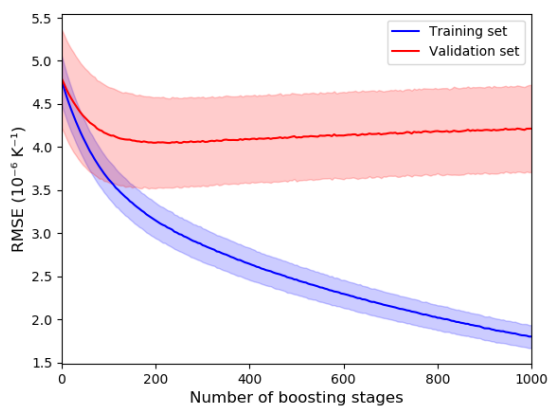


Supporting information for:
Prediction of Thermal Properties of Zeolites
through Machine Learning

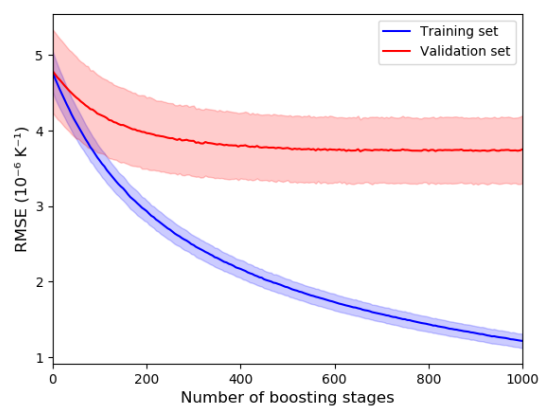
Maxime Ducamp and François-Xavier Coudert*

*Chimie ParisTech, PSL Research University, CNRS, Institut de Recherche de Chimie
Paris, 75005 Paris, France*

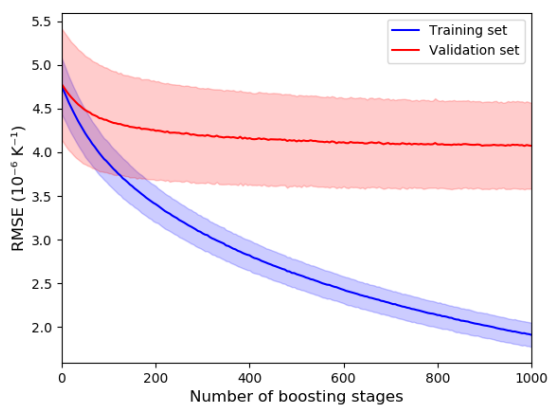
E-mail: fx.coudert@chimieparistech.psl.eu



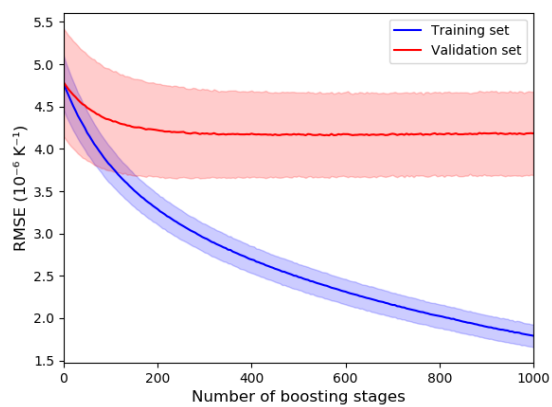
(a)



(b)



(c)



(d)

Figure S1: Learning curves of the thermal expansion using all the types of descriptors for the thermal expansion. Lines correspond to the mean RMSE values and shaded surfaces correspond to the standard deviation.

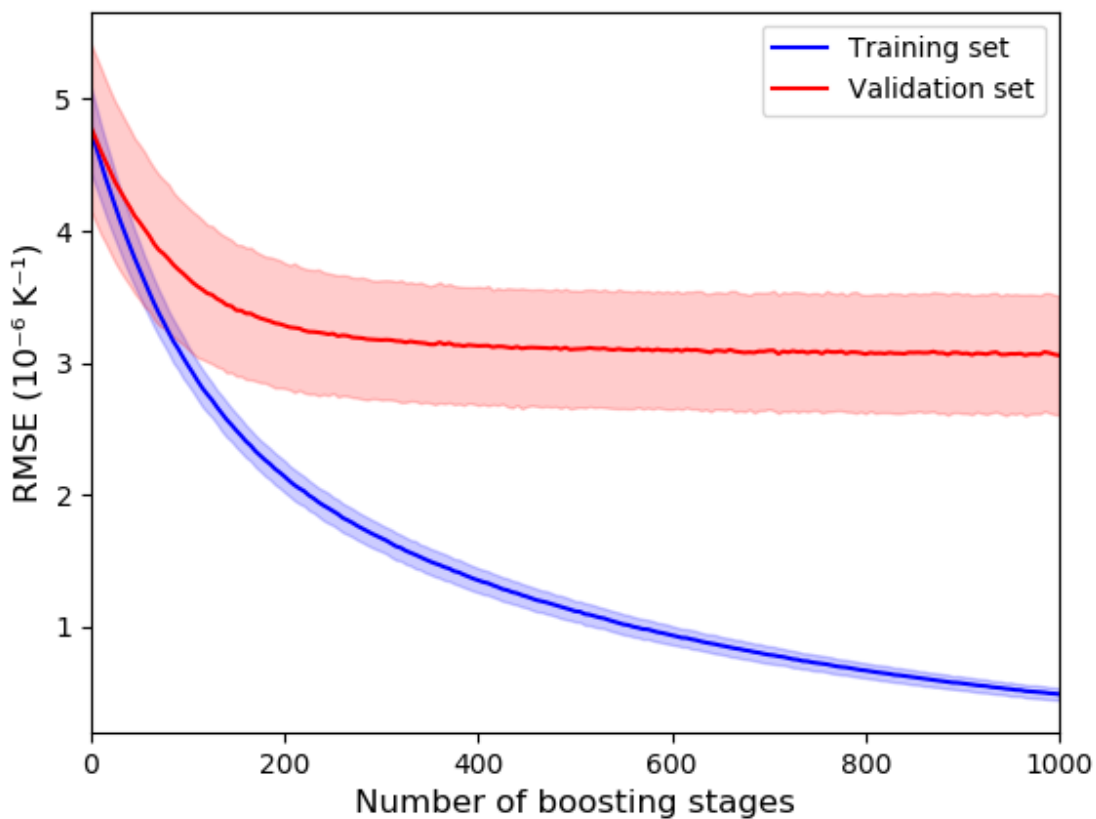


Figure S2: Learning curves of the thermal expansion using the non reduced SOAP features. Lines correspond to the mean RMSE values and shaded surfaces correspond to the standard deviation.

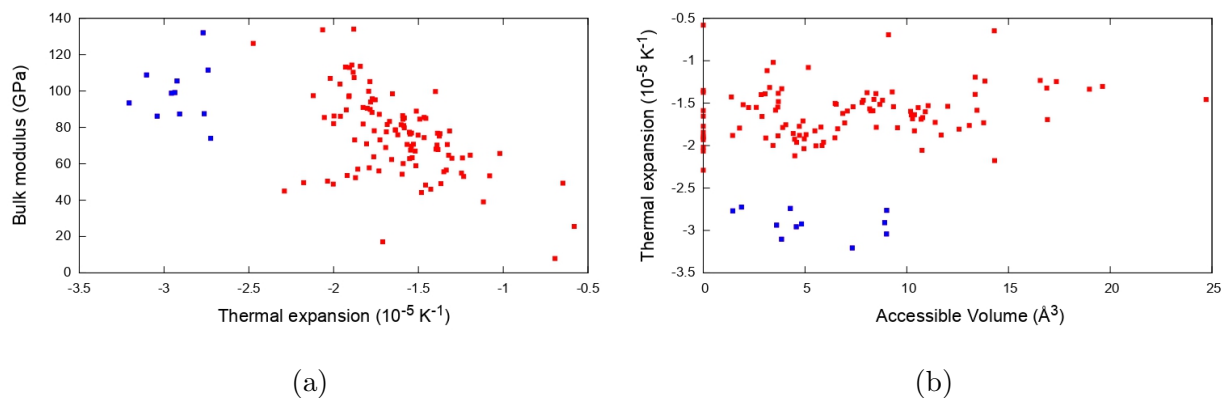


Figure S3: Comparisons of the thermal expansion values of zeolites with their bulk modulus (left panel) and accessible volume (right panel). Blue squares represent the frameworks with lower thermal expansion values.

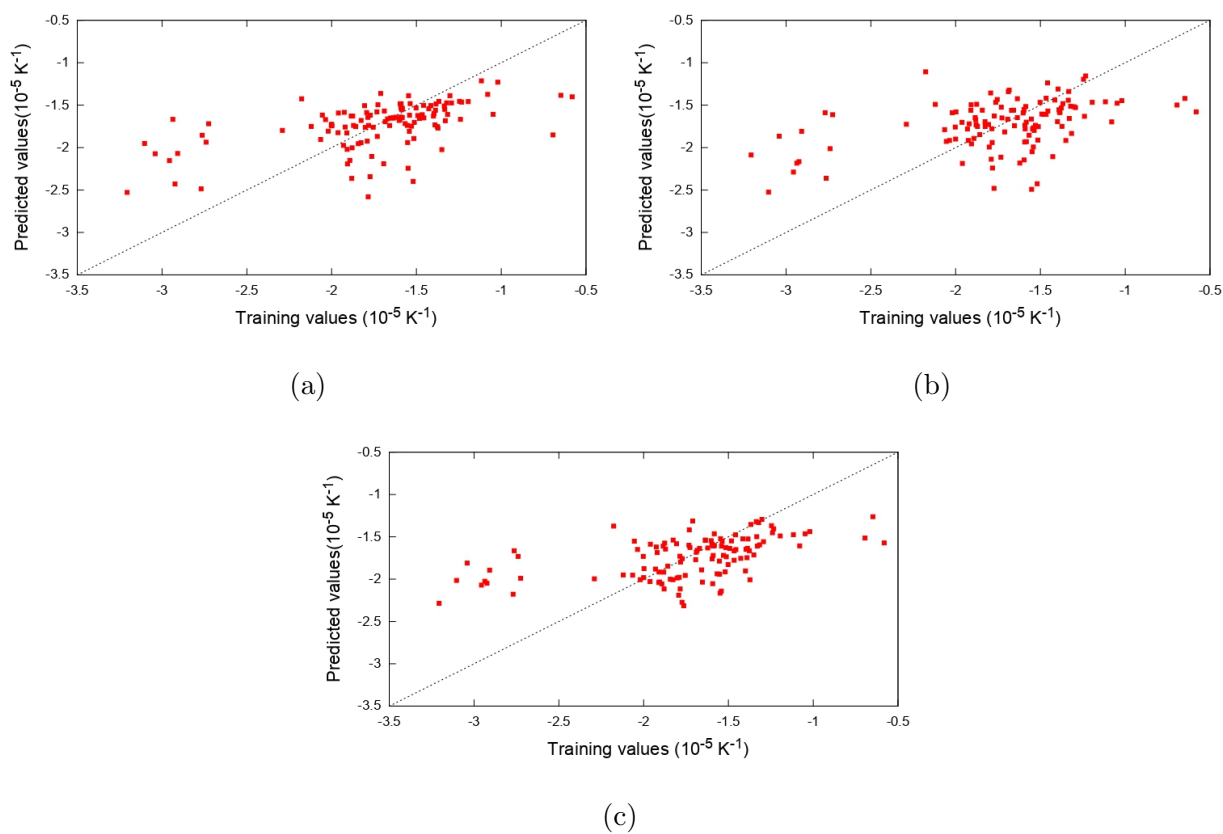
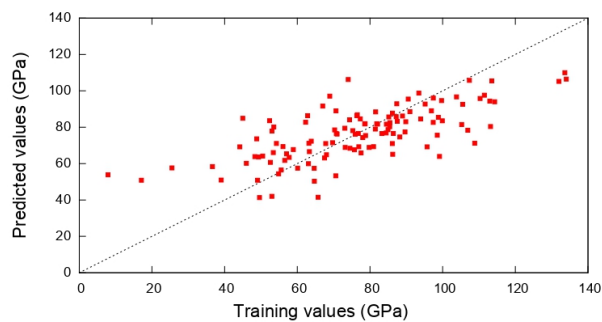
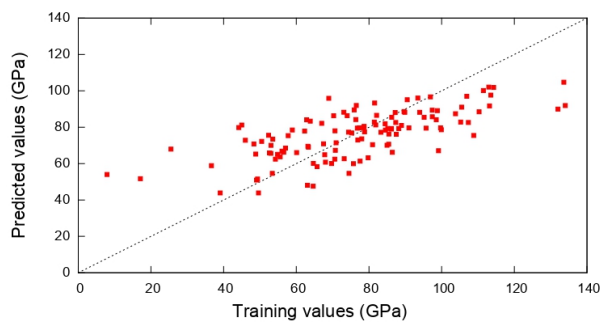


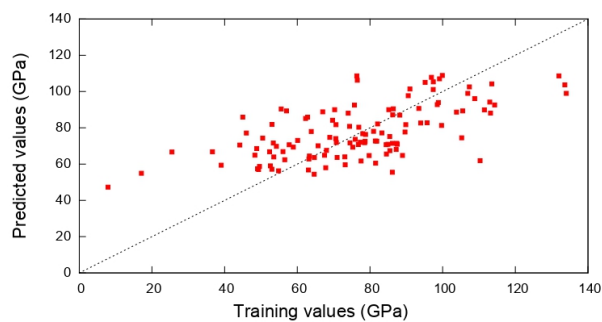
Figure S4: GBR cross-validated predictions on the thermal expansion using three different types of descriptors: (a) SOAP (RMSE = $3.75 \cdot 10^{-6} \text{ K}^{-1}$), (b) Volumetrical (RMSE = $4.16 \cdot 10^{-6} \text{ K}^{-1}$) and (c) Topological descriptors (RMSE = $4.13 \cdot 10^{-6} \text{ K}^{-1}$).



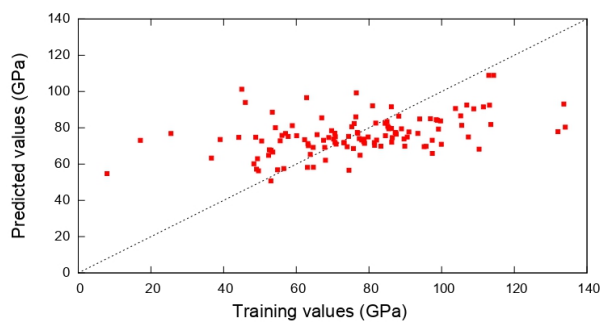
(a)



(b)



(c)



(d)

Figure S5: GBR cross-validated predictions on the bulk modulus K_0 using the four different types of descriptors: (a) Geometrical (RMSE = 16 GPa), (b) SOAP (RMSE = 15.9 GPa), (c) Topological descriptors (RMSE = 17.6 GPa) and (d) Volumetric (RMSE = 20.9 GPa).

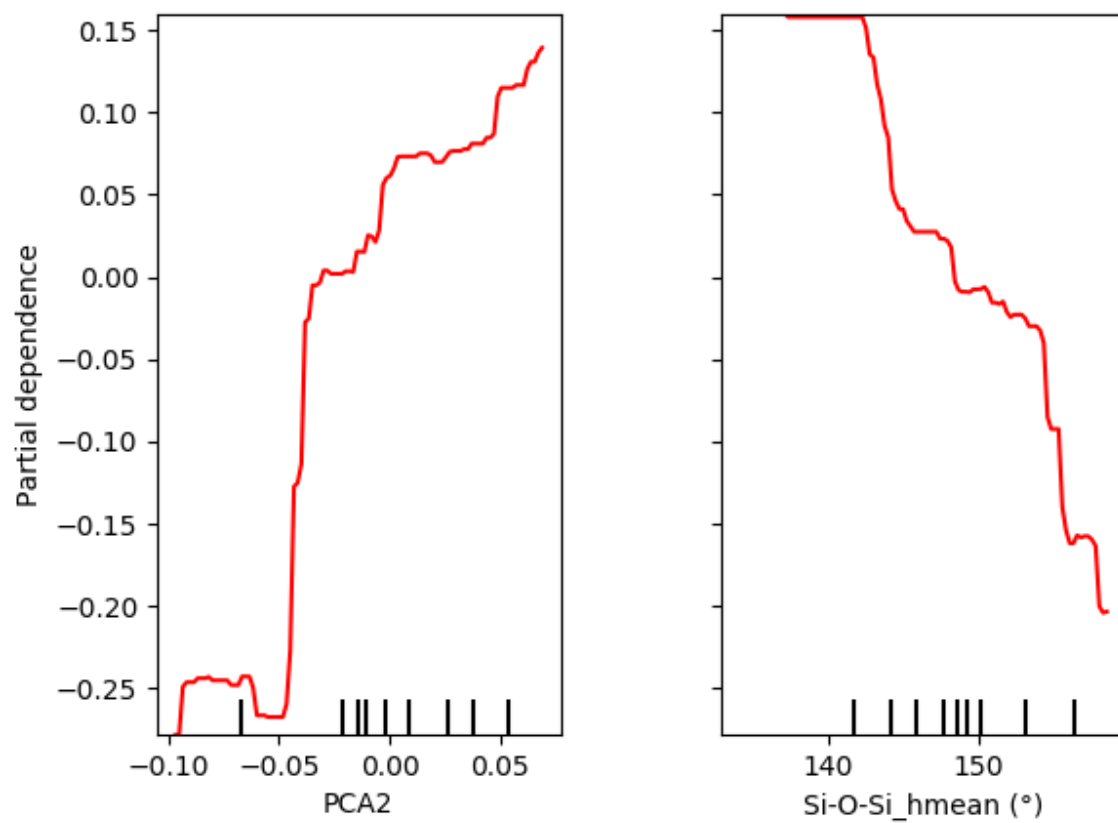


Figure S6: Partial dependence plots of the thermal expansion response with the second principal component of SOAP features and the harmonic mean of the Si-O-Si angles/

Table S1: Complete list of features used for each type of descriptors

Geometrical descriptor	Unit	Volumetric descriptor	Unit
Si-O mean	Å	Accessible volume (AV)	Å ³
Si-O harmonic mean	Å	Non accessible volume (NAV)	Å ³
Si-O geometrical mean	Å	AV + NAV	Å ³
Si-O minimum value	Å	Accessible surface area (ASA)	Å ²
Si-O maximum value	Å	Non accessible surface area (NASA)	Å ²
Si-O variance	Å	ASA + NASA	Å ²
Si-O-Si mean	°	Density	g.cm ⁻³
Si-O-Si harmonic mean	°	Volume	Å ³
Si-O-Si geometrical mean	°	Largest free sphere	Å
Si-O-Si minimum value	°	Largest included sphere	Å
Si-O-Si maximum value	°	Number of channels contributing to ASA	-
Si-O-Si variance	°	Number of channels contributing to AV	-
SOAP descriptors	Unit	Topological descriptors	Unit
Principal components of the SOAP analysis from the first to the 12th (PCAn, n=1, 2, ... ,12)	-	Number of neighbours in the successive spheres of coordination from the 2nd to the 13th (COORDn, n=2, 3, ..., 13)	-

Table S2: List of descriptors used when mixing different types of descriptors in the model. 3 descriptors of each type was used for a total of 12 descriptors.

Descriptor	Unit
Si-O-Si harmonic mean	°
Si-O mean	Å
Si-O-Si variance	°
Density	g.cm ⁻³
Volume	Å ³
Accessible + non accessible surface area	Å ²
Principal component 1	-
Principal component 2	-
Principal component 3	-
Number of neighbours in second sphere of coordination	-
Number of neighbours in third sphere of coordination	-
Number of neighbours in fourth sphere of coordination	-

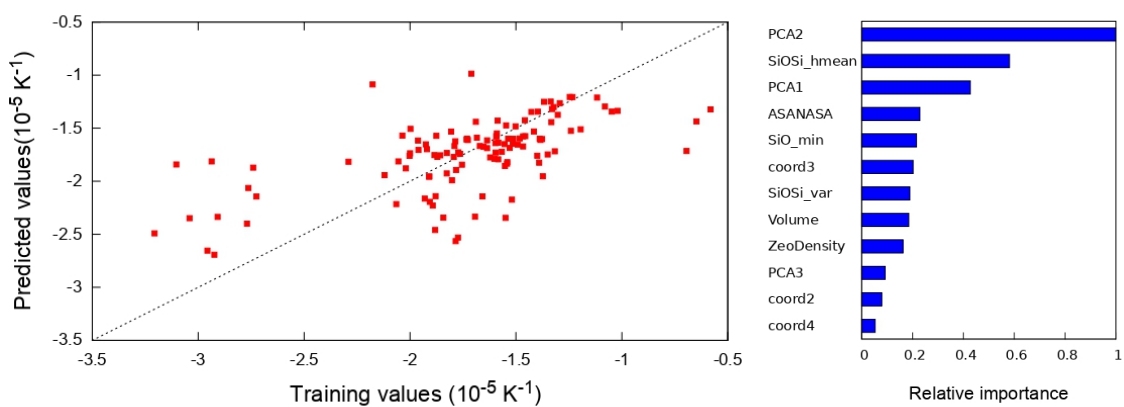


Figure S7: Left: GBR cross-validated predictions on the thermal expansion using a combination of the best 3 descriptors of each type and using non-optimised structures. Right: Feature importance histogram obtained with different types of descriptors and non-optimised structures