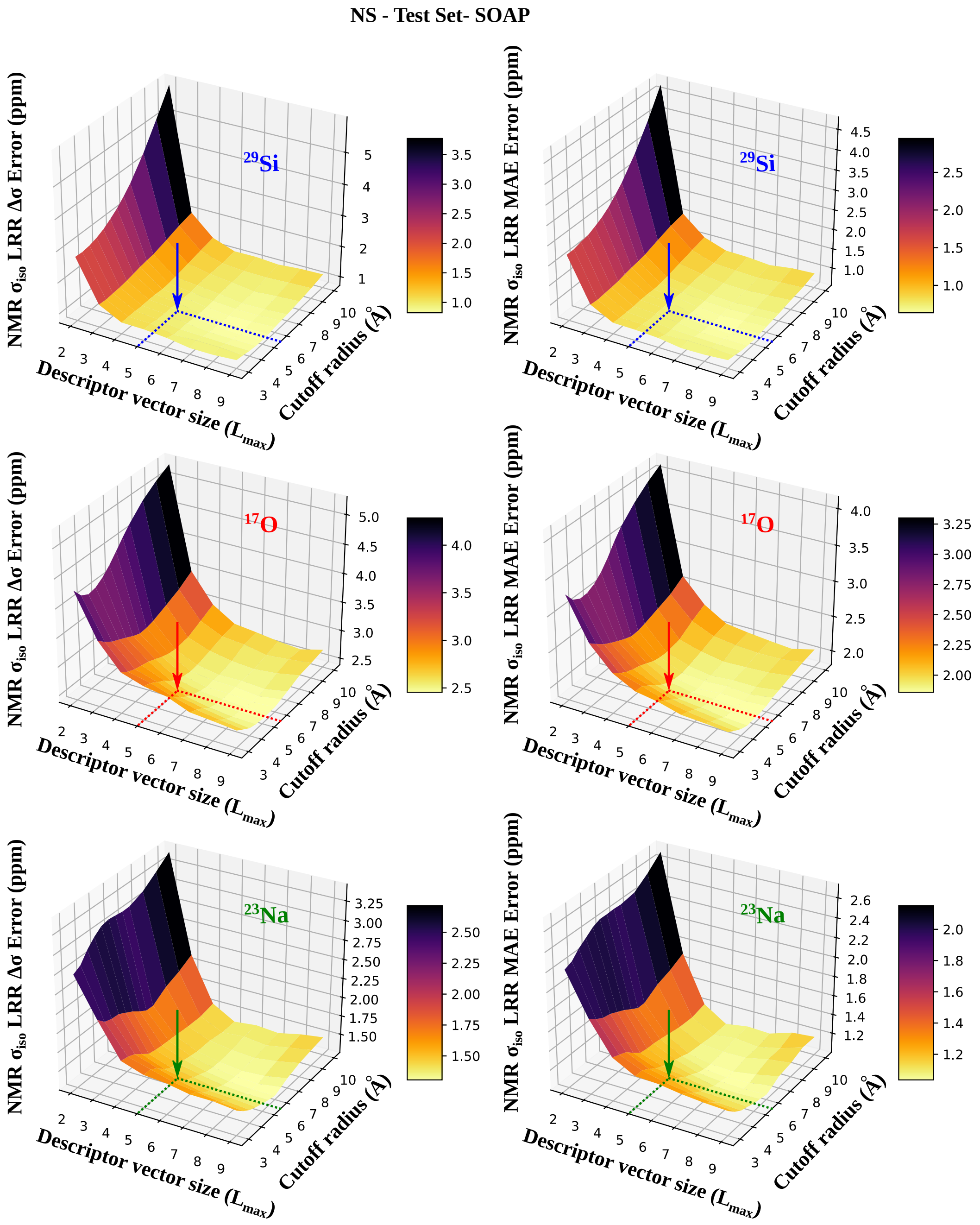

Fig. S8 LRR NMR σ_{iso} predictions errors (Left: $\Delta\sigma$; Right: mean absolute error) as a function of the SOAP descriptors sizes (Lmax) and cutoff radius (R$_c$). The reference set considered (50 structures) is composed of all NS systems at room temperature (300K). $\Delta\sigma$ error is the FWHM of the distribution of absolute LRR deviations from DFT-GIPAW calculated data. The vertical arrows indicate the SOAP parameters choice used in most of our work (L_{max} = 5 and R_c = 5.5 Å). The resulting LRR errors $\Delta\sigma$, RMSE and MAE for the points indicated by the arrows are, respectively, 1.0, 1.0 and 0.8 ppm for ^{29}Si; 2.6, 2.6 and 1.9 ppm for ^{17}O; 1.5, 1.5 and 1.2 ppm for ^{23}Na

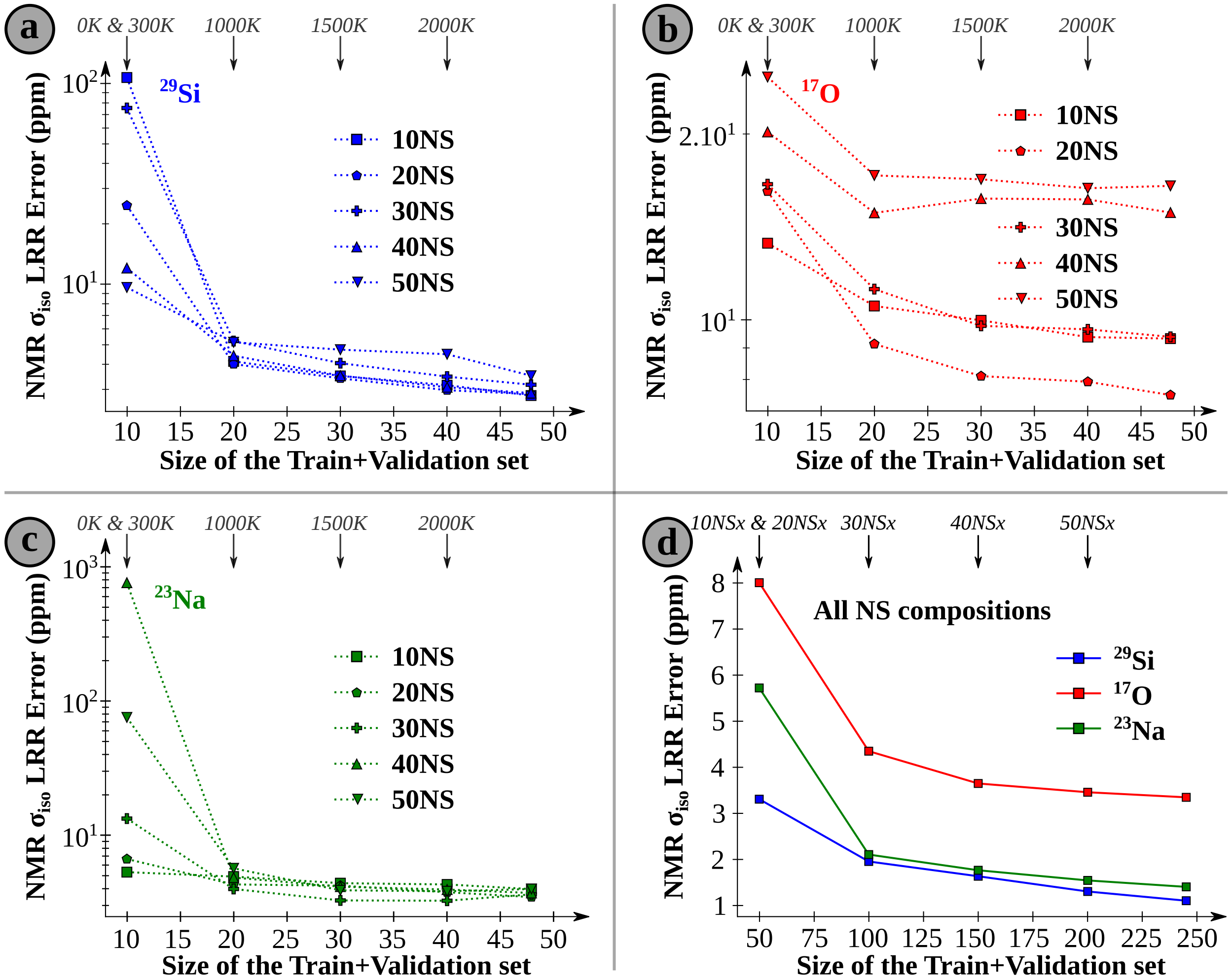

Fig. S9 (a,b,c) Learning curves of LRR-SOAP σ_{iso} predictions for each NS composition ({10 to 50}NS) at all temperatures (0K, 300K, 1000K, 1500K and 2000K) and 2 structures at 2000K as test set. The results are shown for each nucleus: ^{29}Si (a), ^{17}O (b) and ^{23}Na (c). **(d)** Learning Curves for all NS compositions mixed (test set is 5 structures 50NS at 300K). The LRR error reported ($\Delta\sigma$) is the FWHM of the distribution of absolute LRR deviations from DFT-GIPAW calculated data. The vertical arrows indicate the temperatures of the systems appended to the training/validation set for increasing values of the reference set size (x)

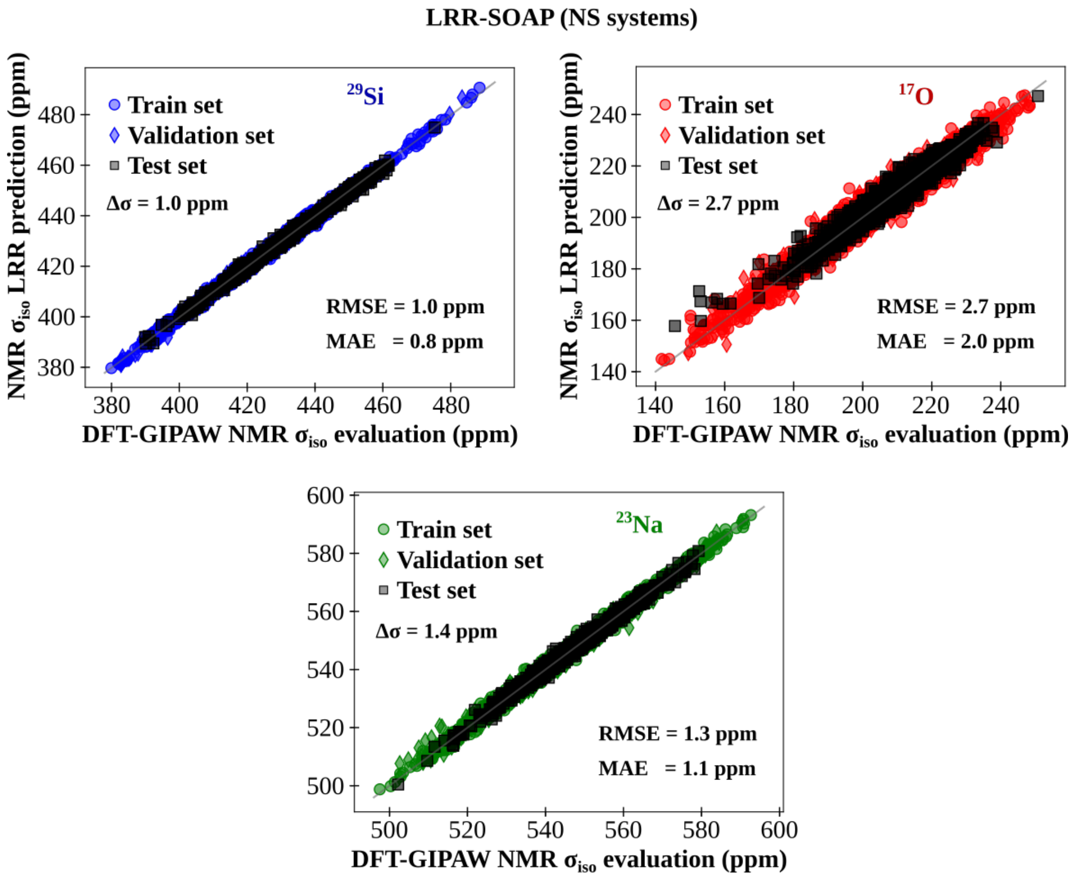

Fig. S10 LRR-SOAP NMR isotropic magnetic shieldings predictions for 0K and 300K NS systems (100 structures of all compositions from 10 to 50 NS mixed in the reference set). The grey oblique line indicates the exact matching between LRR and DFT-GIPAW estimations. The LRR error reported ($\Delta\sigma$) is the FWHM of the distribution of absolute LRR deviations from DFT-GIPAW calculated data. The corresponding root-mean square errors (RMSE) and mean absolute errors (MAE) are also reported in each case

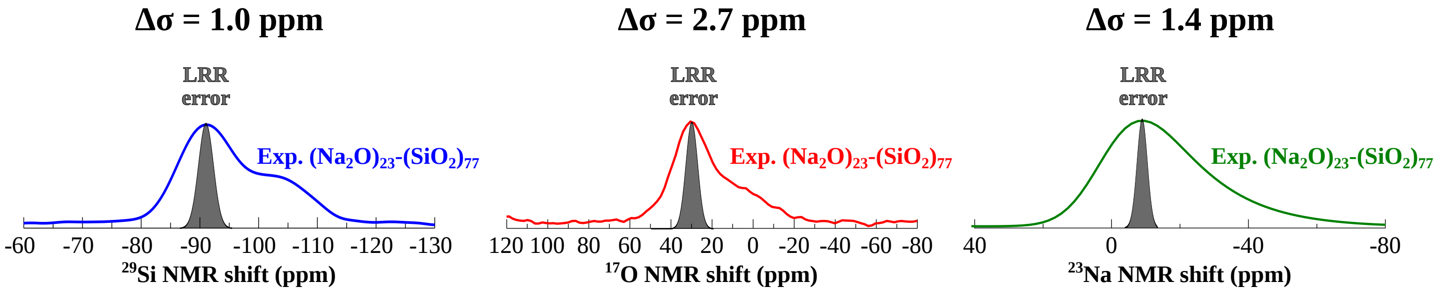

Fig. S11 LRR-SOAP errors distributions (grey Gaussians) of σ_{iso} test set predictions superimposed with the experimental NMR spectra (colored solid lines) of a typical $(Na_2O)_{23}$-$(SiO_2)_{77}$ glass. The LRR error reported ($\Delta\sigma$) is the FWHM of the distribution of absolute LRR deviations from DFT-GIPAW calculated data

NAS - Test Set- SOAP

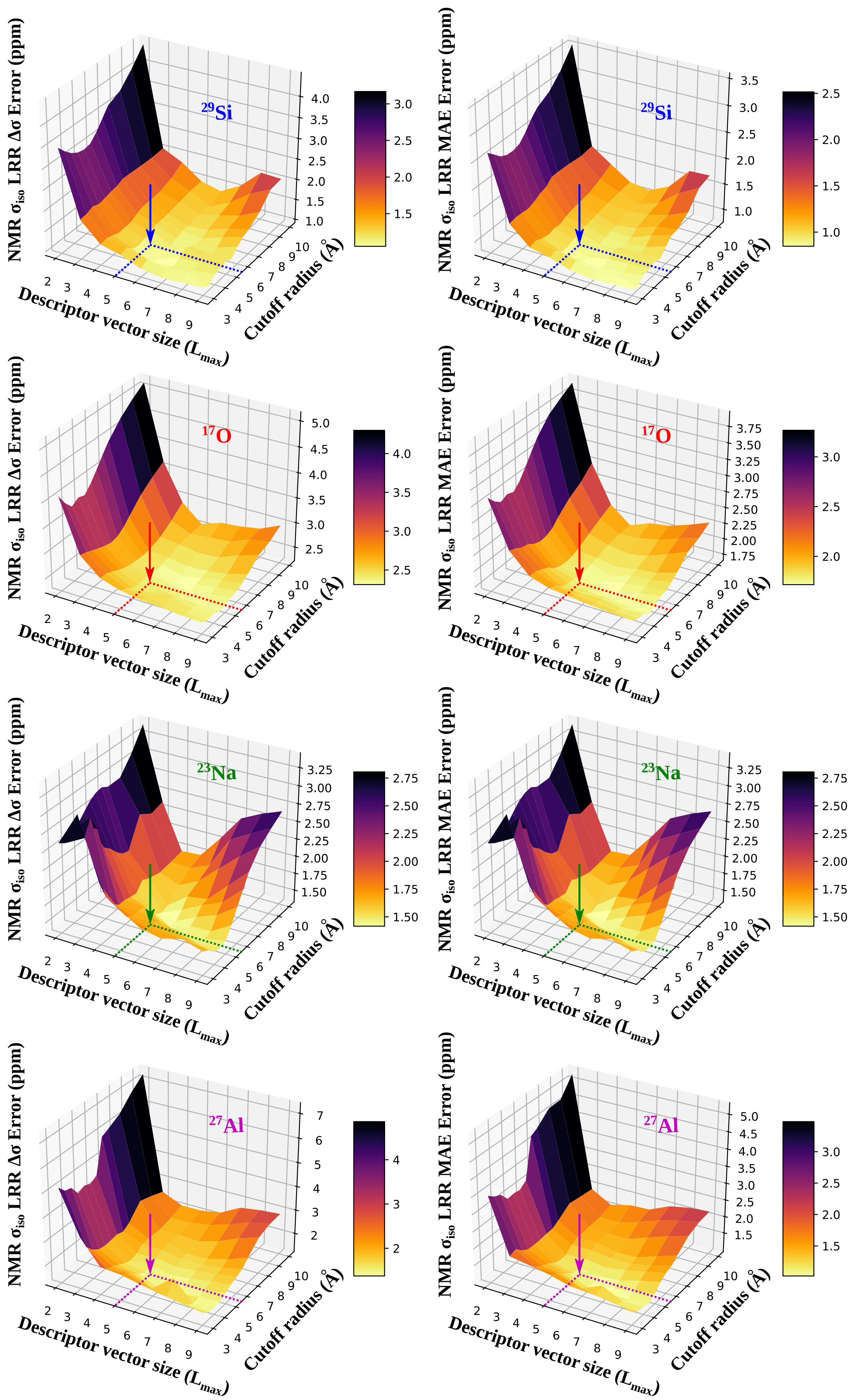

Fig. S12 LRR-SOAP NMR σ_{iso} predictions errors (Left: $\Delta\sigma$; Right: mean absolute error) as a function of the SOAP descriptors sizes (Lmax) and cutoff radius (R$_c$). The reference set (30 structures) considered is composed of all NAS systems at room temperature (300K). $\Delta\sigma$ error is the FWHM of the distribution of absolute LRR deviations from DFT-GIPAW calculated data. The resulting LRR errors $\Delta\sigma$, RMSE and MAE for the points indicated by the arrows region are, respectively, 1.3, 1.3 and 1.0 ppm for ^{29}Si; 2.4, 2.4 and 1.8 ppm for ^{17}O; 1.6, 1.6 and 1.1 ppm for ^{23}Na; 1.6, 1.6 and 1.2 ppm for ^{27}Al

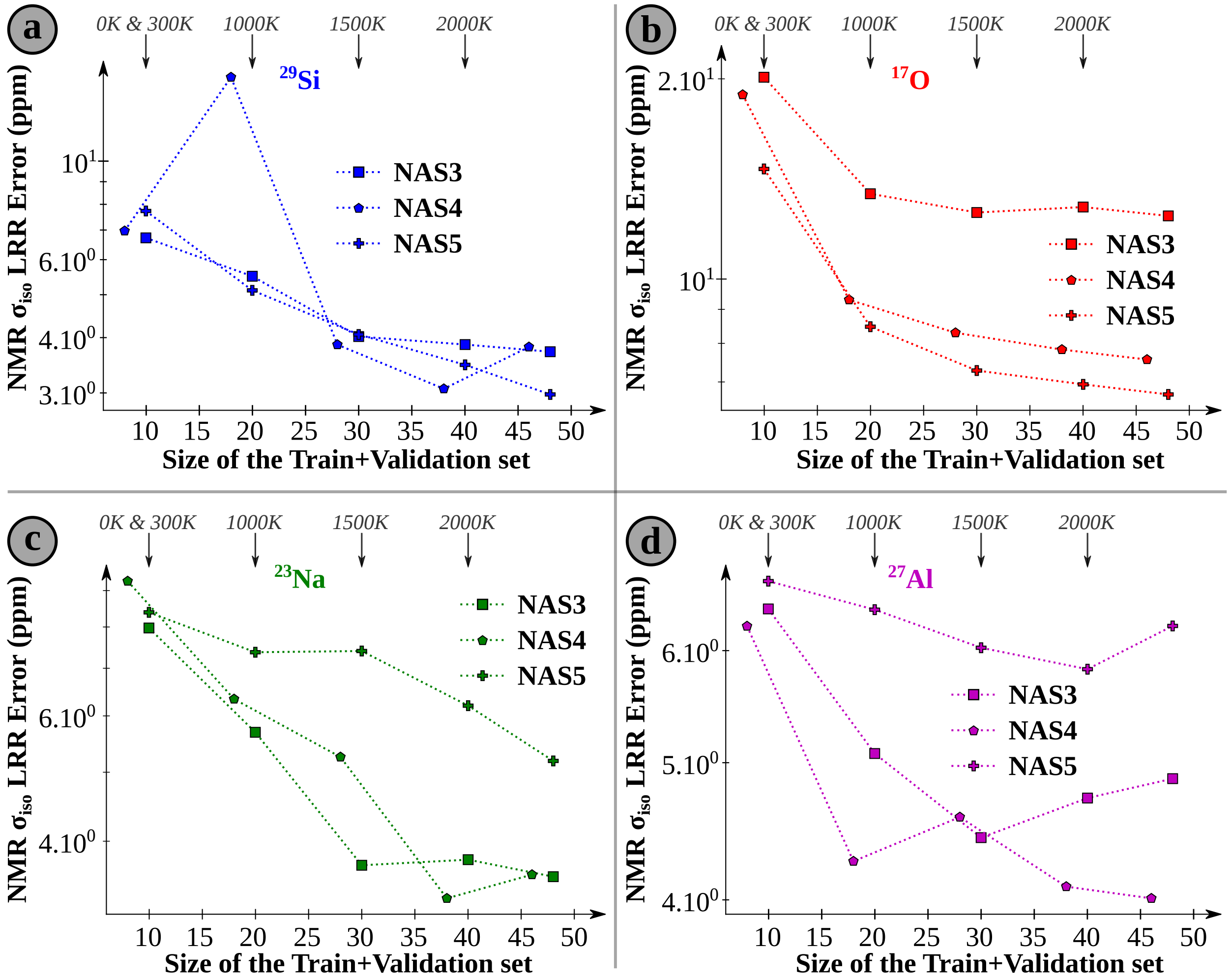

Fig. S13 Learning curves of LRR-SOAP σ_{iso} predictions in the case of sodo-aluminosilicates (each NAS system separately) for ^{29}Si **(a)**, ^{17}O **(b)**, ^{23}Na **(c)** and ^{27}Al **(d)**, for a training/validation set including successively 0K, 300K, 1000K, 1500K and 2000K. The test set is composed of 2 structures at 2000K. The LRR error reported ($\Delta\sigma$) is the FWHM of the distribution of absolute LRR deviations from DFT-GIPAW calculated data

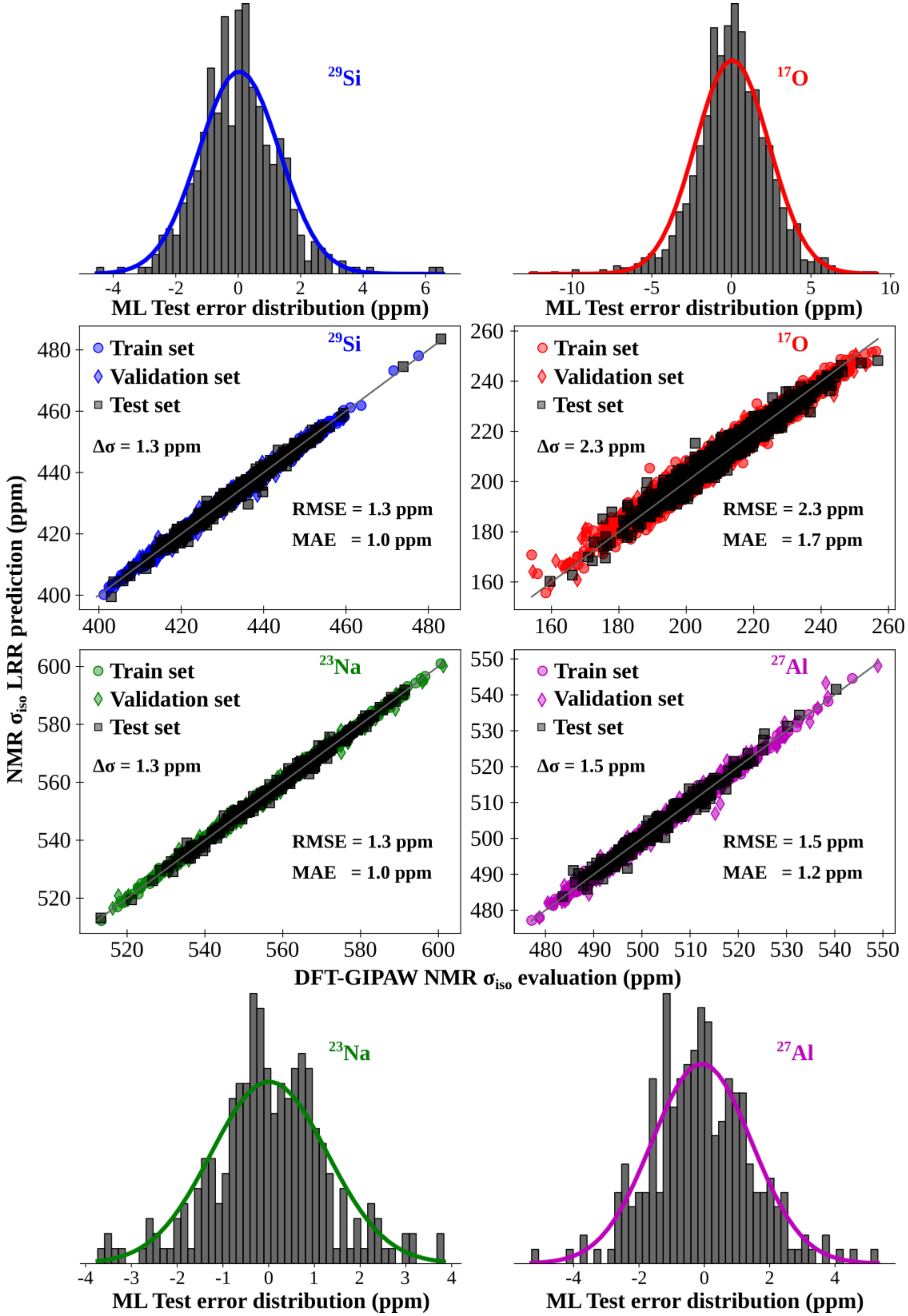

Fig. S14 LRR-SOAP σ_{iso} results for 0K and 300K NAS systems (58 structures of all compositions, NAS3, NAS4 and NAS5 mixed in the reference set). The grey oblique line indicates the exact matching between LRR and DFT-GIPAW estimations and the bar plots show the LRR-SOAP test error distributions fitted by a Gaussian function (solid colored lines). The LRR error reported ($\Delta\sigma$) is the FWHM of the distribution of absolute LRR deviations from DFT-GIPAW calculated data. The corresponding root-mean square errors (RMSE) and mean absolute errors (MAE) are also reported in each case

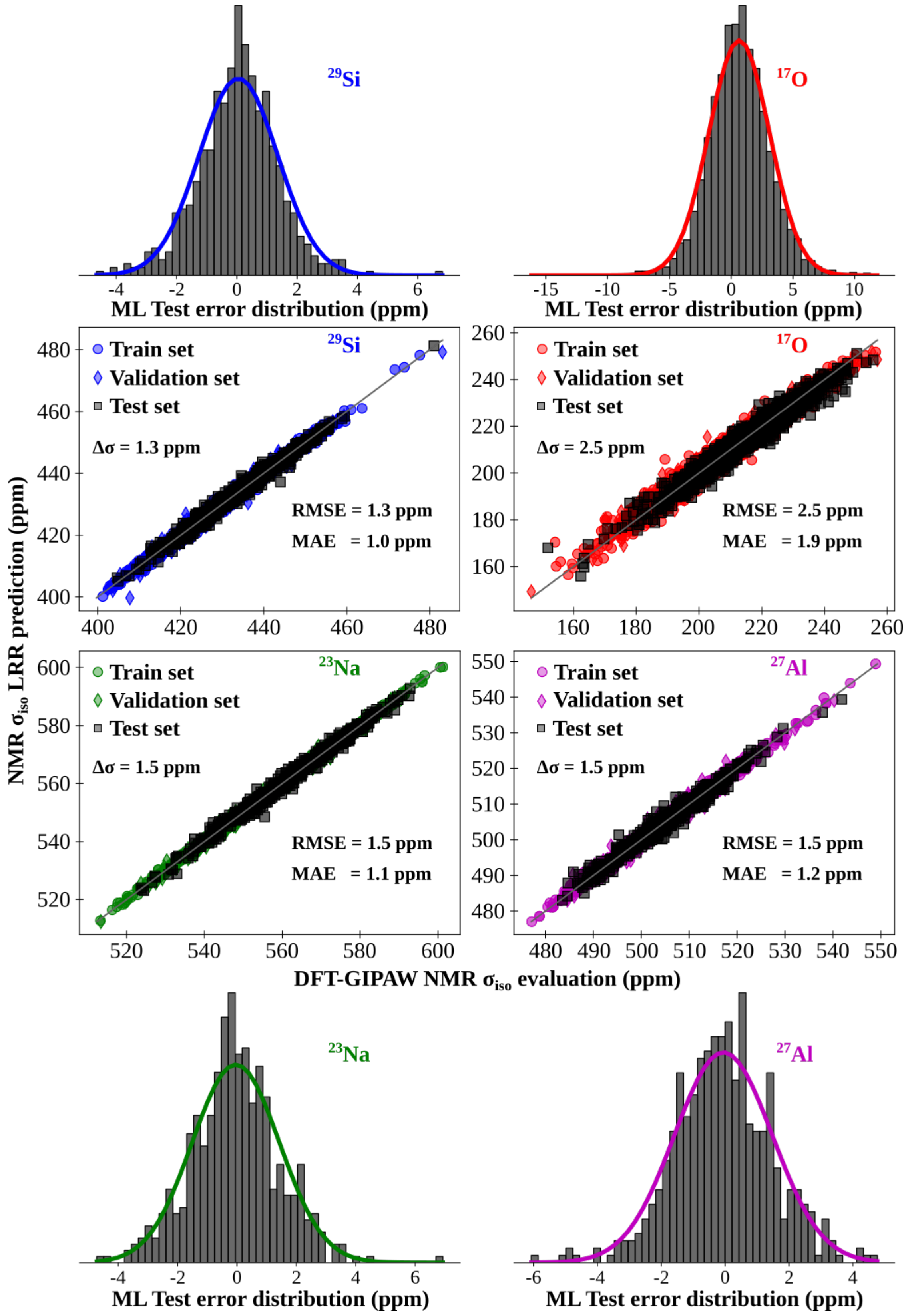

Fig. S15 LRR-SOAP σ_{iso} predictions for 0K and 300K NAS systems (62 structures at all compositions) with a specific test set: four larger structures (two NAS3-L and two NAS4-L of 700 atoms) together with their small size counterpart (two NAS3 and two NAS4 of 350 atoms). The train/validation set is composed of the remaining 54 structures of all NAS compositions. The grey oblique line indicates the exact matching between LRR and DFT-GIPAW estimations and the bar plots show the LRR test error distributions fitted by a Gaussian function (solid colored lines). The LRR error reported ($\Delta\sigma$) is the FWHM of the distribution of absolute LRR deviations from DFT-GIPAW calculated data. The corresponding root-mean square errors (RMSE) and mean absolute errors (MAE) are also reported in each case