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

Abstract

The use of machine learning for the prediction of physical and chemical properties of crystals based on their structure alone is currently an area of intense research in computational materials science. In this work, we studied the possibility of using machine learning-trained algorithms in order to calculate the thermal properties of siliceous zeolite frameworks. We used as training data the thermal properties of 120 zeolites, calculated at the DFT level, in the quasi-harmonic approximation. We compared the statistical accuracy of trained models (based on the gradient boosting regression technique) using different types of descriptors, including *ad hoc* geometrical features, topology, pore space, and general geometric descriptors. While geometric descriptors were found to perform best, we also identified limitations on the accuracy of the predictions, especially for a small group of materials with very highly negative thermal expansion coefficients. We then studied the generalizability of the technique, demonstrating that the predictions were not sensitive to the refinement of framework structures at a high level of theory. Therefore, the models are suitable for the exploration and screening of large-scale databases of hypothetical frameworks, which we illustrate on the PCOD2 database of zeolites containing around 600,000 hypothetical structures.

Introduction

Zeolites are open three-dimensional framework structures composed of corner-sharing TO_4 tetrahedra, where T is a tetrahedrally coordinated framework atoms, most typically Al or Si. Zeolites are widely used in industrial applications as molecular sieves, for fluid adsorption and heterogeneous catalysis.[1] Their microscopic structure is a periodic three-dimensional network of corner-sharing TO_4 tetrahedra, where T is most commonly Al or Si. Out of an infinite number of such arrangements mathematically possible, 244 fully ordered zeolitic topologies have been currently determined experimentally, as approved by the Structure Commission of the International Zeolite Association.[2] The properties of these porous materials depend in part on the topology of their frameworks, and also on their chemical composition, i.e., the nature of the atoms that make up the framework (and the extra-framework ions, when they are present). Computational prediction of zeolite physical and chemical properties is therefore a research topic of high interest, and a large body of work has been published on the calculation of properties such as fluid adsorption[3] and transport,[4] gas mixture separation,[5,6] catalytic activity,[7–9] etc.

Recently, a new trend has emerged in the computational prediction of properties, it is the large-scale computational screening of materials databases to identify promising candidates for specific applications.[10,11] In particular, there is a body of research trying to identify materials featuring unusual (or counter-intuitive, or "abnormal") physical or chemical properties, also called metamaterials.[12] Framework materials, a category to which zeolites belong to, frequently display structural responses under external stimulation, that range from counter-intuitive to thermodynamically forbidden.[13] Many zeolitic materials, in particular, display features such as negative Poisson's ratio[14,15] (also called auxeticity[16]), negative linear compressibility,[17,18] pressure-induced softening,[19] and negative thermal expansion.[20,21] Negative thermal expansion, or NTE, appears to be systematic in pure-silica zeolitic frameworks,[21] and is of interest for potential use in ceramic, optical and electronic applications,[22] as well as in the creation of composite materials with zero thermal expansion for precision instruments

and microelectronics.[23,24]

In the past decades, we have seen the continued development of computational chemistry methods and the rapid increase in CPU speed and high-performance computing resources. The combination of the two has made possible the systematic prediction of materials physical and chemical properties from crystalline structure, even to the point of high-throughput studies that can deal with hundreds or thousands of structures. However, such large-scale studies involve massive costs in terms of CPU time (and associated carbon emissions). In particular, highly accurate calculation methodologies at the quantum level remain expensive, and not all properties can be studied in a high-throughput setting.

One of the possible avenues to accelerate the pace of materials discovery is therefore to leverage the data available, either experimentally or from state-of-the-art computational approaches, through machine learning.[25–27] Among the many tasks that machine learning (ML) is being applied to, one of the most popular in chemical and materials sciences is the creation of properties prediction algorithms based on structure and/or chemical composition,[28] and the identification of materials with specific properties.[29] It has been applied to a wide diversity of problems, including prediction of potential energy surfaces and stability,[30] electronic properties,[31] magnetic properties,[32] mechanical properties[33,34] and stability,[35] catalytic activity,[36] and the characterization of energetic materials,[37] to name only a few examples. We refer the reader to Refs. 28 and 25 for more comprehensive reviews of the principles and applications, respectively.

In the field of zeolites, where a large number of hypothetical frameworks are known,[38] the applications of machine learning for the identification and characterization of frameworks has long been recognized.[39,40] More recently, Helfrecht et al. have analysed the structural diversity of hypothetical zeolites databases, showing the power of smooth overlap of atomic position (SOAP) descriptors in order to classify local environments and identify zeolite building blocks.[41,42] ML algorithms have also been applied to the design of zeolite templates (or organic structure directing agents)[43] and more broadly in synthesis,[44] as well as the predic-

tion of catalytic activity.[7,45] In the past few years, our group and others have demonstrated the capability of ML to predict various mechanical properties of zeolitic frameworks. First, we focused on average volumetric properties, such as bulk and shear moduli.[34,46] Later, we have shown how the use of ML techniques, integrated within a multi-scale modeling strategy including quantum chemical calculations, can speed up the discovery of zeolites with complete auxeticity,[47] a subclass of mechanical metamaterials with very rare behavior.[48] These predictions have yet to be directly verified experimentally, due to the difficulty in systematic characterization of crystal elastic constants (or mechanical behavior under anisotropic stress), but in cases where data is available, the mechanical properties are usually found to be in good agreement with DFT values.[21,49]

In this work, we used a previously calculated database of thermal properties of pure silica zeolites and investigated the feasibility of its use for machine learning purposes. In the next sections, we first introduce the computational methods, including a brief presentation of the databases we used, as well as the algorithm we chose and a definition of all the materials descriptors. We then assess the accuracy of different models trained on our DFT-calculated database, and compare the results using different types of materials descriptors. Finally, we discuss the use of non DFT-optimised structures in the training of the model, and confront our database of zeolites with a much larger one: the PCOD2 database of 600,000 hypothetical zeolite structures.

Computational methods

DFT calculations

The present work is based on zeolite properties data calculated at the quantum chemical level, previously obtained in the group. We used the data on thermal and mechanical properties of pure SiO_2 zeolites from these DFT calculations as a basis for the ML study reported herein. We refer the reader to Ref. 21 for a full methodological description, but provide here a short

summary for the sake of convenience.

The DFT calculations were done on 134 structures for which we optimized the geometry. Those calculations were performed starting from the IZA models for 190 zeolites with fewer than 150 atoms in their unit cell; from those 190, 134 structures achieved convergence within reasonable time constraints. For these structures, we calculated the thermal properties using the quasi-harmonic approximation. Of those, 120 calculations converged with a sufficient range of volume and temperature. Compared to the computation of third and higher order terms of the energy, this technique allows for the determination of certain thermal properties while keeping an affordable computational cost. The principle behind it is to obtain the relationship between the volume V and the frequencies of phonons ω_k — this dependence is missing in the harmonic approximation — by making several harmonic frequency calculations at different volumes. Once this relationship is known, one can determine the equilibrium volume at each temperature by minimizing the Helmholtz free energy $F(V,T)$ with respect to the volume following the equation:

$$F(V,T) = U_0(V) + U_{\text{vib}}(V,T) - TS \tag{1}$$

where U_0 is the zero-temperature lattice energy. U_{vib}, the vibrational part of the energy, can be written as:

$$U_{\text{vib}}(V,T) = E_0(V) + k_B T \sum_k \ln\left(1 - e^{\frac{\hbar\omega_k(V)}{k_B T}}\right) \tag{2}$$

where $E_0(V)$ is the zero-point energy of the system, k_B is the Boltzmann constant, $\hbar$ is the reduced Planck constant and ω_k is the volume-dependent vibration frequency.

Through this method we were able to obtain several properties such as the thermal expansion and the bulk modulus, as well as the later's dependence on pressure and temperature. We compared the already synthesized SiO_2 zeolites with the theoretical ones and, while it was not straightforward to define synthesis conditions, we observed that the bulk modulus K_0 and its derivative K_0' seems of great importance in determining the synthetic feasibility

for a given structure. Indeed the high rigidity of a framework makes it more stable and we noticed that a overly large negative K_0' apparently leads to the structure not being synthesizable, as no experimentally known framework was found to exhibit this behaviour. The negative K_0' being associated with a mechanism of amorphization under pressure provides an additional rationale for this finding.

Deem database

To compare our results with predictions made on a much larger number of zeolitic structures, we used the PCOD2 database of hypothetical zeolites created by Deem and co-workers.[38,50] The development of this database started in 1992[51] with 2,000 structures and continued its growth reaching around 600,000 structures in 2006,[52] which is the number of structures we studied here. The database, no longer accessible on its original academic website, was mirrored from our own local archives and made available at the following stable URL: `https://doi.org/10.5281/zenodo.4030232`

These structures were obtained from combined Monte-Carlo simulations, simulated annealing and refinement using a classical force field, the Sanders-Leslie-Catlow (SLC) interatomic pair potential. 10% of the 2.7 millions unique structures obtained in their work were found to have an energy per SiO_2 unit that is within 30 kJ/mol from the well-known α-quartz dense phase. Many of these structures were expected to be stable and achievable through synthesis, and the authors emphasized that their database could drive the discovery and synthesis of novel materials as well as support identification of materials through powder pattern searching and matching. In addition, they also calculated the stiffness tensor of second-order elastic constants — with the same force field approach. As observed before however,[34] the results obtained on mechanical properties were not very accurate, as K values reported range from $-27{,}000$ (unstable) to 20,500 GPa (unphysically high). Nevertheless, this structural database of hypothetical zeolites has been widely used in the exploration and systematic prediction of structure/property relationships.[42,53,54]

Machine learning

In order to apply a machine learning strategy to the task at hand, we chose a gradient boosting regression (GBR)[55,56] for the algorithm, as it has been proven to be quite robust and efficient for small datasets, like in our case.[57] This algorithm is a stage-wise additive model which trains decision trees that are built in a greedy fashion to minimize the loss function, which was chosen to be a least squares function in our case. We have used this methodology in the past for the prediction of physical properties in dense and porous frameworks, including mechanical stiffness,[34] and anisotropic elastic properties such as negative Poisson's ratio.[47]

Table 1: Hyperparameters for the gradient boosting regression

Parameter	**Value**
Number of boosting stages	250 [a] / 500 [b]
Learning rate	0.01
Minimum samples split	2
Maximum depth	2
Minimum samples leaf	2
Subsample	0.4
max features	square root of total features
loss function	least squares

[a] Used for the geometrical descriptors only to avoid over-fitting observed in the learning curves.
[b] Used for all other descriptors.

We used the GBR implementation from the `scikit-learn` Python package.[58] We used a 3-fold cross-validation procedure as implemented in Sci-kit learn package which we repeated 50 times in order to obtain relevant accuracy scores and errors by averaging them over all the simulations. To choose and validate the hyperparameters, we used cross-validation and chose as a measure of accuracy the root mean square error (RMSE). In particular, we focused on the impact of the number of boosting stages, which — although GBR is generally said to be fairly robust to overfitting — we found to be an important hyperparameter. The learning curves for this hyperparameter, for the prediction of thermal expansion coefficient based different sets of descriptors, are displayed in Figure S1. We can clearly see that even though we took precautions to avoid over-fitting when defining the other hyperparameters,

this phenomenon is happening for a large number of boosting stages. For the set of *ad hoc* geometrical descriptors, we therefore chose to use 250 boosting stages. For other descriptors, the phenomenon is not as pronounced, and we decided to keep a higher number of decision trees (namely, 500). We report the final set of hyperparameters used for this study on Table 1.

Choice of descriptors

Besides the choice of the algorithm and hyperparameters, one other critical aspect of the machine learning methodology for prediction of materials properties is the choice of descriptors. Descriptors should reflect the nature of the input data properly (in this case, the atomic structure of materials) and allow for the differentiation of the different materials. Moreover, it is also important that the descriptors share some kind of link with the targeted labels. Within these constraints, and based on the existing literature in the field,[28] we identified several sets of possible materials descriptors for use in our model. We then proceeded to compare the performance (or statistical accuracy) of the models based on different descriptors. For each set, we chose a total of 12 descriptors which, taking into account the size of our training data set and our choice of hyperparameters, allowed a reasonable description of our systems while avoiding over-fitting in our simulations. Here we give a brief explanation of each type of descriptor and how we obtained them. The complete list of descriptors used for each type is available in supporting information in Table S1, and the discussion of the relative merits of the different sets is detailed in the text.

Ad hoc geometrical descriptors

Zeolitic frameworks are a three-dimensional assembly of corner-sharing SiO_4 tetrahedra. Therefore, it is natural to characterize the local geometry of each atomic environment by the set of simple parameters such as Si–O distances and Si–O–Si angles. It was further demonstrated in a large number of works, including some of the earlier studies on structure–property relationships in zeolitic frameworks,[59,60] that these two parameters are of great

importance for understanding the physical and chemical properties of this family of materials.

In prior work, we have shown that such *ad hoc* geometrical descriptors — designed from the chemical intuition and our knowledge about the systems at hand — can be used in supervised machine learning for the prediction of mechanical properties.[34,47] For this reason, we included them in the present study: from each optimized zeolite structure, we calculated the distribution of bond distances and angles using the pymatgen python package.[61] We then used as descriptors some statistical features of these distributions: different means, variance, extremal values, etc.

Topological descriptors

One of the conclusion of our previous systematic study[21] was that the topology plays a key role on thermal properties of zeolitic frameworks. Indeed, we reported a large span of the values for the thermal expansion and the bulk modulus of these frameworks, although the composition for all zeolites is identical. This led us to think that topological features could form interesting descriptors to include in our current study. Although several representations of framework topology are possible, we chose here to use the coordination sequence, i.e., the number of neighbours in each successive coordination spheres of the Si atoms.

Because zeolites are four-connected nets, the first term of the sequence is always 4; we therefore decided to take into account from the second sphere to the 13^{th} sphere — averaging over all starting Si atoms. Information on topology such as the number of neighbours were obtained from the optimized zeolite structures using the CrystalNets julia package.[62] This package designed for the identification and manipulation of crystal nets representation and topology has been developed in our group and made available on Github at `https://github.com/coudertlab/CrystalNets.jl`

Volumetric descriptors

Because zeolites are nanoporous materials, another possible choice of structural descriptors would be focused on the characterization of their pore volume, and more generally, nonlocal or volumetric information. Indeed, as described earlier, the quasi-harmonic approximation we used to determine the thermal properties of zeolites is considering the harmonic expression of the Helmholtz free energy to which we added the vibrational part of the energy (which depends on the volume). Thus volumetric descriptors such as the density, the accessible volume or the surface area could also play an important role in the prediction through machine learning. These quantities were obtained from the optimized zeolite structures using the Zeo++ software package.[63,64] The surface area, accessible volume and volume being dependent of the choice of unit cell, we normalized them per number of SiO_2 units. For the sake of consistency, we chose used as descriptor the calculated (DFT-optimized) crystallographic density, instead of the topological density reported by the IZA. This choice is not crucial to the conclusions of this work, however, as the reported and calculated density values are almost identical.

Smooth Overlap of Atomic Positions (SOAP)

The SOAP is an encoding method introduced by Bartók and co-workers,[41,65] which is a descriptor of local geometry, describing the environment around a given point (usually an atomic position). By projecting the local geometry onto orthornormal basis functions based on spherical harmonics, it is invariant by rotation and permutation of atoms. It has been used in particular to determine the similarity of two neighbourhood environments, and to identify features that differentiate molecular structures from one another,[66,67] including in the specific case of zeolites.[42] It has also been used to encode atomic environments for machine learning inter-atomic potentials, due to its powerful and rich material representation.[68] Finally, it is also used as descriptor for regression tasks, namely the prediction of physical or chemical properties.[69,70]

One important parameter of this method is the cut-off, representing the distance until which all the environment is included in the description. Initial tests showed that a suitable value for zeolitic systems is around 6 Å, as this distance includes the nearest and next-nearest neighbours, including then characteristics of both Si–O distances and Si–O–Si angles. This was verified by computing the SOAP descriptors for different values of cut-off and running machine learning predictions based on these different sets. For values of cut-off below 6 Å, we observed a decrease of the RMSE with increasing cut-off, while the accuracy remained constant for values higher than 6 Å. Finally, in order to reduce the large dimensionality of the SOAP descriptors and bring them to a comparable set to other descriptors, we used the Principal Component Analysis (PCA) technique, which determines through an algorithm the most important components and projects the data on them — reducing the dimensionality while keeping as much variation as possible. We chose here the first 12 components from the PCA analysis.

Results and discussion

ML model based on geometric descriptors

In this section, we will discuss the general performance of ML models based on simple geometric descriptors, as used in past studies for mechanical properties,[34,47] for the prediction of thermal properties of frameworks. To assess the statistical accuracy of our models, we used a cross-validation strategy on our data set. Hyperparameters (in particular the number of boosting stages) were tested systematically in order to avoid overfitting of the models, as detailed in the Methods sections. We display on Figure 1 the results of a GBR model for thermal expansion based on geometrical descriptors. We can see that values of thermal expansion are overall well predicted in the center range as our data are concentrated between around -2 and -1 K^{-1}. However it can be seen that large deviations are present for several points outside of this range. It seems that there are outlier materials (especially for very

negative values of thermal expansion), and that there are too few values on the extreme sides of the data set to allow for the model to train efficiently on them, which is why we are witnessing such high errors on these points.

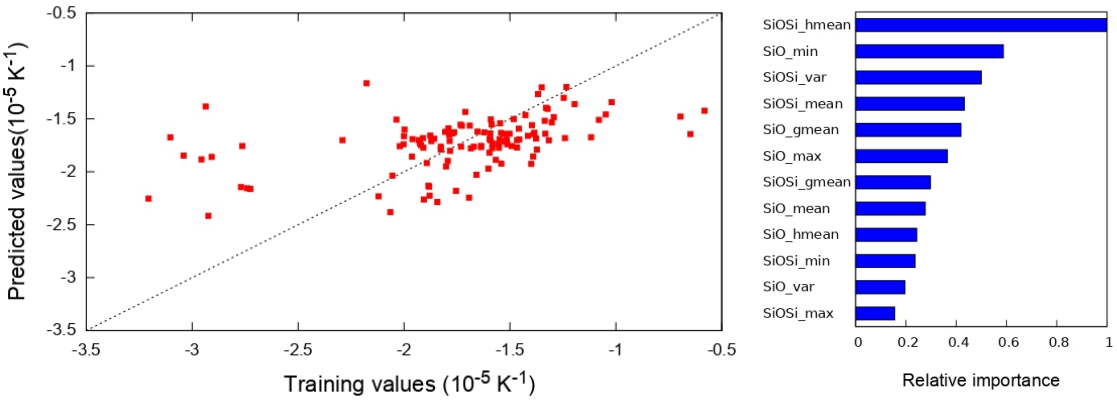


Figure 1: Left: GBR cross-validation prediction of the thermal expansion α, based on geometrical descriptors. Right: relative importance of each descriptor.

To quantify this dispersion we used the RMSE (root-mean-square error), for which we found an averaged value of 4.24 10^{-6} K^{-1} corresponding to an error of 20%, which is reasonable considering the small data set used in this study — and in line with accuracy of ML models based for other macroscopic physical properties based on local geometric descriptors.[48] We also confirmed that, removing the points with low values of thermal expansion (outliers for $\alpha < -2.5$ 10^{-5} K^{-1}), the RMSE lowers to 2.56 10^{-6} K^{-1}, giving much better predictive power. To try to find the microscopic origin of the specific behavior of these materials, we investigated their frameworks. As shown in supporting information Figure S3, while their thermal expansion deviates away from the average value, no other property or feature showed any systematic difference compared to the rest of the frameworks. Visual inspection of their structures did not reveal any particularity either, and we therefore believe that the specific behavior observed may be unphysical, and find its root in high order terms, which are not included under the quasi-harmonic approximation and thus not taken into account in our DFT calculations.

It is also important to note that the deviation observed in the prediction of thermal

expansion also finds its origin (in part) in noise in the training data itself, that comes from the uncertainty of the DFT calculations. The properties obtained by DFT calculations were obtained through a rather long and difficult process which comes with a certain degree of uncertainty. In order to perform a systematic study of thermal properties (to build the database), we had to fix a certain set of parameters for all the frameworks (ranges of temperature, volume expansion, number of points in numerical derivatives, etc): more fine-tuned parameters for each zeolite could have resulted in a better accuracy for the calculated properties, but was not feasible in a systematic approach. We note that a deviation of the same magnitude is also observed with other types of descriptors, for which representations can be found in Figure S4. We note here that, in this study, our main interest is in evaluating the feasibility of the ML models and the comparison of descriptors, for a physical property (thermal expansion) that has never been studied in framework materials at that scale before.

Finally, we estimated the relative importance of each descriptor in the training process and report the results on the right panel of Figure 1. It can be seen that the first descriptor in terms of importance is related to the Si–O–Si angles (harmonic mean of angle values). Going down the list, the angles appear to be of higher importance for the prediction of thermal expansion than the Si–O distances. This confirms the physical intuition, because thermal expansion is dominated by low-frequency vibration modes, which typically involve tetrahedral rotations of SiO_4 units and Si–O–Si angle bending.

This is further confirmed with the partial dependence plots depicted in Figure 2. With this representation, we can see how a chosen property (here thermal expansion) responds as a function of some specific descriptors (left: the Si–O–Si angle harmonic mean; right: minimal Si–O distance). It can be clearly seen that the Si—O–Si angle exhibits a nearly linear dependence with the predicted thermal expansion, starting from around 142°, whereas the impact of the Si–O distance is smaller in amplitude, and without a clear monotonic trend. Knowing that the minimal Si–O distance was found to be the second most important feature when training the model, it highlights the fact that angles have much more importance than

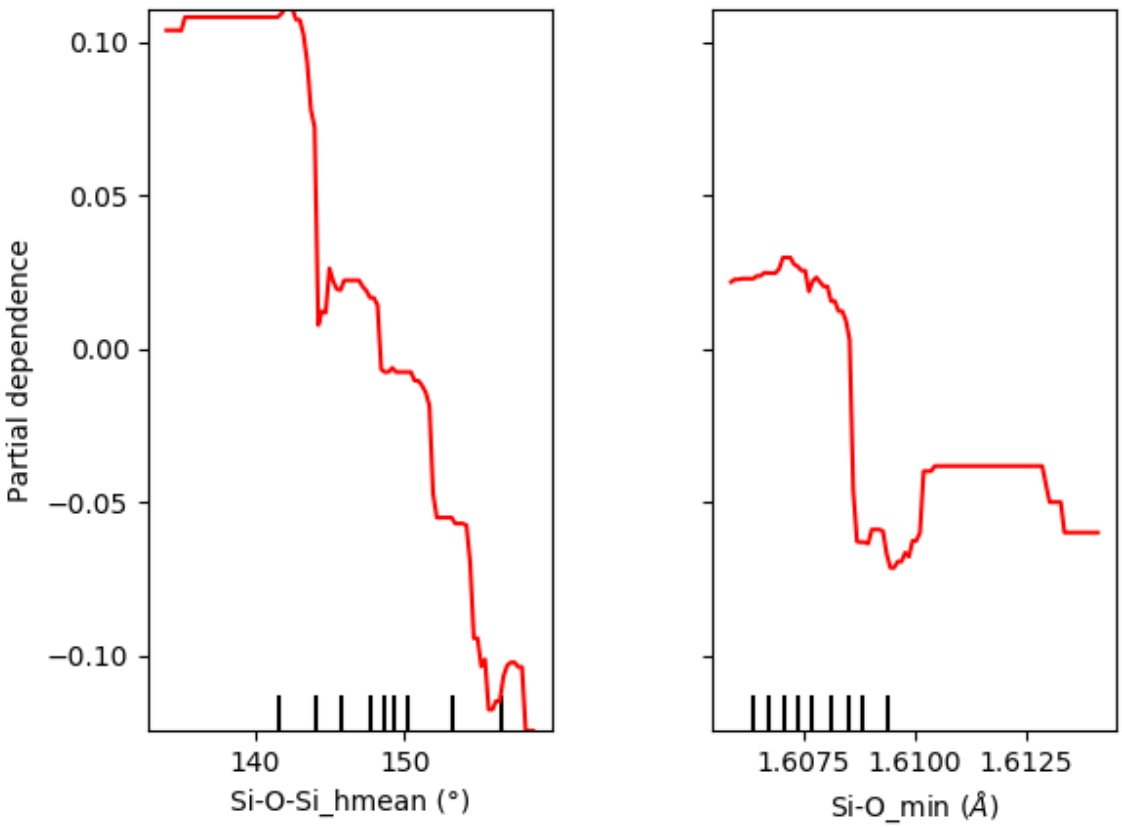


Figure 2: Partial dependence plots of the (left panel) SiOSi_hmean and (right panel) SiO_min. Right panel shows the comparison between the these two descriptors.

distances in the prediction of the thermal expansion.

Comparison of different descriptors

In this section, we want to compare the performance of different sets of descriptors for the description of thermal properties. Among all four types of descriptors tested (see Figure S4), the use of principle components of SOAP descriptors seems to provide the best-performing description, with a RMSE value of 3.75 10^{-6} K^{-1}. This is smaller than the *ad hoc* geometric descriptors (angles and distances) as shown above (4.24 10^{-6} K^{-1}): we comment this by noting that the SOAP method gives a more comprehensive description of the local geometry of the structure, including effects more complex than just the angles and distances. On the

other hand, we see that topological and volumetric descriptors are on par with geometric features, showing RMSE values of 4.13 10^{-6} K^{-1} and 4.16.10^{-6} K^{-1}, respectively.

This conclusion is strengthened when we look at the prediction of other physical properties. As examples, we present in supporting information the predictions of the bulk modulus (using the same methodology). There again, SOAP descriptors allow for the most accurate description, with a RMSE value of 15.9 GPa, lower than for the topological and volumetric descriptors (RMSE of 17.6 GPa and 20.9 GPa respectively). It can also be observed on the representations (Figure S5) that volumetric descriptors result in a really poor prediction of bulk modulus, as the cloud of points strays away from the perfect prediction represented by the dashed line and tends to form an horizontal line. This conclusion is counter-intuitive, as we could have expected the density or porosity-related metrics to be directly linked to the stiffness of the materials. However, we show here that using solely these metrics is not sufficient to efficiently train a model. On the other hand, we note that *ad hoc* geometrical descriptors to an accuracy close to that obtained with the SOAP descriptors (RMSE of 16.0 GPa). This highlights once more the importance of Si–O–Si angles in determining the properties of zeolites, as we observe a similar accuracy compared to SOAP descriptors which contain much more information on the structure than just the angles.

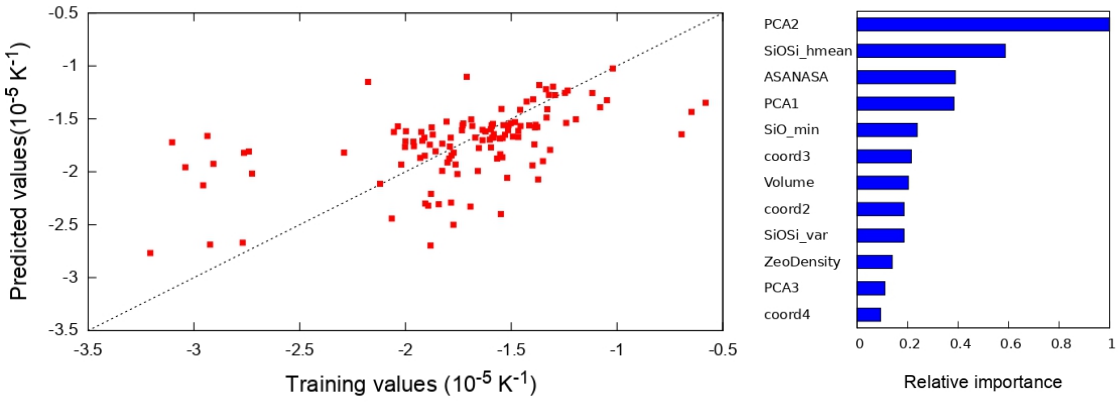


Figure 3: Left: GBR cross-validated predictions on the thermal expansion using a combination of the best 3 descriptors of each type. Right: Feature importance histogram obtained when training the model with different type of descriptors.

Having looked at all the types of descriptor separately, we decided then to train a model combining this time different types in order to compare them in the same model. We chose to use the three best descriptors of each type in determining the thermal expansion which gave us a new set of 12 descriptors (Table S2). We kept the same hyperparameters, after confirming through learning curves that these allowed us to avoid over-fitting. We reported the cross-validated prediction of the thermal expansion using this new set of descriptors in Figure 3, along with the relative features importance. What we observe first is that, while the accuracy is slightly better than before with a RMSE of $3.64.10^{-6}$ K^{-1}, the improvement is marginal, and mainly located for the low values of thermal expansion — many of which were outliers in the previous models.

Looking at the features importance, we see that the second principal component of SOAP descriptors is the most important in determining the thermal expansion — as was already the case with only SOAP features. This demonstrates that, even if we reduced the dimensionality of these features with a principal component analysis which causes a loss of some of the information, this method still represents a highly competitive way of describing a structure. Figures depicting the partial dependence of the second principal component of SOAP features as well as the Si–O–Si harmonic mean can be found in supporting information (Figure S6). These representations show that the dependence of the thermal expansion on those two descriptors is quite similar. Both of them exhibit a strong linear dependence, unlike with all the other descriptors used here, for which the dependence is either weak or almost null. As already observed previously, Si–O–Si angles represent one of the major features to incorporate in the description of the structure of zeolites. Here we demonstrated that this descriptor is essential for the prediction of the thermal expansion, and we believe that it would be useful to include it for the determination of other properties of zeolitic materials.

Applying ML models to non-DFT optimised structures

We have proven above that the prediction of thermal properties is possible, with reasonable accuracy, from the structure of a zeolitic framework optimized at the DFT level. While interesting in itself, the need for a DFT-optimized structure can strongly limit the applicability of the ML model in a high-throughput screening scenario. Indeed, while the structure optimization is computationally less intensive than the quasi-harmonic calculation of thermal properties (and can be performed in hundreds of structures[47]), it is not scalable to the size of available hypothetical databases of zeolitic materials (which can contain hundreds of thousands of structures).

Therefore we found interesting to investigate the performance of a ML model trained on non DFT-optimised structures to predict the same thermal and mechanical properties (still using for those properties the data obtained by DFT). This is, in effect, a test of the sensitivity of the ML model to the accuracy of the geometries used as input. For this, we created a data set containing the same 120 structures as the DFT-optimised data set but using the zeolites structures from the IZA database (obtained by a distance least-squares refinement technique) and used them directly as a training set without optimising them. We used the same hyperparameters as well as the same combination of descriptors as the above section, as we proved that the results were slightly better than with just only one type of descriptor. All the features required were retrieved on the non-optimised structures following the same procedure described in the computational methods. We reported the cross-validated predictions along with the feature importance plot in supporting information.

We obtain for this new ML model a RMSE of $3.63.10^{-6}$ K^{-1} for thermal expansion, close to the previous values observed. The same behavior is also observed on the representation, with the same group of isolated outliers corresponding to the lowest values of thermal expansion. This shows that non-optimised structures do not differ too much from the optimised ones, and that these differences in geometry do not strongly impact the performance of ML models. Furthermore, we can see on the feature importance plot (Figure S7)

that the importance order is quite similar. Indeed the second principal component is the most important followed by the harmonic mean of the Si–O–Si angles, just like the previous model with optimised structures. This is due to the similarity of the features between non-optimised and optimised structures. Therefore, this validates the prediction of physical properties from structures optimised at a level lower than DFT (for example, force field optimised structures), making it possible to investigate very large-scale zeolitic data sets, such as the PCOD2 database created by Deem et al..[38,50]

Deem database

We now apply our ML model to the PCOD2 database,[38,50] which contains around 600,000 hypothetical pure silica zeolite structures, obtained from combining Monte-Carlo simulations, simulated annealing, and structure refinement using a classical force field (the Sanders–Leslie–Catlow (SLC) interatomic pair potential[71]). It constitutes a great tool for the machine learning-based exploration of new synthesizable structures, or the identification of candidate zeolites with targeted properties. For example, we have used it in the past to identify new frameworks with auxetic behavior.[47] In this work, our interest lies in comparing our data set of calculated zeolites with the full database of hypothetical structures and ultimately, trying to predict the distribution of thermal expansion and bulk modulus of the PCOD2 database.

In order to do this, we first calculated different descriptors for the whole PCOD2 database. For around 600,000 structures, the computational effort is as follows (timing reported for nonparallel, single-CPU calculations): a couple of days for the bond distances and angles, one week for the SOAP features and around two weeks for the topological descriptors. Due to the size of some of the systems within this database, volumetric descriptors could not be retrieved as the time needed to calculate them was excessively long. We reported histograms of values of Si–O distances and Si–O–Si angles on Figure 4 for both our training data set, and the entire PCOD2 database — with occurrences normalised for the sake of clarity. First, we can see that the span of values for both distances and angles is close between IZA structures

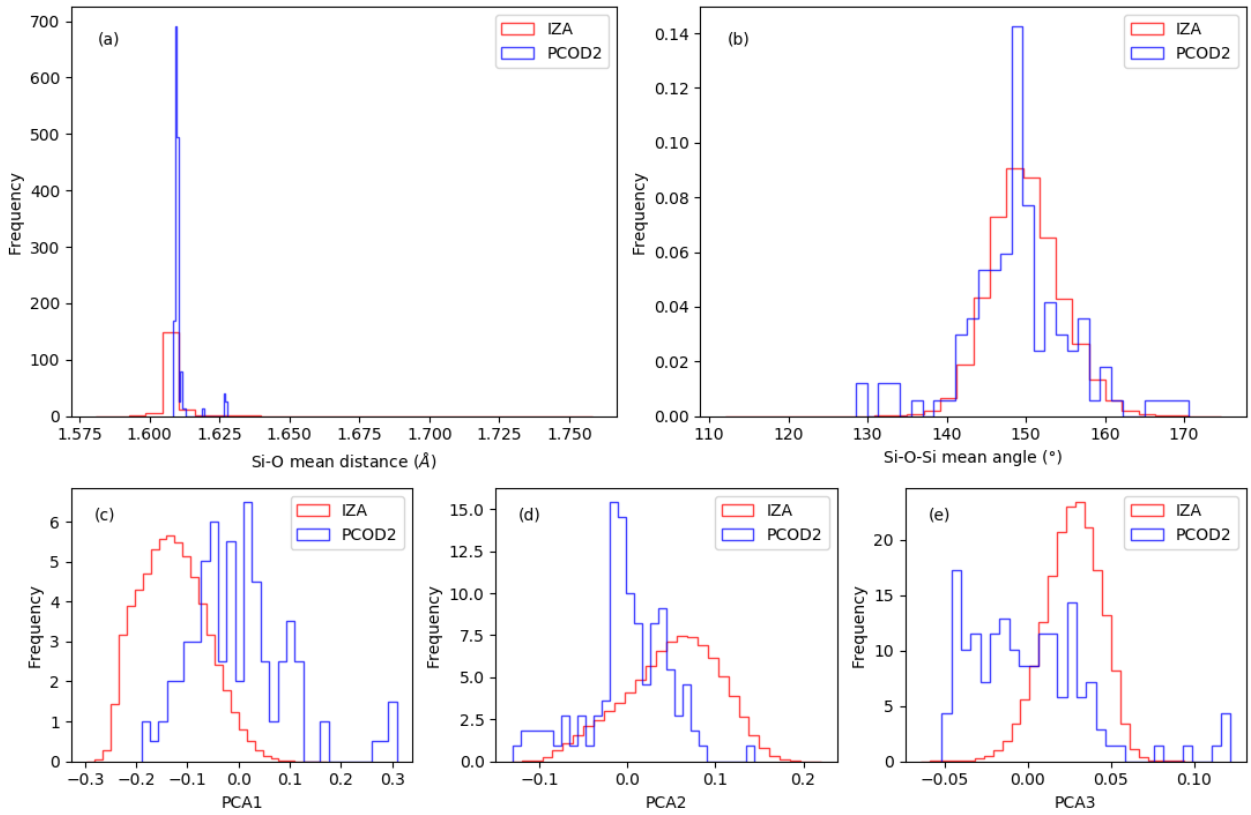


Figure 4: Distribution of descriptor values of our data set compared to the Deem database. Descriptors represented are: (a) Si–O mean distance; (b) Si–O–Si mean angle; (c), (d) and (e) first three components of principal component analysis of SOAP features. Due to the difference in data set size, frequencies have been normalized for both data set.

and the PCOD2 database. This is strongly encouraging for a prediction model because the geometrical descriptors show approximately the same distribution, meaning that a model trained on these descriptors should be generalizable without extrapolation. Si–O distances do not vary much, with values being concentrated between 1.60 and 1.62 Å — in line with our observations that Si–O distances do not have a great importance when training ML models. In comparison, Si–O–Si angles exhibit a larger range of values, with a total span of up to 20°. This explains the importance of these angles in the predictions, as they account for a lot of the diversity between the different structures.

All these observations are in contrast with the case of the principal components of SOAP features. Indeed one can see that for the first three principal components there is a shift in the distribution of IZA structures compared to the PCOD2 database. This would mean

that the frameworks we are currently working on do not represent all of the possible environments present in the database of theoretical structures: the geometrical diversity of the PCOD2 database is higher. This could also mean that some of the structures in the PCOD2 database are outside of the realm of "feasible zeolite structures", as defined by the convex hull of features from the experimentally known frameworks (in the IZA database). This is an interesting new take on the question of experimental feasibility of frameworks,[42,49] which we intend to pursue further, but is outside of the scope of our current work. We can only conclude that a prediction on the PCOD2 database using the principal components of SOAP features as descriptors would probably result in a poor prediction.

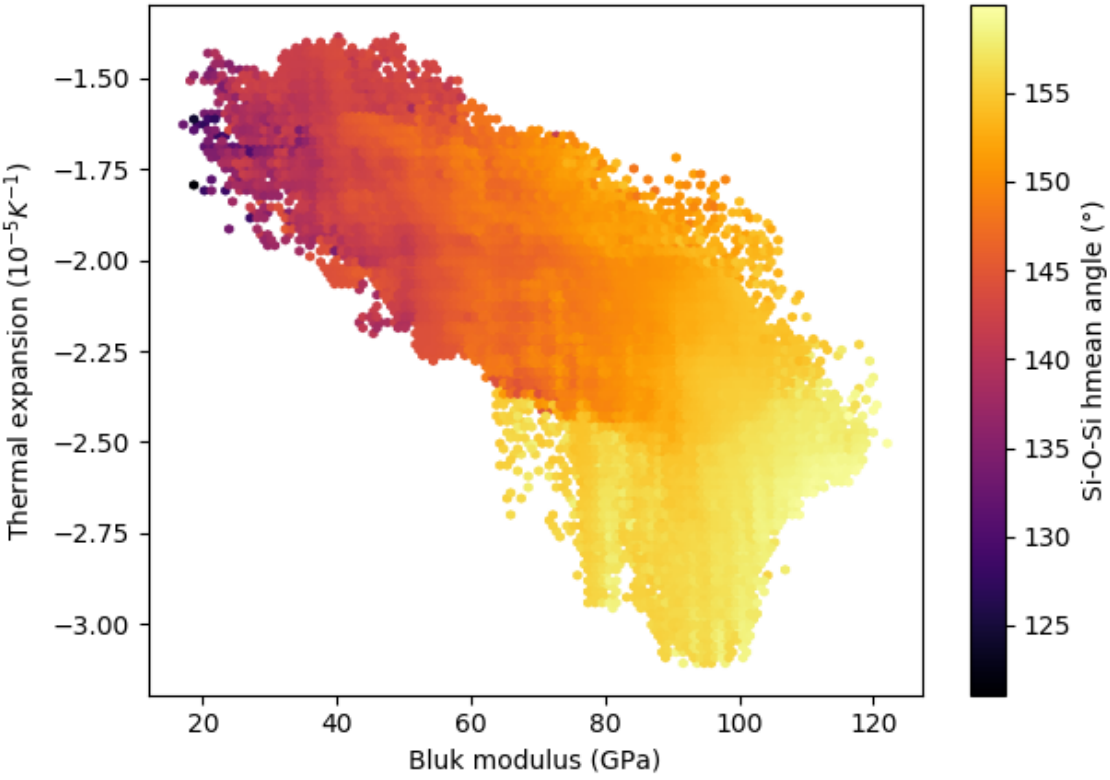


Figure 5: GBR prediction of the PCOD2 database on the thermal expansion and bulk modulus using the geometrical descriptors and our data set as training set.

Therefore, we decided instead to use the *ad hoc* geometrical descriptors and train a model to try to predict the properties of the database of theoretical zeolite structures. We used the

same set of statistic metrics on bond and angles as described earlier in this paper and used them to predict the thermal expansion and bulk modulus on the full PCOD2 database. Results are shown in Figure 5 — where they are plotted as a heat map, with the harmonic mean of Si–O–Si angle, using a gradient of color as a third dimension. We can observe some general trends on the PCOD2 structures, with a statistical correlation between thermal expansion and bulk modulus: structures with a higher bulk modulus are often found to correspond to larger (negative) values of negative thermal expansion (NTE). We can actually see a hint of this behavior, already, in our training data set with properties computed at the DFT level (Figure S3). We find that both physical properties are apparently linked to the same geometrical feature, namely, the Si–O–Si angle — which we found previously to be the most important feature when training ML models. Materials with higher values of Si–O–Si angles have both higher bulk modulus, and more pronounced NTE. This correlation between mechanical and thermal behavior through a relatively simple geometric feature is an interesting new development, and would have to be confirmed — for example, through systematic calculations of representative structures within the PCOD2 database.

Conclusions

In this work, we used our previously calculated database to train a series of machine learning models for the prediction of thermal properties of zeolites: including thermal expansion, as well as pressure and temperature dependence of bulk moduli. After identifying the optimal hyperparameters for our gradient boosting regression models, we compared the accuracy of models built on different types of structure descriptors: *ad hoc* geometric descriptors (based on Si–O distances and Si–O–Si angles), generic structures descriptors based on Smooth Overlap of Atomic Positions (SOAP), descriptors related to the geometrical characteristics of the porous network, and others related to the topology of the four-connected zeolitic net.

From these comparisons, we gained insight into the structure–property relationships in

zeolite frameworks. With regards to geometric descriptors, we saw that the description of angles is much more important than distances for the zeolite frameworks as angles were found to be the most important parameter when training the model. We also found that "agnostic" geometric descriptors such as SOAP outperform the *ad hoc* descriptors that we had identified, with lower root-mean-squared and mean-absolute errors. Moreover, even with SOAP descriptors, the most important component was found to be directly related to Si–O–Si angles, confirming our initial analysis. These conclusions are true regardless of the specific thermal property under study (i.e.n the thermal expansion, the bulk modulus or its derivatives), but we hypothesize that this would apply more broadly for a large scope of physical properties of zeolite frameworks.

Finally, we applied our predictive models to the PCOD2 database of hypothetical structures of zeolites, to confront our data set of DFT-calculated structures with a much larger set. Comparing the descriptors for both data set showed that the distributions of Si–O distances and Si–O–Si angles are within the same range, which hints that our database is well suited to predict properties of zeolites using the geometrical descriptors. We noted also that Si–O–Si angles exhibit a much larger variation of values compared to the distances. This is certainly one of the reason why the metrics on angles are much more important than metrics on distances. In the case of SOAP descriptors, distributions seem to be shifted compared to the PCOD2 database, suggesting that the hypothetical environments may be too diverse compared to feasible zeolites — and also showing that even though using SOAP descriptors resulted in the best predictions throughout this work, they are not suited for a prediction on the PCOD2 database. The geometrical descriptors were then used to predict the thermal expansion and bulk modulus of the PCOD2 database, which highlighted a trend where zeolites with a higher bulk modulus tends to have larger negative thermal expansion. We also find this to be linked to the Si–O–Si angles as the higher the angles the higher the bulk modulus.

These conclusions demonstrate the possibility for prediction of thermal properties of

framework materials based on purely geometrical characteristics, and expand the range of application of ML models on such systems, based on high-accuracy reference data obtained by quantum chemical calculations. Our study demonstrates the importance and impact of the choice of materials descriptors, and can be in future work extended to other physical properties, or to a broader range of materials — include variations in chemical composition, which would then need to be encoded into an entirely new class of features.

Acknowledgement

We acknowledge financial support from the Agence Nationale de la Recherche under project "MATAREB" (ANR-18-CE29-0009-01) and access to high-performance computing platforms provided by GENCI grant A0110807069.

Supporting Information Available

Learning curves, additional correlation plots, complete list of features of each set of descriptors

References

(1) Auerbach, S. M.; Carrado, K. A.; Dutta, P. K. *Handbook of Zeolite Science and Technology*; CRC Press: Boca Raton, 2003.

(2) International Zeolite Association, *Database of Zeolite Structures*, available online at `http://www.iza-structure.org/databases/` (accessed Dec 15, 2021).

(3) Fuchs, A. H.; Cheetham, A. K. Adsorption of Guest Molecules in Zeolitic Materials: Computational Aspects. *J. Phys. Chem. B* **2001**, *105*, 7375–7383.

(4) Smit, B.; Maesen, T. L. M. Molecular Simulations of Zeolites: Adsorption, Diffusion, and Shape Selectivity. *Chem. Rev.* **2008**, *108*, 4125–4184.

(5) Kim, J.; Lin, L.-C.; Martin, R. L.; Swisher, J. A.; Haranczyk, M.; Smit, B. Large-Scale Computational Screening of Zeolites for Ethane/Ethene Separation. *Langmuir* **2012**, *28*, 11914–11919.

(6) Fang, H.; Kulkarni, A.; Kamakoti, P.; Awati, R.; Ravikovitch, P. I.; Sholl, D. S. Identification of High-CO_2-Capacity Cationic Zeolites by Accurate Computational Screening. *Chem. Mater.* **2016**, *28*, 3887–3896.

(7) Thornton, A. W.; Winkler, D. A.; Liu, M. S.; Haranczyk, M.; Kennedy, D. F. Towards computational design of zeolite catalysts for CO_2 reduction. *RSC Adv.* **2015**, *5*, 44361–44370.

(8) Van Speybroeck, V.; Hemelsoet, K.; Joos, L.; Waroquier, M.; Bell, R. G.; Catlow, C. R. A. Advances in theory and their application within the field of zeolite chemistry. *Chem. Soc. Rev.* **2015**, *44*, 7044–7111.

(9) Sastre, G. Confinement effects in methanol to olefins catalysed by zeolites: A computational review. *Front. Chem. Sci. Eng.* **2016**, *10*, 76–89.

(10) Colón, Y. J.; Snurr, R. Q. High-throughput computational screening of metal–organic frameworks. *Chem. Soc. Rev.* **2014**, *43*, 5735–5749.

(11) Daglar, H.; Keskin, S. Recent advances, opportunities, and challenges in high-throughput computational screening of MOFs for gas separations. *Coord. Chem. Rev.* **2020**, *422*, 213470.

(12) Bertoldi, K.; Vitelli, V.; Christensen, J.; van Hecke, M. Flexible mechanical metamaterials. *Nat. Rev. Mater.* **2017**, *2*, 523.

(13) Coudert, F.-X.; Evans, J. D. Nanoscale metamaterials: Meta-MOFs and framework materials with anomalous behavior. *Coord. Chem. Rev.* **2019**, *388*, 48–62.

(14) Siddorn, M. J. Classifying and Identifying Negative Poisson's Ratio. An Examination of the Auxeticity in Zeolitic Materials. Ph.D. thesis, University of Exeter, Exeter, 2014.

(15) Grima, J. N.; Jackson, R.; Alderson, A.; Evans, K. E. Do Zeolites Have Negative Poisson's Ratios? *Adv. Mater.* **2000**, *12*, 1912–1918.

(16) Evans, K. E.; Alderson, A. Auxetic Materials: Functional Materials and Structures from Lateral Thinking! *Adv. Mater.* **2000**, *12*, 617–628.

(17) Cairns, A. B.; Goodwin, A. L. Negative linear compressibility. *Phys. Chem. Chem. Phys.* **2015**, *17*, 20449–20465.

(18) Dudek, K. K.; Attard, D.; Caruana-Gauci, R.; Wojciechowski, K. W.; Grima, J. N. Unimode metamaterials exhibiting negative linear compressibility and negative thermal expansion. *Smart Mater. Struct.* **2016**, *25*, 025009.

(19) Nearchou, A.; Cornelius, M.-L. U.; Jones, Z. L.; Collings, I. E.; Wells, S. A.; Raithby, P. R.; Sartbaeva, A. Pressure-induced symmetry changes in body-centred cubic zeolites. *R. Soc. Open Sci.* **2019**, *6*, 182158.

(20) Miller, W.; Smith, C. W.; Mackenzie, D. S.; Evans, K. E. Negative thermal expansion: a review. *J. Mater. Sci.* **2009**, *44*, 5441–5451.

(21) Ducamp, M.; Coudert, F.-X. Systematic Study of the Thermal Properties of Zeolitic Frameworks. *J. Phys. Chem. C* **2021**, *125*, 15647–15658.

(22) Evans, J.; Mary, T.; Sleight, A. Negative thermal expansion materials. *Phys. B: Condens. Matter* **1997**, *241-243*, 311–316.

(23) Yang, X.; Xu, J.; Li, H.; Cheng, X.; Yan, X. In Situ Synthesis of ZrO_2/ZrW_2O_8 Composites With Near-Zero Thermal Expansion. *J. Am. Ceram. Soc.* **2007**, *90*, 1953–1955.

(24) Zhou, C.; Zhou, Y.; Zhang, Q.; Meng, Q.; Zhang, L.; Kobayashi, E.; Wu, G. Near-zero thermal expansion of ZrW_2O_8/Al–Si composites with three dimensional interpenetrating network structure. *Compos. B Eng.* **2021**, *211*, 108678.

(25) Butler, K. T.; Davies, D. W.; Cartwright, H.; Isayev, O.; Walsh, A. Machine learning for molecular and materials science. *Nature* **2018**, *559*, 547–555.

(26) Schmidt, J.; Marques, M. R. G.; Botti, S.; Marques, M. A. L. Recent advances and applications of machine learning in solid-state materials science. *npj Comput. Mater.* **2019**, *5*, 484.

(27) Schleder, G. R.; Padilha, A. C. M.; Acosta, C. M.; Costa, M.; Fazzio, A. From DFT to machine learning: recent approaches to materials science–a review. *J. Phys. Mater.* **2019**, *2*, 032001.

(28) Chibani, S.; Coudert, F.-X. Machine learning approaches for the prediction of materials properties. *APL Mater.* **2020**, *8*, 080701.

(29) Kauwe, S. K.; Graser, J.; Murdock, R.; Sparks, T. D. Can machine learning find extraordinary materials? *Comput. Mater. Sci.* **2020**, *174*, 109498.

(30) Bartók, A. P.; De, S.; Poelking, C.; Bernstein, N.; Kermode, J. R.; Csányi, G.; Ceriotti, M. Machine learning unifies the modeling of materials and molecules. *Sci. Adv.* **2017**, *3*, e1701816.

(31) Schütt, K. T.; Glawe, H.; Brockherde, F.; Sanna, A.; Müller, K. R.; Gross, E. K. U. How to represent crystal structures for machine learning: Towards fast prediction of electronic properties. *Phys. Rev. B* **2014**, *89*, 1875.

(32) Singh, H. K.; Zhang, Z.; Opahle, I.; Ohmer, D.; Yao, Y.; Zhang, H. High-Throughput Screening of Magnetic Antiperovskites. *Chem. Mater.* **2018**, *30*, 6983–6991.

(33) de Jong, M.; Chen, W.; Notestine, R.; Persson, K.; Ceder, G.; Jain, A.; Asta, M.; Gamst, A. A Statistical Learning Framework for Materials Science: Application to Elastic Moduli of k-nary Inorganic Polycrystalline Compounds. *Sci Rep* **2016**, *6*, 15004.

(34) Evans, J. D.; Coudert, F.-X. Predicting the Mechanical Properties of Zeolite Frameworks by Machine Learning. *Chem. Mater.* **2017**, *29*, 7833–7839.

(35) Moghadam, P. Z.; Rogge, S. M.; Li, A.; Chow, C.-M.; Wieme, J.; Moharrami, N.; Aragones-Anglada, M.; Conduit, G.; Gomez-Gualdron, D. A.; Van Speybroeck, V. et al. Structure-Mechanical Stability Relations of Metal-Organic Frameworks via Machine Learning. *Matter* **2019**, *1*, 219–234.

(36) Kitchin, J. R. Machine learning in catalysis. *Nature Catal.* **2018**, *1*, 230–232.

(37) Elton, D. C.; Boukouvalas, Z.; Butrico, M. S.; Fuge, M. D.; Chung, P. W. Applying machine learning techniques to predict the properties of energetic materials. *Sci Rep* **2018**, *8*, 11793.

(38) Deem, M. W.; Pophale, R.; Cheeseman, P. A.; Earl, D. J. Computational Discovery of New Zeolite-Like Materials. *J. Phys. Chem. C* **2009**, *113*, 21353–21360.

(39) Yang, S.; Lach-hab, M.; Vaisman, I. I.; Blaisten-Barojas, E. Identifying Zeolite Frameworks with a Machine Learning Approach. *J. Phys. Chem. C* **2009**, *113*, 21721–21725.

(40) Carr, D. A.; Lach-hab, M.; Yang, S.; Vaisman, I. I.; Blaisten-Barojas, E. Machine learning approach for structure-based zeolite classification. *Micropor. Mesopor. Mater.* **2009**, *117*, 339–349.

(41) De, S.; Bartók, A. P.; Csányi, G.; Ceriotti, M. Comparing molecules and solids across structural and alchemical space. *Phys. Chem. Chem. Phys.* **2016**, *18*, 13754–13769.

(42) 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.

(43) Daeyaert, F.; Ye, F.; Deem, M. W. Machine-learning approach to the design of OSDAs for zeolite beta. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 3413–3418.

(44) Moliner, M.; Román-Leshkov, Y.; Corma, A. Machine Learning Applied to Zeolite Synthesis: The Missing Link for Realizing High-Throughput Discovery. *Acc. Chem. Res.* **2019**, *52*, 2971–2980.

(45) Ting, K. W.; Kamakura, H.; Poly, S. S.; Takao, M.; Siddiki, S. M. A. H.; Maeno, Z.; Matsushita, K.; Shimizu, K.-i.; Toyao, T. Catalytic Methylation of *m*-Xylene, Toluene, and Benzene Using CO_2 and H_2 over TiO_2-Supported Re and Zeolite Catalysts: Machine-Learning-Assisted Catalyst Optimization. *ACS Catal.* **2021**, *11*, 5829–5838.

(46) Kim, N.; Min, K. Accelerated Discovery of Zeolite Structures with Superior Mechanical Properties via Active Learning. *J. Phys. Chem. Lett.* **2021**, *12*, 2334–2339.

(47) Gaillac, R.; Chibani, S.; Coudert, F.-X. Speeding Up Discovery of Auxetic Zeolite Frameworks by Machine Learning. *Chem. Mater.* **2020**, *32*, 2653–2663.

(48) Chibani, S.; Coudert, F.-X. Systematic exploration of the mechanical properties of 13 621 inorganic compounds. *Chem. Sci.* **2019**, *10*, 8589–8599.

(49) Coudert, F.-X. Systematic investigation of the mechanical properties of pure silica zeolites: stiffness, anisotropy, and negative linear compressibility. *Phys. Chem. Chem. Phys.* **2013**, *15*, 16012.

(50) Pophale, R.; Cheeseman, P. A.; Deem, M. W. A database of new zeolite-like materials. *Phys. Chem. Chem. Phys.* **2011**, *13*, 12407.

(51) Deem, M. W.; Newsam, J. M. Framework crystal structure solution by simulated annealing: test application to known zeolite structures. *J. Am. Chem. Soc.* **1992**, *114*, 7189–7198.

(52) Earl, D. J.; Deem, M. W. Toward a Database of Hypothetical Zeolite Structures. *Ind. Eng. Chem. Res.* **2006**, *45*, 5449–5454.

(53) Li, Y.; Yu, J. New Stories of Zeolite Structures: Their Descriptions, Determinations, Predictions, and Evaluations. *Chem. Rev.* **2014**, *114*, 7268–7316.

(54) Pophale, R.; Daeyaert, F.; Deem, M. W. Computational prediction of chemically synthesizable organic structure directing agents for zeolites. *J. Mater. Chem. A* **2013**, *1*, 6750.

(55) Friedman, J. H. Greedy function approximation: A gradient boosting machine. *Ann. Stat.* **2001**, *29*, 1189–1232.

(56) Friedman, J.; Hastie, T.; Tibshirani, R. *The elements of statistical learning - Data mining, inference and prediction*, 2nd ed.; Springer: New York, 2009.

(57) Caruana, R.; Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. Proceedings of the 23rd international conference on Machine learning - ICML '06. 2006.

(58) 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.

(59) Wang, X.; Simard, M.; Wuest, J. D. Molecular Tectonics. Three-Dimensional Organic Networks with Zeolitic Properties. *J. Am. Chem. Soc.* **1994**, *116*, 12119–12120.

(60) Wragg, D. S.; Morris, R. E.; Burton, A. W. Pure Silica Zeolite-type Frameworks: A Structural Analysis. *Chem. Mater.* **2008**, *20*, 1561–1570.

(61) Ong, S. P.; Richards, W. D.; Jain, A.; Hautier, G.; Kocher, M.; Cholia, S.; Gunter, D.; Chevrier, V. L.; Persson, K. A.; Ceder, G. Python Materials Genomics (pymatgen): A

robust, open-source python library for materials analysis. *Comput. Mater. Sci.* **2013**, *68*, 314–319.

(62) *Identification and Classification of Crystal Topologies*, Lionel Zoubritzky, Internship report, École normale supérieure — PSL University, Spring 2021.

(63) Willems, T. F.; Rycroft, C. H.; Kazi, M.; Meza, J. C.; Haranczyk, M. Algorithms and tools for high-throughput geometry-based analysis of crystalline porous materials. *Micropor. Mesopor. Mater.* **2012**, *149*, 134–141.

(64) Martin, R. L.; Smit, B.; Haranczyk, M. Addressing Challenges of Identifying Geometrically Diverse Sets of Crystalline Porous Materials. *J. Chem. Inf. Model.* **2011**, *52*, 308–318.

(65) Bartók, A. P.; Kondor, R.; Csányi, G. On representing chemical environments. *Phys. Rev. B* **2013**, *87*, 184115.

(66) Engel, E. A.; Anelli, A.; Ceriotti, M.; Pickard, C. J.; Needs, R. J. Mapping uncharted territory in ice from zeolite networks to ice structures. *Nature Commun.* **2018**, *9*, 135701.

(67) Nicholas, T. C.; Goodwin, A. L.; Deringer, V. L. Understanding the geometric diversity of inorganic and hybrid frameworks through structural coarse-graining. *Chem. Sci.* **2020**, *11*, 12580–12587.

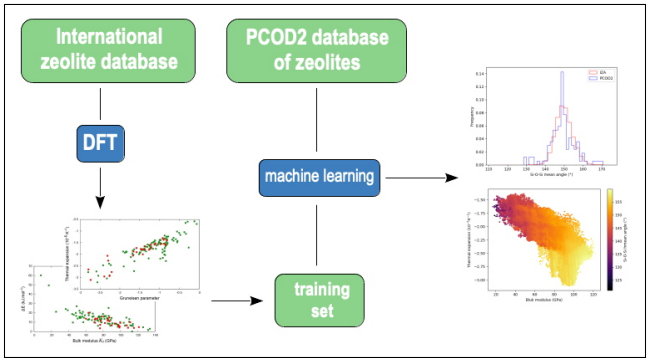
(68) Byggmästar, J.; Hamedani, A.; Nordlund, K.; Djurabekova, F. Machine-learning interatomic potential for radiation damage and defects in tungsten. *Phys. Rev. B* **2019**, *100*, 144105.

(69) 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.

(70) 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.

(71) Sanders, M. J.; Leslie, M.; Catlow, C. R. A. Interatomic potentials for SiO_2. *J. Chem. Soc. Chem. Commun.* **1984**, 1271–1273.

TOC Graphic


International zeolite database
PCOD2 database of zeolites
DFT
machine learning
training set