Supplementary Information

MODNet: accurate and interpretable property predictions for limited materials datasets by feature selection and joint-learning

Pierre-Paul De Breuck,[1,*] Geoffroy Hautier,[1,2] and Gian-Marco Rignanese[1,†]

[1] *UCLouvain, Institute of Condensed Matter and Nanosciences (IMCN), Chemin des Étoiles 8, B-1348 Louvain-la-Neuve, Belgium*

[2] *Thayer School of Engineering, Dartmouth College, Hanover, New Hampshire 03755, USA*

(Dated: May 11, 2021)

* pierre-paul.debreuck@uclouvain.be

† gian-marco.rignanese@uclouvain.be

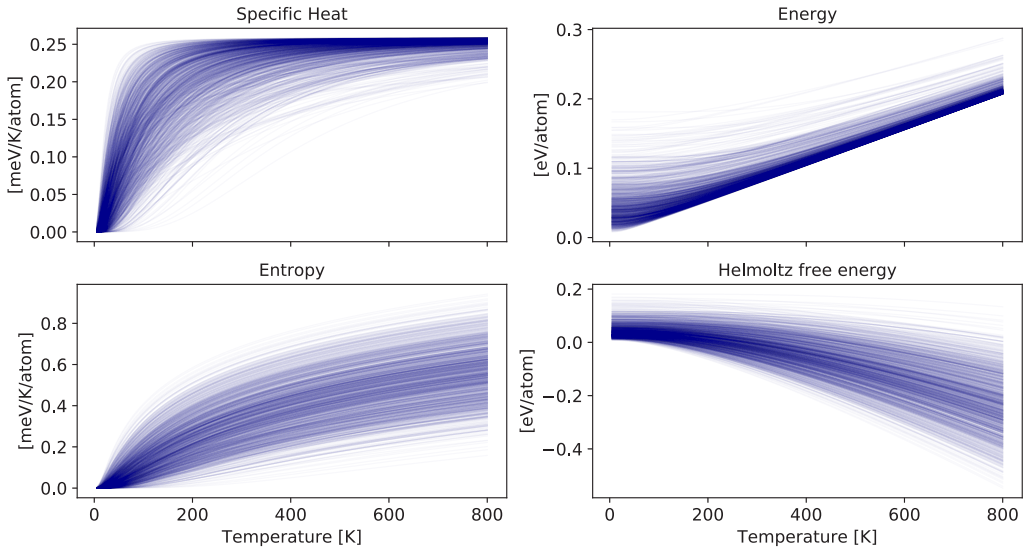
I. SUPPLEMENTARY METHODS AND DISCUSSION

A. Datasets

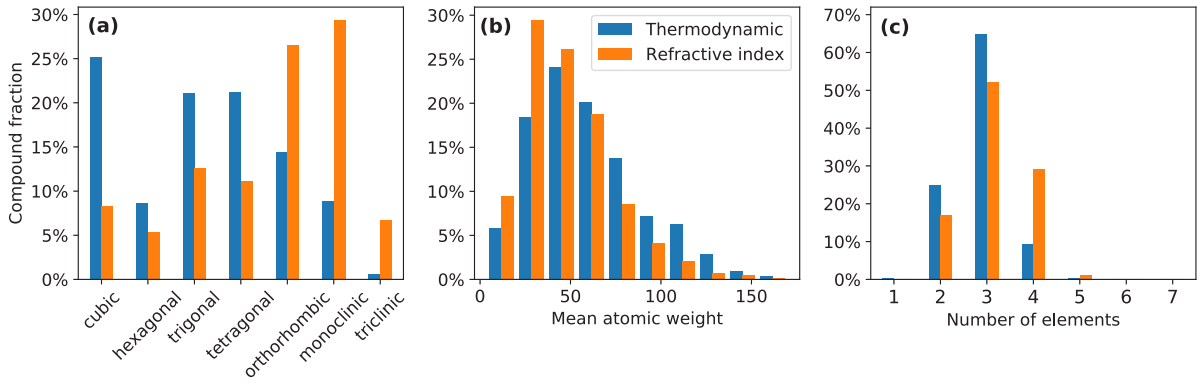
The different models used in this work were trained on four different properties: formation energy, band gap, refractive index and vibrational thermodynamics. We briefly present these latter two datasets as they are new to the machine learning field. Petretto *et al.* [1] computed the vibrational properties for 1,245 inorganic compounds, in the harmonic approximation based on Density Functional Perturbation Theory (DFPT). In particular, this dataset contains the following thermodynamic properties: vibrational entropy, Helmholtz free energy, internal energy and heat capacity from 5 to 800 K in steps of 5 K. Supplementary Figure 1 graphically represents these four properties from 5 to 800 K for all materials contained in the dataset in meV/atom or meV/K/atom. A wide variety of values is reached, with different Debye temperatures as can be seen from the specific heat. This is important to notice as it indicates that there is no significant bias, giving us confidence for generalizing on unseen data.

The refractive index for 4040 compounds were computed by Naccarato *et al.* (see Ref. [2]) relying on Density Functional Theory (DFT) and high-throughput methods. Typical values encountered in the dataset ranges from 1 to 6, with 60% below 2.

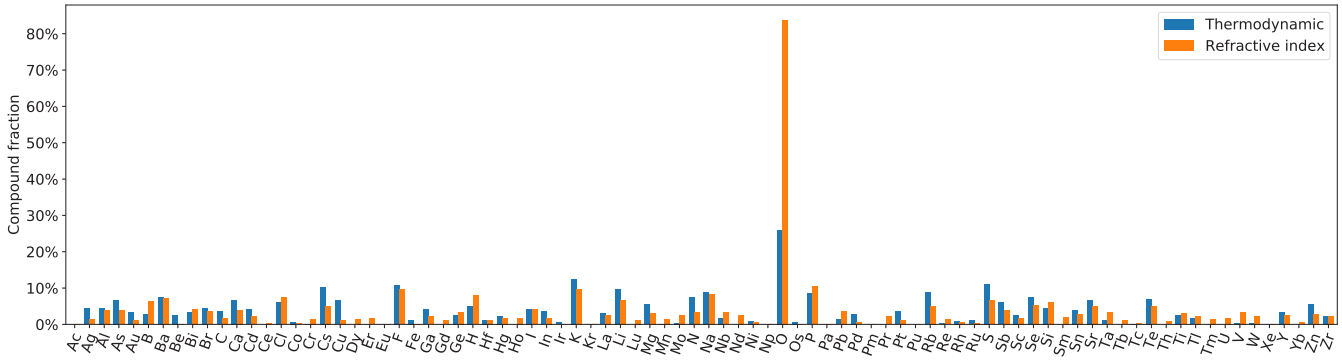
Supplementary Figures 2, 3 and 4 contain various histograms representing the data distribution for both properties. Various crystalline compounds are present ranging from simple mono-elemental compounds to complex semiconductors. The thermodynamic data (resp. refractive index) cover the 7 (resp. 7) symmetry groups, 84 (resp. 165) space groups and 64 (resp. 86) elements. Most compounds are ternary alloys and there are no (resp. almost no) materials with more than 5 different elements. The mean atomic mass has a large range, confirming the variety of elements present in these two datasets. Note that the refractive index dataset is biased towards oxides, with 27 % of all materials containing oxygen atoms.



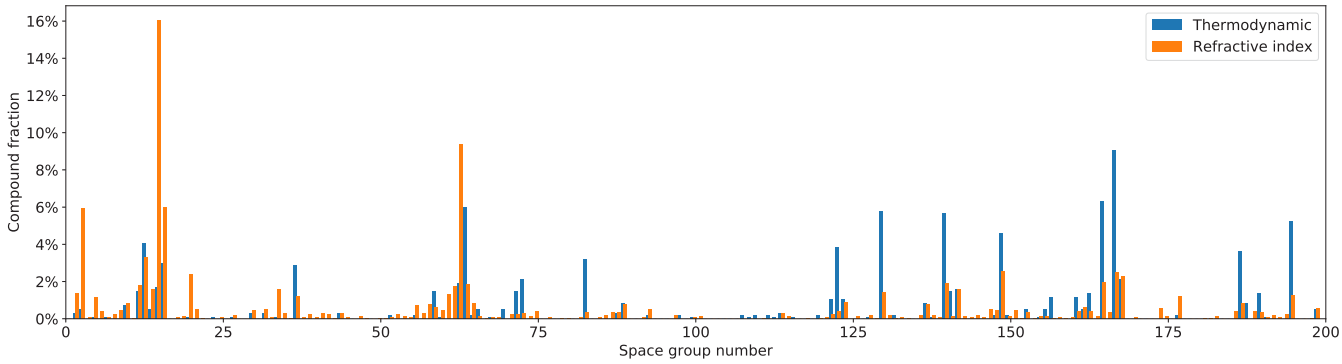
Supplementary Figure 1. Visual representation of values encountered in the vibrational thermodynamic dataset (from Ref. [1]). Four vibrational properties from 5 to 800 K for 1245 materials are present. A wide range of values are reached, with clear differences in Debye temperatures.



Supplementary Figure 2. Relative distribution of (a) the crystal system, (b) mean atomic mass and (c) number of elements for both the thermodynamic and refractive index datasets.



Supplementary Figure 3. Relative amount of samples containing a particular element for both the thermodynamic and refractive index datasets. In total, they cover 64 and 86 elements respectively.



Supplementary Figure 4. Space group distribution for both the thermodynamic and refractive index datasets. In total they cover 84 and 165 spacegroups respectively.

B. Identifying compound-property relations with feature selection

A detailed description of the features used in Figure 4 (and accompanying discussion) is given here.

- **AGNIFingerPrint:** This feature represents the local environment of an atom by computing the integral of the product of the radial distribution function (RDF) with a Gaussian window function. It was originally proposed by Botu *et al.* to fit empirical energy potentials [3]. It is defined as:

$$G_i(\eta) = \int R_i(r) e^{-(\frac{r_{ij}}{\eta})^2} dr \tag{1}$$

where i is the considered site, $R(r)$ the RDF, r_{ij} the distance between site i and neighbor j and η giving the width of the Gaussian window. In practice, this integral is discretized and a cut-off function is used, see the original work for more information [3].

When computed for a compound, the mean *AGNIFingerPrint* over all sites is taken.

- **p-valence range:** Range of p-electrons, i.e. the number of p-electrons is first computed for each element in the compound, and then the difference between the highest and lowest number is taken.
- **Maximum GS bandgap** Highest bandgap among the ground state (GS) constituent elemental compounds. In other words, for each element in the composition, the corresponding most stable single-elemental structure is taken with a corresponding band gap. Then the maximum over these different elemental bandgaps is taken.
- **Density:** Mass per unit volume of the compound.
- **Minimum $\Delta\chi$:** Here $\Delta\chi$ represents the difference in electronegativity between a given cation and the mean electronegativity of all anions in the compound. This is first computed for each cation. Then the minimum is taken over all cations.
- **Generalized RDF:** This featurizer generalizes the RDF by using non-rectangular bins (Gaussian binning is used here), see Ref. [4]. It is therefore very similar to the previously explained *AGNIFingerprint*. Here the integral of the RDF with a Gaussian of mean 4.0 Å and width 1.0 Å is taken.
- **HOMO:** The highest occupied molecular orbital (HOMO) is estimated from the atomic orbital energies of the composition (see Ref. [5]).
- **Maximum Packing Efficiency:** This feature (see Ref. [6]) represents the packing efficiency of the crystal by measuring the volume fraction of occupied atoms. The algorithm first determines the largest radius for each atom, which equals the distance from the center of the atom to the closest Voronoi face. Then, the packing efficiency is computed as the following volume ratio:

$$\frac{1}{V_{cell}} \sum_i \frac{4}{3} \pi r_i^3 \tag{2}$$

where V_{cell} is the volume of the unit cell and r_i the largest radius (as defined before) of each atom i in the unit cell.

Note that, exceptionally, some features are unable to be computed for a few compounds, due to missing information needed by the underlying routines. When this happens, the corresponding value is set to -1, which explains negative values for non-negative properties in Fig. 4.

C. Model testing procedure

Supplementary Table I summarizes training, validation and test set sizes for each property used in this work.

Supplementary Table I. Training, validation and test sizes used for training the models on the different datasets, see Table I of the main paper. All properties (except the complete formation and band gap energy datasets) use a 10-fold cross validation scheme. Validation sizes always correspond to 1/10th of the total train-validation set.

Property	Training size	Validation size	Test size
E_f	504	56	4990
	60 000	6 667	1 822
E_g	504	56	4 990
	60 000	6 667	1 822
E_g^{nz}	504	56	4 990
	36 720	4 080	4 628
n	3 240	360	425
S_{305K}	990	110	125

D. Hyperparameter optimization of MODNet when learning on multiple properties

Hyperparameter optimization is done following a 10-fold validation scheme on the previously presented dataset. The network is trained on 90% of the data, then validated on the remaining 10% by computing the MAE of each property (4 thermodynamic quantities, each at 20 temperatures). This is repeated ten times, and the mean over the ten folds is taken.

It is important to note that the network is optimized on all targets together. Therefore, different MAEs corresponding to different properties are combined to form the following criteria:

- **WMAE**: A Weighted summed MAE of the different properties such that the property-temperature curves are in the same range:

$$\text{WSMAE} = \sum_{p\in(S,C_v,H,U)} w_p \; MAE_p, \qquad (w_S, w_{Cv}, w_H, w_U) = (0.3, 1, 0.004, 0.001) \tag{3}$$

- **MAX WSMAE**: The WSMAE, but only taking the 5% worst predicted samples.

- **MAPE**: A mean absolute percentage error over all properties.

First, the learning rate of MODNet was optimized to 0.01 from the set {0.1, 0.1, 0.005, 0.001}, by fixing all other hyperparameters on arbitrarily chosen values.

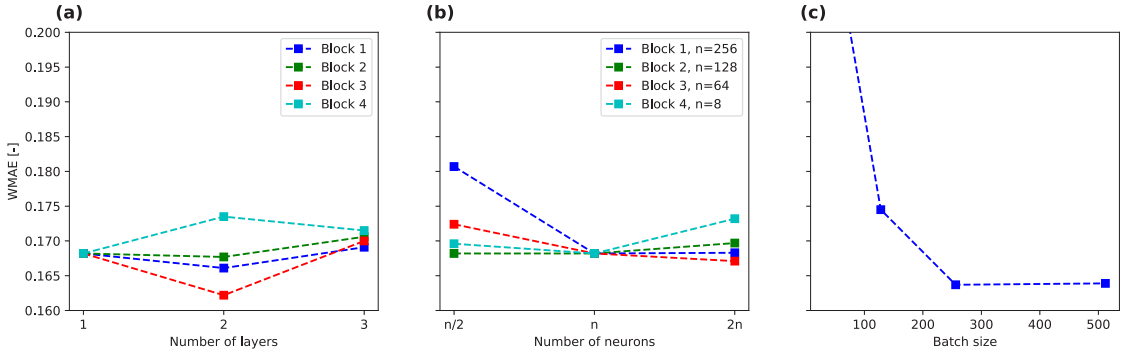
We then optimize the global loss function, by studying the effect of the following hyperparameters: input scaling, output scaling, loss function and loss weights. These two latter parameters are very important as the peculiarity of our network consists in optimizing the predictions of *multiple* targets. Therefore, each target should have an equal contribution in the loss functions. For example, the vibrational energy and enthalpy are typically 3 orders of magnitude larger then the vibrational entropy and specific heat, which causes an imbalance in the summation of the loss function. Different solutions exist, such as scaling the targets or weighting targets in the loss function.

Supplementary Table III presents the MAE (over 10 folds) for various combinations of previous loss-parameters. For scaling, a min-max or standard scaling were tested. For the loss function, a Mean Square Error (MSE) or MAE were compared. Furthermore, weighting the 4 different properties such that the average property-temperature curves are in the same range was also compared against scaling. From the results (see Supplementary Table III), it was chosen to use a min-max scaling of the input variables, a MSE loss function and weighting of the different properties instead of scaling.

Once the learning rate, scaling, loss function and weighting are fixed, the number of layers as well as the number of neurons can be optimized. We choose two layers per block (except for the last one), with 256, 128, 64 and 8 neurons in the successive blocks, see Fig. 6. Adding (or removing) a layer, as well as doubling or halving the number of neurons does not improve accuracy as can be seen in Fig. 5. From the same figure, the batch size was fixed to

256. Other architectures that were tested are listed in Supplementary Table IV with the corresponding performance criteria. Finally, the number of features were optimized, see next section.

The final architecture is depicted in Fig. 6. A rectified linear unit function (ReLU) is used as activation for each layer. Learning is performed using an Adam optimizer ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $decay = 0$) on 600 epochs.



Supplementary Figure 5. Optimization of the MODNet architecture . The effect of adding or removing layers (a) at each block is shown as well as reducing or increasing the number of neurons (b). The effect of changing the batch size is also shown (c). When one hyperparameter is changed, others are kept fixed at respectively 256, 128, 64 and 8 neurons (and 1 layer) per block.

Supplementary Table II. Architecture and hyperparameters of the MODNet for each property of Table I. Architecture is given in number of neurons per layer separated by a comma.

Property	Training size	Architecture	Activation	Loss	Batch size	Learning rate	# optimal features
E_f	504	256, 128	ReLU	MAE	64	0.001	1000
	60,000	1024, 256, 128	ELU	MAE	64	0.001	2000
E_g	504	256, 128, 64, 64, 32, 32	ELU	MAE	64	0.001	1000
	60,000	256, 128	ELU	MAE	64	0.001	1000
n	36,720	128, 64, 32	ELU	MAE	64	0.001	1000

Supplementary Table III. Scaling and loss optimization of MODNet when learning on the four vibrational properties and formation energy. The validation error (three custom defined criteria, see text, and S_{305K}) is reported for each set of hyperparameters.

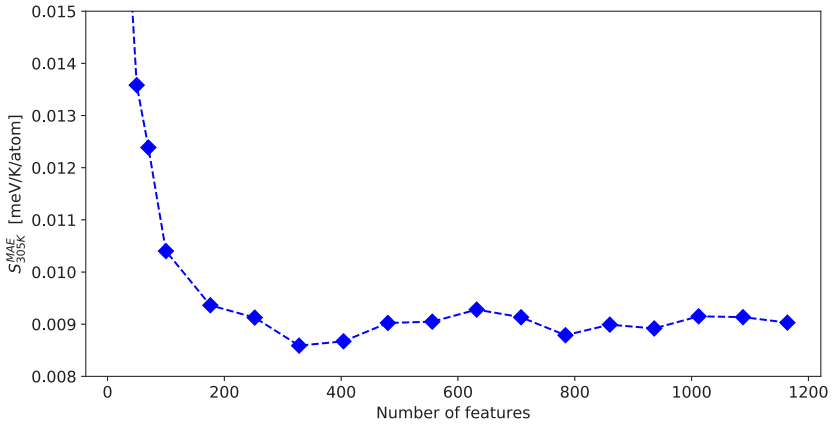
Input scale	Output scale	Loss function	Loss weights [a]	WMAE	MAX WMAE	MAPE	S_{305K} [meV/K/atom]
min-max	No	MSE	Yes, squared	0.167	0.435	0.0653	0.00860
Standard	No	MSE	Yes, squared	0.185	0.482	0.0760	0.00959
No	No	MSE	Yes, squared	0.416	1.220	0.158	0.0207
min-max	No	MAE	Yes	0.166	0.447	0.0683	0.00881
min-max	min-max	MAE	No	0.172	0.454	0.0700	0.00894
min-max	min-max	MAE	No	0.164	0.441	0.0680	0.00890
min-max	min-max	MSE	No	0.259	0.630	0.109	0.01253
min-max	min-max	MSE	No	0.1677	0.440	0.0666	0.00894

Supplementary Table IV. Supplementary architectures for optimization of MODNet when learning on the four vibrational properties and formation energy. The validation error (three custom defined criteria, see text, and S_{305K}) is reported for each set of hyperparameters.

Block 1	Block 2	Block 3	Block 4	WMAE	WMAE MAX	MAPE	S_{305K} [meV/K/atom]
256,256	128,128	64,64	8	0.166	0.434	0.0666	0.00873
256,256	128,128	64,64	8,8	0.169	0.441	0.0679	0.00881
128,128,128	64,64,64	32,32,32	8,8	0.182	0.468	0.0730	0.00945
512	512	128	64	0.1646	0.429	0.0668	0.00866
256, 128, 32	32, 32	32, 64	16	0.174	0.455	0.0712	0.00892
512	128	64	8, 8, 8	0.1698	0.441	0.0688	0.00876
512	128	64	4, 4, 4	0.1651	0.433	0.0658	0.00840
512	128	16	3, 3, 3	0.1702	0.440	0.0686	0.00863

E. Optimal number of features for m-MODNet

The effect of the number of optimal features is assessed here. This is done by computing the MAE for S_{305K} of m-MODNet as a function of the number of relevant features. As one can observe in Fig. 6, the MAE is monotonically decreasing until ∼330 features, which shows that newly added features are indeed relevant to the prediction, followed by a gentle increase in MAE. One can thus safely ignore the remaining features. Note that the observed increase after the minimum is quite weak, and taking all computed features would not be dramatic. From our knowledge, this is because 2000 features is still small for a network such as the m-MODNet which drastically limits overfitting by its joint-nature. In contrast, the single target MODNet benefits much more from feature selection as was stated in the main paper. One could go beyond by computing even more features. For example, we recommend computing multiple functions ($\frac{1}{x}$,x^2, $log(x)$, e^x, ...), called feature engineering (see Ref. [7] for an empirical analysis), which would certainly improve predictions. Nevertheless, reducing the number of features is still beneficial in our case, as it slightly reduces the validation error, reduces computation time and moreover enables to identify (and understand) important features.



Supplementary Figure 6. MAE of S_{305K} as a function of the number of features selected by our feature selection algorithm for m-MODNet.

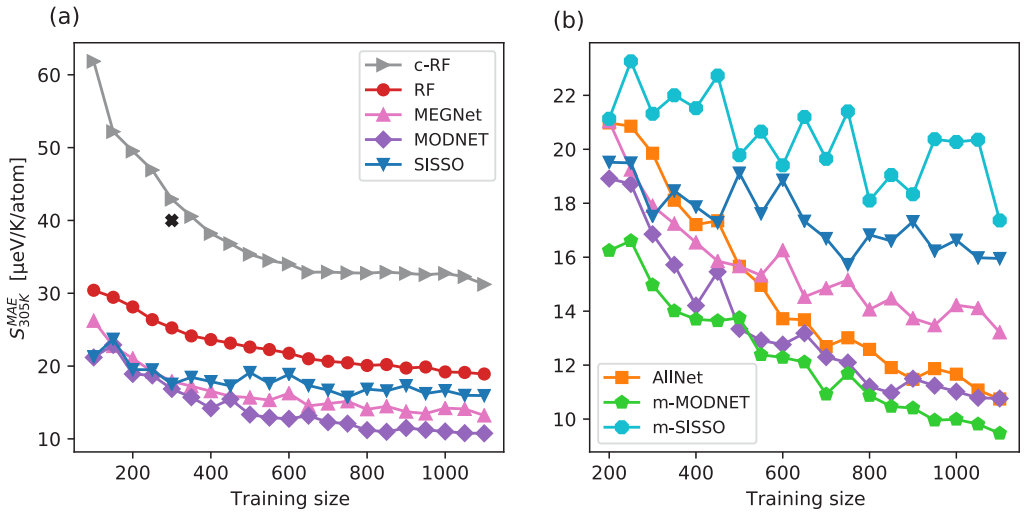
F. Practical considerations of SISSO

The screening and sparsifying operator (SISSO) model has been fitted on various properties in this work for comparison purposes. Some important practical aspects are detailed in this section.

The SISSO framework typically starts from a few tens of features, iteratively applies operators on them in order to generate an enormous amount of combined features before proceeding to the SISSO screening itself. In our work, we automatically generate thousands of features for each material. Using all of them as starting point in the SISSO-approach is in fact unfeasible as the amount of secondary and ternary features grow too quickly. Therefore, one should first select the most appropriate primary features before running the SISSO framework.

We chose to proceed as follow. A first SISSO is executed without any operators (i.e. rung set to 0), and with a SIS-subspace of 3 features. The 50-dimensional model is then used to select the corresponding 50 most significant features. Then a second SISSO is executed on these 50 features with a rung of 2 and SIS-space of 20 features. Then the best model equal or below 5D is chosen before being evaluated on the hold-out test set. In other words we use a double SISSO, with the first one used to do feature selection.

G. MODNet and vibrational thermodynamic data: supplemental results



Supplementary Figure 7. Test MAE on the vibrational entropy at 305K (S_{305K} in μeV/K/atom) as a function of the training size for various strategies (see main text for a detailed description). (a) Random Forest (RF) approaches are compared to neural-network models (Nets) and SISSO. The black cross indicates the RF result of Legrain *et al.* [8] based on 300 materials. (b) Best performing models and multi-target approaches are confronted.

Figure 2 in the main letter contains a quantitative comparison of different ML models on the vibrational entropy at 305K. The error density distribution was given for three training sizes. For sake of completeness, we report here, in Supplementary Figure 7, the mean absolute error for various training sizes (ranging from 100 to 1100 in steps of 50) for the same ML methods (please refer to the main text for the different approaches). Moreover, the expected error from a previous work by Legrain *et al.* on vibrational thermodynamics based on a RF on the composition alone trained on 300 materials, is indicated by a black cross on this figure [8].

From Supplementary Figure 7 (a), it can be seen that neural-network models perform better than RF approaches whatever the size of the dataset. This can probably be explained by the rather complex non-linear nature of the regression problem. Furthermore, MODNet leads to smaller errors than MEGNet confirming our previous finding that it is more appropriate for small datasets. For the RF approaches, adding structural features is clearly beneficial,

as explained in the main text. SISSO has a comparable mean error than best methods around and below 300 samples. However beyond this point, no significant improve in mean error is seen, in contrast with other algorithms.

In Fig. 7(b), best performing methods are confronted. Whatever the training size in the range available here, adding feature selection (compare AllNet to MODNet) reduces the prediction error. This is especially true at low sample sizes, where helping the learning algorithm by giving the right descriptors is crucial. The MEGNet model with transfer learning performs somewhat better than AllNet below 500 samples but it is outperformed beyond this point by all neural network models based on physical descriptors. Finally, our joint-learning approach that adds all other thermodynamic properties (m-MODNet) shows in average a slight improve in accuracy (∼8%). This is in contrast with m-SISSO, which does not show an improve in accuracy with respect to singe target learning.

[1] G. Petretto, S. Dwaraknath, H. P. C. Miranda, D. Winston, M. Giantomassi, M. J. van Setten, X. Gonze, K. A. Persson, G. Hautier, and G.-M. Rignanese, Sci. Data **5**, 180065 (2018).

[2] F. Naccarato, F. Ricci, J. Suntivich, G. Hautier, L. Wirtz, and G.-M. Rignanese, Phys. Rev. Materials **3**, 044602 (2019).

[3] V. Botu and R. Ramprasad, Int. J. Quantum Chem. **115**, 1074 (2015).

[4] A. Seko, H. Hayashi, K. Nakayama, A. Takahashi, and I. Tanaka, Phys. Rev. B **95**, 144110 (2017).

[5] S. Kotochigova, Z. H. Levine, E. L. Shirley, M. D. Stiles, and C. W. Clark, Phys. Rev. A **55**, 191 (1997).

[6] L. Ward, R. Liu, A. Krishna, V. I. Hegde, A. Agrawal, A. Choudhary, and C. Wolverton, Phys. Rev. B **96**, 024104 (2017).

[7] J. Heaton, in *SoutheastCon 2016* (IEEE, Norfolk, VA, USA, 2016) pp. 1–6.

[8] F. Legrain, J. Carrete, A. van Roekeghem, S. Curtarolo, and N. Mingo, Chem. Mater. **29**, 6220 (2017).