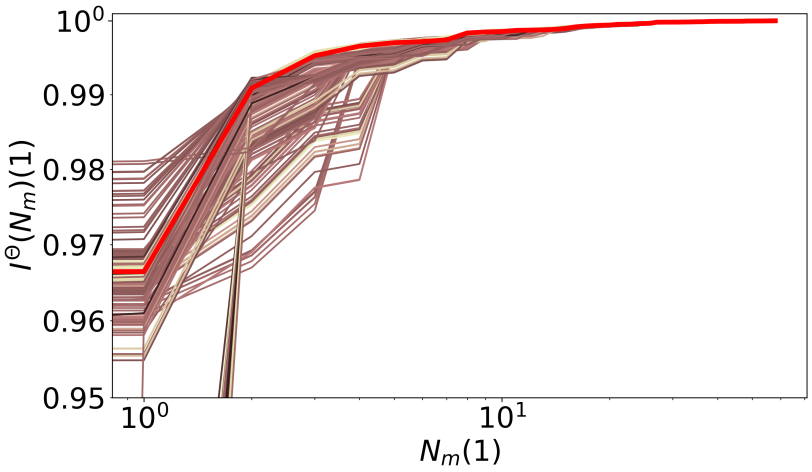
Efficient Gaussian Process Regression for prediction of molecular crystals harmonic free energies - supplementary information

Marcin Krynski[1,2]*, Mariana Rossi[1,3]

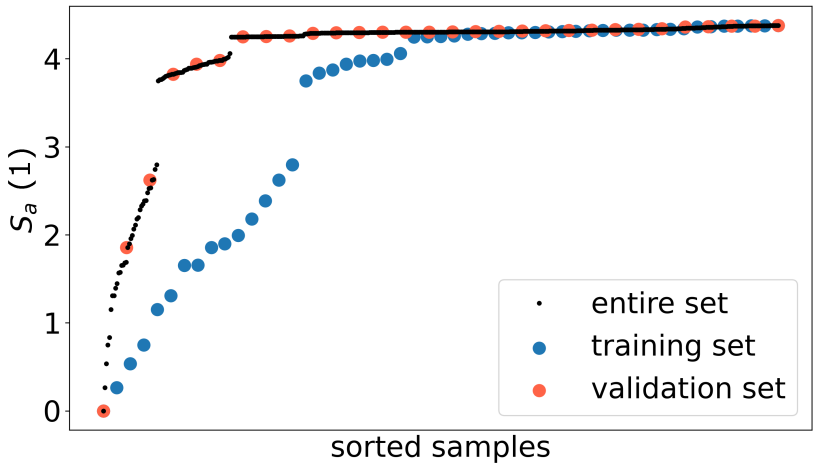
August 5, 2021

1. Fritz Haber Institute of the Max Planck Society, Berlin, Germany
2. Current address: Warsaw University of Technology, Faculty of Physics, Warsaw, Poland
3. MPI for the Structure and Dynamics of Matter, Hamburg, Germany
* marcin.krynski@pw.edu.pl

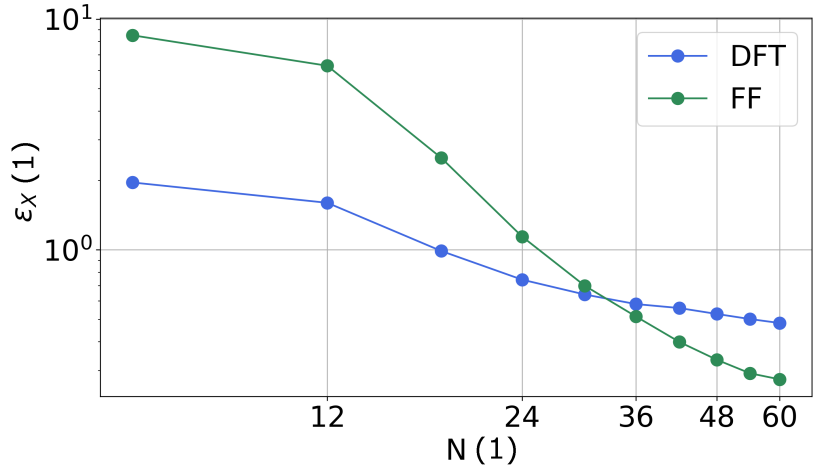
Supplementary Figures



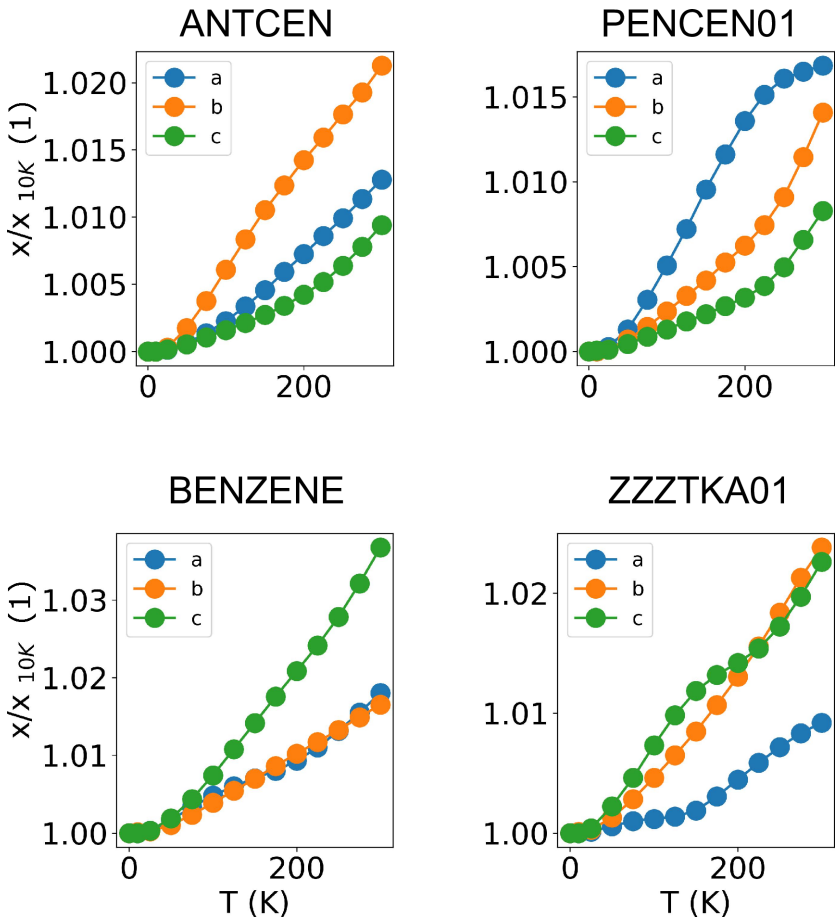
Supplementary Figure 1: Convergence of the $I^{\Theta}(N_m)$ calculated for all potential training set candidates. Colours from yellow to brown marks sets of low to high $I^{\Theta}(N_m)$ convergence. The red colour is marking the selected training set.



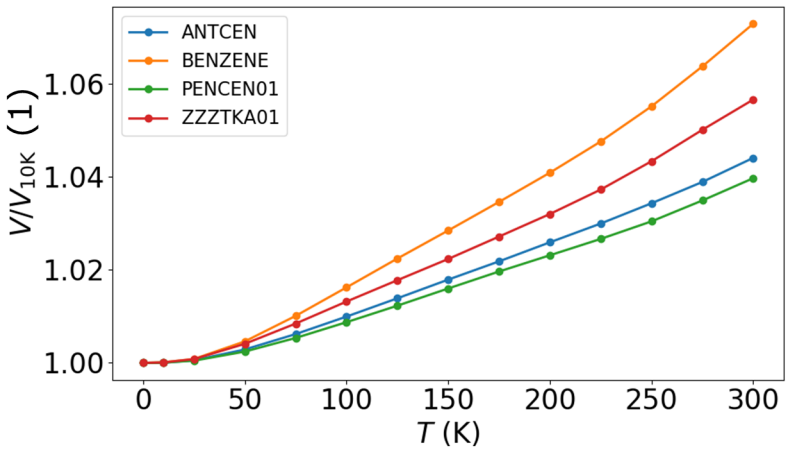
Supplementary Figure 2: Sorted vector S_a calculated for each of discussed sets: validation, training and the entire set. Each dot represent one molecular crystal. Presented in the Figure 2 sorted vector S_a for the training set differs greatly from this of the entire set. This is caused by the necessity for the training set to cover proportionally more outliers than those present in the entire set.



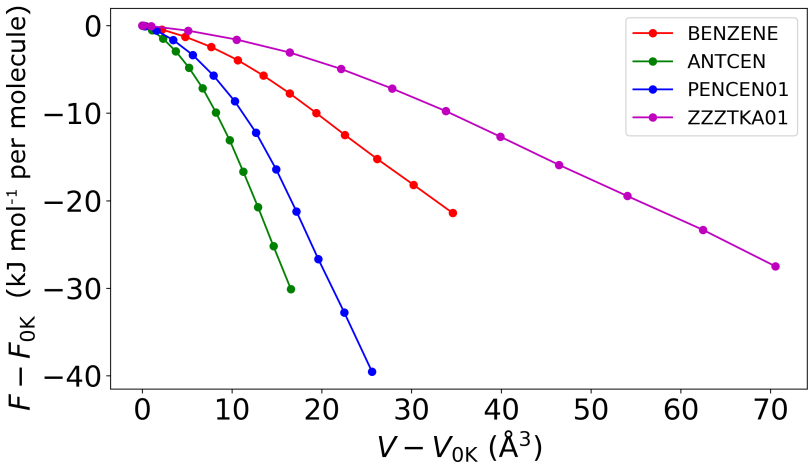
Supplementary Figure 3: Prediction error $\epsilon_X = 100 \times \sqrt{\frac{\frac{1}{N_X}\sum_i^{N_X}(F'(X_i)-F(X_i))^2}{\frac{1}{N_X-1}\sum_j^{N_X}\left(\bar{F}-F(X_i)\right)^2}}$ (where $\boldsymbol{X}$ is the subset of N_X crystal structures for which the prediction is performed) calculated for the validation and the training set for both, *ab initio* and classical models at 300K.



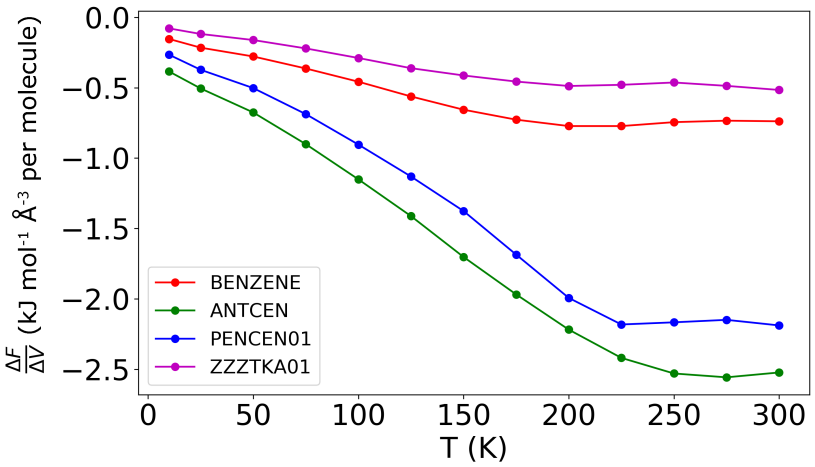
Supplementary Figure 4: Temperature evolution of the relative lattice parameters calculated for four investigated molecular crystals.



Supplementary Figure 5: Temperature evolution of the relative volume calculated for four investigated molecular crystals.



Supplementary Figure 6: Relative free energy as a function of relative volume calculated within 0-300K range.



Supplementary Figure 7: Temperature variation of the the thermal pressure $P_{th}(T) = \frac{\Delta F(T)}{\Delta V(T)}$.

Supplementary Tables

Supplementary Table 1: Results of the GPR hyper-parameters optimisation together with the mean absolute error of the harmonic free-energy calculated for all discussed descriptors based on both, *ab initio* and classical model. All presented values were obtained based on the entire training set ($N_m = 60$).

descriptor	σ	l	σ_ϵ	$F_{MAE}^{N_m=60}(kJ/mol/atom)$
SOAP_{DFT}	0.077	72	0.03	0.038
MBTR_{DFT}	0.077	414	0.03	0.040
ACSF_{DFT}	0.074	18	0.03	0.063
SOAP_{FF}	0.097	22	0.03	0.020
MBTR_{FF}	0.097	157	0.01	0.033
ACSF_{FF}	0.096	42	0.04	0.072

Supplementary Table 2: Lattice energy and free energy at 0K, 100K, 200K and 300K for chosen crystals structures of different families and polymorphs, calculated with respect to the structure showing the lowest free energy at 300K within a specific category. Structures identifiers are presented according to the CCDC data base format. All values are presented in $(kJ/mol/molecule)$.

CCDC id.	E_{lat}	F_{0K}	F_{100K}	F_{200K}	F_{300K}
ANTCEN	0.63	0.64	0.56	0.42	0.29
ANTCEN01	-1.10	-0.55	-0.15	0.53	1.27

Continued on next page

Supplementary Table 2 – *Continued from previous page*

CCDC id.	E_{lat}	F_{0K}	F_{100K}	F_{200K}	F_{300K}
ANTCEN07	0.27	0.28	0.26	0.23	0.20
ANTCEN08	-1.17	-0.57	-0.14	0.62	1.43
ANTCEN09	-1.20	-0.62	-0.16	0.62	1.47
ANTCEN10	-1.02	-0.57	-0.21	0.41	1.08
ANTCEN11	-0.81	-0.47	-0.18	0.30	0.82
ANTCEN12	-0.50	-0.28	-0.09	0.23	0.57
ANTCEN13	-0.12	-0.00	0.09	0.24	0.41
ANTCEN14	0.34	0.36	0.35	0.32	0.30
ANTCEN17	0.42	0.43	0.41	0.36	0.31
ANTCEN19	0.00	0.00	0.00	0.00	0.00
ANTCEN20	-1.46	-0.86	-0.40	0.39	1.24
ANTCEN21	-1.39	-0.87	-0.45	0.26	1.03
ANTCEN22	-1.48	-0.84	-0.35	0.48	1.39
ANTCEN23	-1.53	-0.97	-0.48	0.37	1.28
BENZEN	1.19	0.40	0.16	0.01	0.29
BENZEN01	0.22	0.00	0.05	0.38	1.15
BENZEN03	4.97	6.69	7.57	9.36	11.65
BENZEN04	4.97	6.69	7.57	9.36	11.65
BENZEN11	0.02	0.89	1.34	2.41	4.00
BENZEN12	1.17	2.97	3.70	5.31	7.51
BENZEN13	-0.08	0.47	0.81	1.67	3.02
BENZEN15	0.76	0.19	0.06	0.10	0.56
BENZEN16	2.66	3.08	3.67	4.87	6.45
BENZEN17	2.83	3.37	3.99	5.25	6.91
BENZEN18	1.38	0.50	0.22	0.01	0.21
BENZEN19	0.08	0.02	0.14	0.59	1.49
BENZEN20	0.26	0.02	0.05	0.36	1.11
BENZEN25	1.38	0.51	0.23	0.01	0.21
BENZEN26	2.50	1.27	0.77	0.19	0.03
DUCKOB04	0.00	0.00	0.00	0.00	0.00
DUCKOB05	2.53	9.30	10.87	13.97	17.31
DUCKOB06	2.93	9.84	11.43	14.59	17.99
DUCKOB07	7.38	15.62	17.38	20.99	24.93
DUCKOB08	13.60	23.36	25.27	29.29	33.70
DUCKOB09	19.90	31.08	33.09	37.39	42.14
HEPTAN01	0.00	0.00	0.00	0.00	0.00
HEPTAN03	-0.44	0.16	0.29	0.55	0.83
ZZZDKE01	0.00	0.00	0.00	0.00	0.00
ZZZDKE02	-0.58	-0.31	-0.14	0.17	0.51
NAPHTA04	-1.80	-0.86	-0.40	0.39	1.26
NAPHTA12	-1.88	-0.57	0.14	1.36	2.70
NAPHTA15	-1.91	-0.88	-0.37	0.51	1.48
NAPHTA17	-1.71	-0.97	-0.57	0.10	0.83

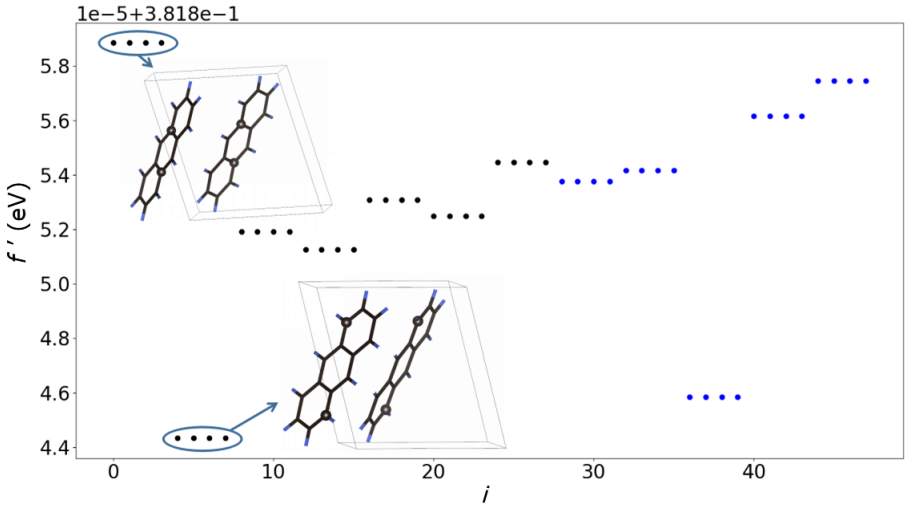
Continued on next page

Supplementary Table 2 – *Continued from previous page*

CCDC id.	E_{lat}	F_{0K}	F_{100K}	F_{200K}	F_{300K}
NAPHTA18	-1.46	-0.92	-0.62	-0.11	0.44
NAPHTA23	-1.93	-0.75	-0.20	0.77	1.83
NAPHTA24	-1.92	-0.77	-0.22	0.74	1.79
NAPHTA36	0.00	0.00	0.00	0.00	0.00
PENCEN	-0.11	0.19	0.41	0.84	1.35
PENCEN01	-2.14	-1.10	-0.55	0.50	1.71
PENCEN05	3.33	3.65	3.27	2.70	2.17
PENCEN10	0.00	0.00	0.00	0.00	0.00
JAYDUI	0.00	0.00	0.00	0.00	0.00
JAYDUI01	7.20	12.32	13.62	16.24	19.05
PYRENE	0.02	0.02	0.02	0.01	0.00
PYRENE01	0.00	0.00	0.00	0.00	0.00
PYRENE02	0.02	0.02	0.02	0.01	0.00
PYRENE03	-1.51	-1.30	-0.91	-0.30	0.32
PYRENE07	-1.08	-1.30	-0.79	0.05	0.92
PYRENE08	1.69	1.66	2.19	3.07	3.89
PYRENE09	-0.31	0.80	1.83	3.45	5.08
PYRENE10	-1.42	-1.17	-0.71	0.02	0.77
ZZZTKA01	-0.40	-0.03	0.12	0.36	0.61
ZZZTKA02	0.00	0.00	0.00	0.00	0.00
TETCEN	0.00	0.00	0.00	0.00	0.00
TETCEN01	-1.14	-0.69	-0.43	0.02	0.51
TETCEN03	2.76	2.81	2.33	1.55	0.77

Supplementary Note 1

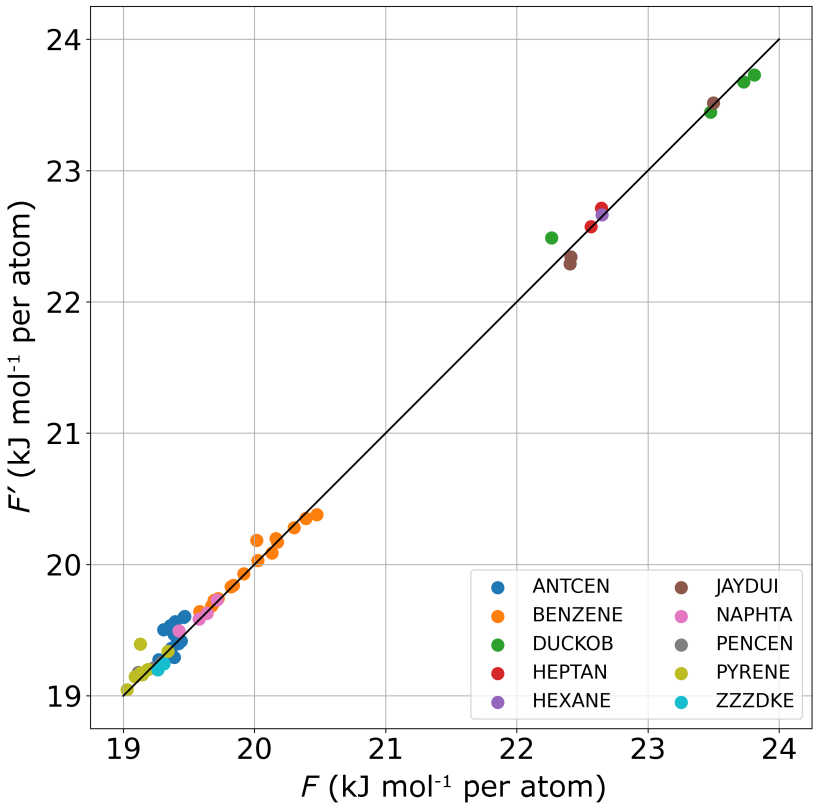
Atom-wise contributions to the free energy can be directly accessed in the presented ML method by analyzing $\sum_{j=1}^{N_{\text{ae}}}(\mathbf{C}^{*\text{T}})_{ij}\alpha_j$ in Eq. 9 of the manuscript. Those values are not supposed to have a physical meaning but can give useful insights. We have analyzed those for all crystals of the validation set and show in Figure 8 these contributions for the ANTCEN19 crystal (others showed similar results). Even if there is a clear grouping of similar atoms with respect to their free energy contribution, the energy difference between atoms is not large. Moreover, there is very little difference between the free energy contribution of carbon and hydrogen atoms. This observation could be related to the fact that the free energy is a strongly non-local, extensive property.



Supplementary Figure 8: Atomic contributions to the free energy obtained during prediction process for ANTCEN19 samples based on the DFT data set at 300K. Black and blue colours represent carbon and hydrogen atoms, respectively. Two inset structures show the first eight carbon atoms and free energy contributions associated with them.

Supplementary Note 2

We have analyzed whether the force-field (FF) structures could be used to predict DFT free energies. In order to obtain an upper limit for the errors that such study would yield, we performed additional calculations of DFT free energies using geometries obtained by relaxations with the AIREBO force-field. We have used the same training (all 60 structures), validation sets and SOAP descriptors as in the manuscript. We performed a new optimization of the hyperparameters of the GPR model with the same approach as presented in the manuscript and obtained values: σ=0.079, l=18 and σ_ϵ=0.04. The predictions were performed at 300K. Figure 9 shows a correlation between predicted and calculated free energies. We have obtained a mean absolute deviation of 0.07 kJ/mol/atom – almost a factor 2 higher when compared to the 0.04 kJ/mol/atom when DFT structures were used. It is expected that once more accurate potential is used, resulting in structures closer to those obtained in DFT, a higher prediction accuracy could be achieved. This method thus shows some promise that the DFT free energies could be directly from FF structures and potentially free energies.



Supplementary Figure 9: Correlation between predicted F' and calculated F free energies at 300 K, with a model trained on force-field structures, but with DFT free-energy predictions. Different crystal families are represented by different colors.