Supporting Information:

A Generic Machine Learning Algorithm for the Prediction of Gas Adsorption in Nanoporous Materials

George S. Fanourgakis,[*,†] Konstantinos Gkagkas,[‡] Emmanuel Tylianakis,[¶] and George Froudakis[*,†]

†*Department of Chemistry, University of Crete, Voutes Campus, GR-70013 Heraklion, Crete, Greece*

‡*Advanced Technology Division, Toyota Motor Europe NV/SA, Technical Center, Hoge Wei 33B, 1930 Zaventem, Belgium*

¶*Department of Materials Science and Technology, University of Crete, Voutes Campus, GR-70013, Heraklion, Crete, Greece*

E-mail: fanourg@uoc.gr; frudakis@uoc.gr

Statistical Metrics for the evaluation of the ML results

The R-squared (R^2), mean absolute error (MAE) and the root mean squared error (RMSE) statistical metrics were used for the evaluation of the performance of the ML algorithms. They were computed using the following expressions

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(A_i - F_i)^2}{\sum_{i=1}^{n}(A_i - F)^2} \tag{1}$$

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|A_i - F_i| \tag{2}$$

$$\text{RMSE} = \sqrt{\frac{\sum_{i=1}^{n}(A_i - F_i)^2}{n}} \tag{3}$$

where F_i are the reference values (i.e. the various gases uptake capacities of CoRE MOFs obtained from the GCMC simulations), F the corresponding average value, while A_i are the predictions of the ML algorithm for each MOF.

Histograms of the gas adsorption

Distributions of the volumetric-based adsorption capacities of the 2932 CoRE MOFs examined in the present study.

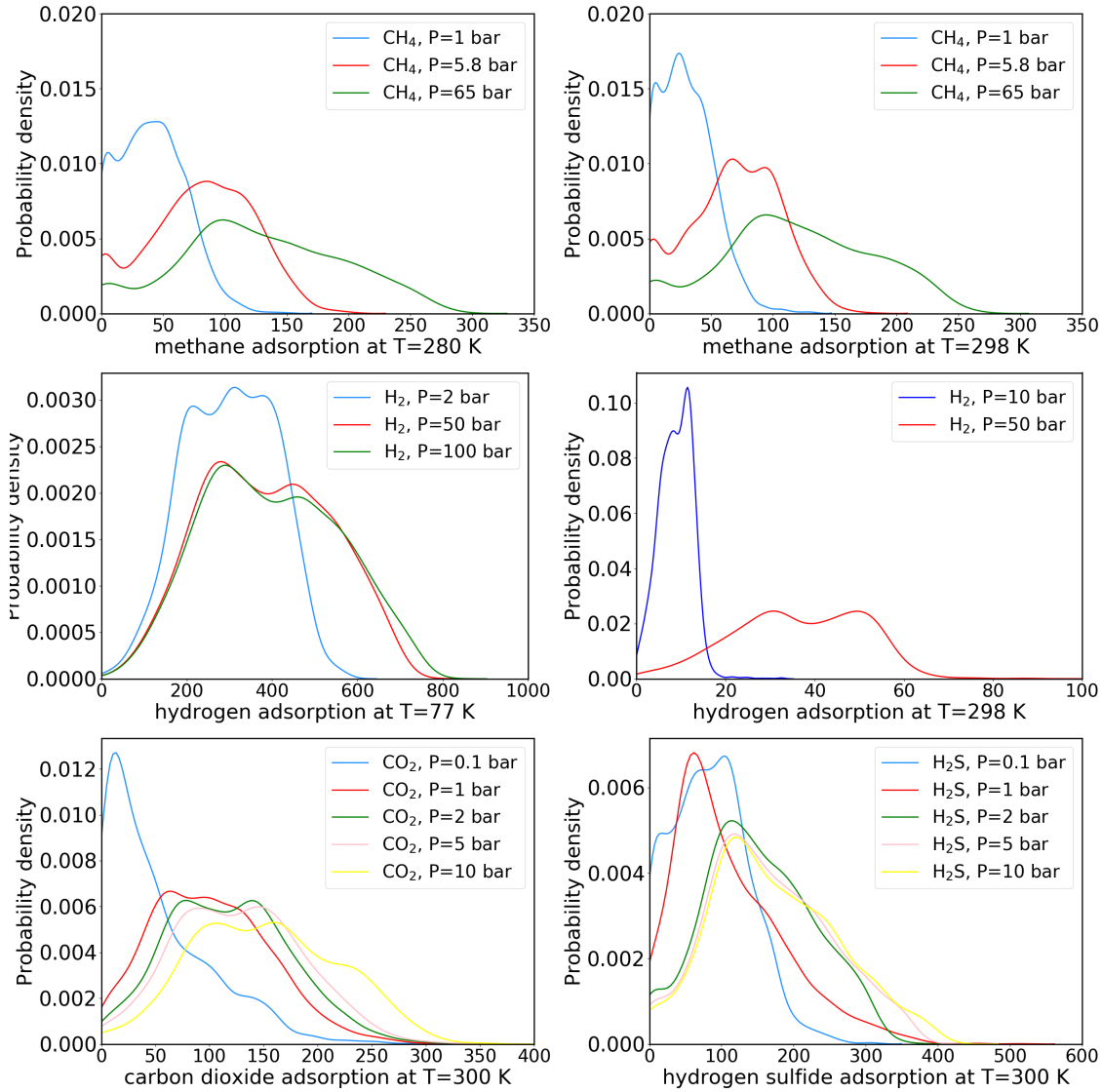

Figure S1: Histograms of the methane, molecular hydrogen, carbon dioxide, and hydrogen sulfide volumetric-based adsorption capacities of the 2932 CoRE MOFs as computed by GCMC simulations. The temperature and pressure of each simulation are indicated in the graphs.

Machine Learning descriptors

In Table S1 information about the standard structural descriptors and the Vprobes and Dprobes is provided. In figure S2 and S3 the corresponding distributions are graphically illustrated. The frequency of appearance of various types of atoms in the 2932 CoRE MOFs examined is shown in figure S4.

Table S1: Descriptors used in the present study in the ML methods. The minimum, maximum and mean value of each descriptor for the 2932 CoRE MOFs are tabulated. The three different types of ML descriptors used in the present study are the standard structural features alone or combined with the Vprobes or the Dprobes.

Descriptors	min	max	mean
Standard Structural Features			
void fraction	0.0	0.89	0.40
pore volume [cm^3 g^{-1}]	0.07	2.09	0.42
density [g cm^{-3}]	0.41	3.91	1.44
grav. surface area [m^2 g^{-1}]	0.0	5132	648.1
pore-limiting diameter [Å]	1.8	23.7	4.33
largest cavity diameter [Å]	2.47	23.9	5.88
Vprobes			
Vprobe-1 (σ =2.5 Å)	0.00	4.7	1.06
Vprobe-2 (σ =3.0 Å)	0.00	11.3	1.45
Vprobe-3 (σ =3.5 Å)	0.00	32.6	2.18
Vprobe-4 (σ =4.0 Å)	0.00	95,0	3.56
Dprobes			
Dprobe-1 (σ =2.5 Å)	0.00	295	4.03
Dprobe-2 (σ =3.0 Å)	0.00	43	2.17
Dprobe-3 (σ =3.5 Å)	0.00	61	2.74
Dprobe-4 (σ =4.0 Å)	0.00	168	8.90

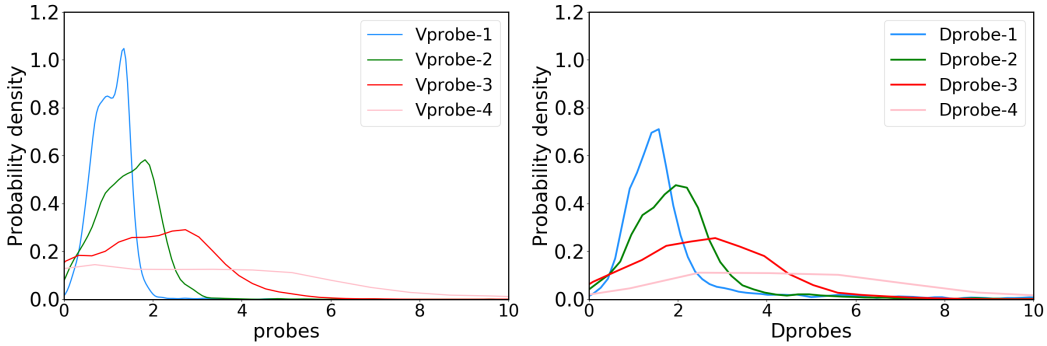

Figure S2: Histograms of the Vprobes and Dprobes for the 2932 CoRE MOFs used in the present study.

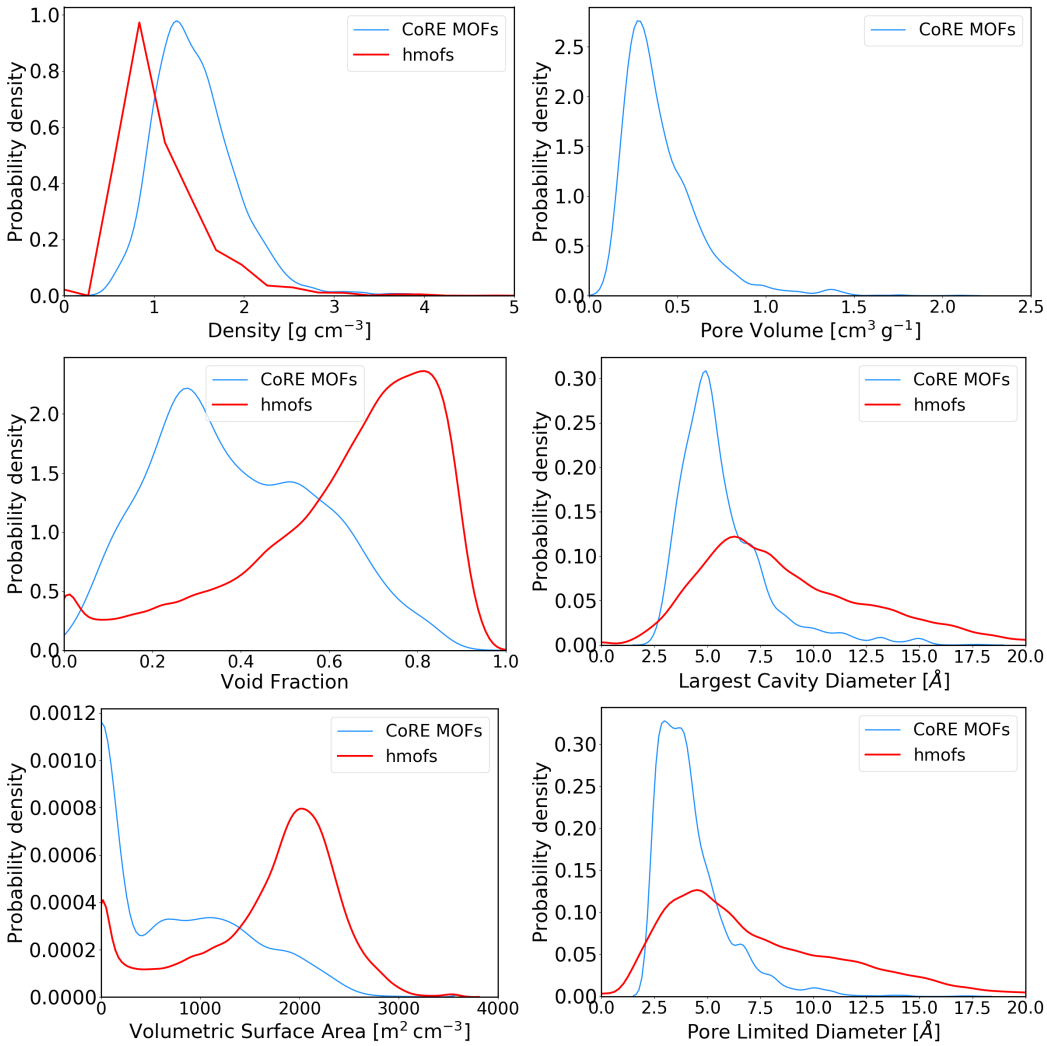

Figure S3: Histograms of the standard structural features for the 2932 CoRE MOFs (blue lines) used in the present study: a) density, b) void fraction, c) pore volume, d) pore-limited diameter, e) gravimetric surface area, f) largest cavity diameter. For comparison, the corresponding quantities for the $\sim 137,000$ hypothetical MOFs (hmofs) are also shown (red lines).

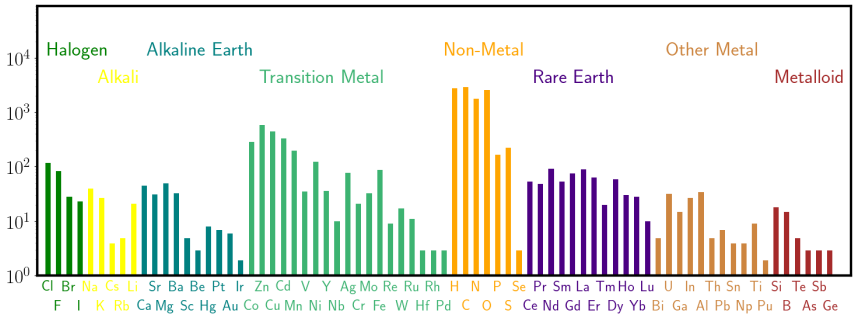

Figure S4: Number of appearances of various chemical species in the 2932 CoRE MOFs examined in this study, demonstrating the chemical diversity of the database. Notice the logarithmic scale in the y-axis.

Evaluation of ML results

Cross validation results

Table S2: Evaluation of the ML performance for different sets of descriptors at various temperatures and pressures, based on the mean absolute error (MAE). The average adsorption capacity of the 2932 MOFs is given in the fourth column. The units are v_{STP}/v.

gas	T [K]	P [bar]	AVER	struc.	struc.+ Vprobes	struc.+ Qprobes	struc.+ Dprobes	struc. + Vprobes + Dprobes
CH_4	280	1	40.5	14.3	6.93	12.8	8.13	6.84
CH_4	280	5.8	80.9	17.9	11	16.5	12.6	10.8
CH_4	280	65	130	18.8	14.9	16.9	15.8	14.7
CH_4	298	1	30.1	11.9	5.08	10.4	6.38	5.04
CH_4	298	5.8	68.3	16.4	9.12	14.8	10.5	9.21
CH_4	298	65	121	17.8	13.6	16.2	14.9	13.6
CO_2	300	0.1	50.2	34	33.5	26.9	19.9	19.7
CO_2	300	1	98.5	33.2	32.1	27.8	21.3	20.8
CO_2	300	2	116	32	30.6	26.3	21.2	20.6
CO_2	300	5	128	30.8	29.4	25.5	21.1	20.6
CO_2	300	10	152	28.7	25.5	23.5	19.7	19.1
H_2S	300	0.1	83	34.6	31.6	29.3	22.8	21.8
H_2S	300	1	137	33.2	28.9	28.1	24.1	23.4
H_2S	300	2	153	31.8	27.4	27.3	23.8	22.9
H_2S	300	5	168	29.5	25.3	26	23.2	22.8
H_2S	300	10	175	29.1	24.4	25.3	22.7	22
H_2	77	2	300	47.8	40.2	41	32.4	31.6
H_2	77	50	390	49.5	39.5	42.2	36.8	34.7
H_2	77	100	402	50.8	40.1	44	38.3	36.3
H_2	298	10	8.9	1.43	1.09	1.18	0.839	0.805
H_2	298	50	36.6	4.77	3.63	3.88	2.66	2.52

Table S3: Similar to Table S2 for the root mean square error (RMSE).

gas	T [K]	P [bar]	AVER	struc.	struc.+ Vprobes	struc.+ Qprobes	struc.+ Dprobes	struc. + Vprobes + Dprobes
CH_4	280	1	40.5	9.54	4.01	8.29	5.13	4.02
CH_4	280	5.8	80.9	11.6	7.19	10.6	8.25	7.13
CH_4	280	65	130	11.8	9.67	10.9	10.3	9.64
CH_4	298	1	30.1	7.87	2.81	6.65	3.83	2.81
CH_4	298	5.8	68.3	10.7	5.85	9.51	6.8	5.9
CH_4	298	65	121	11.2	8.92	10.5	9.66	8.9
CO_2	300	0.1	50.2	22.8	22.3	17.1	12.1	11.9
CO_2	300	1	98.5	22.7	21.6	18.7	14.2	13.8
CO_2	300	2	116	21.7	20.2	17.6	14.1	13.8
CO_2	300	5	128	20.8	19.5	17.2	14.3	13.9
CO_2	300	10	152	18.9	16.6	15.4	13	12.6
H_2S	300	0.1	83	23.3	21.3	19.5	15.3	14.6
H_2S	300	1	137	21.7	19	18.6	16.3	15.7
H_2S	300	2	153	20.5	17.6	17.7	15.7	15.1
H_2S	300	5	168	19.2	16.5	17	15.2	14.7
H_2S	300	10	175	19.4	16.1	17	15.2	14.5
H_2	77	2	300	32.2	26.8	27.5	21.9	21.3
H_2	77	50	390	32.2	25.5	28	23.6	22.8
H_2	77	100	402	32.7	25.5	28.8	24.5	23
H_2	298	10	8.9	0.913	0.696	0.8	0.541	0.512
H_2	298	50	36.6	2.97	2.33	2.6	1.73	1.64

Variation of statistical metrics with training set size

The variation of the R^2, MAE and RMSE with the training set size of the ML predictions for the CO_2, H_2, and H_2S adsorption capacities of the 2932 CoRE MOFs is illustrated at various thermodynamic conditions in the figures S5, S6 and S7, respectively.

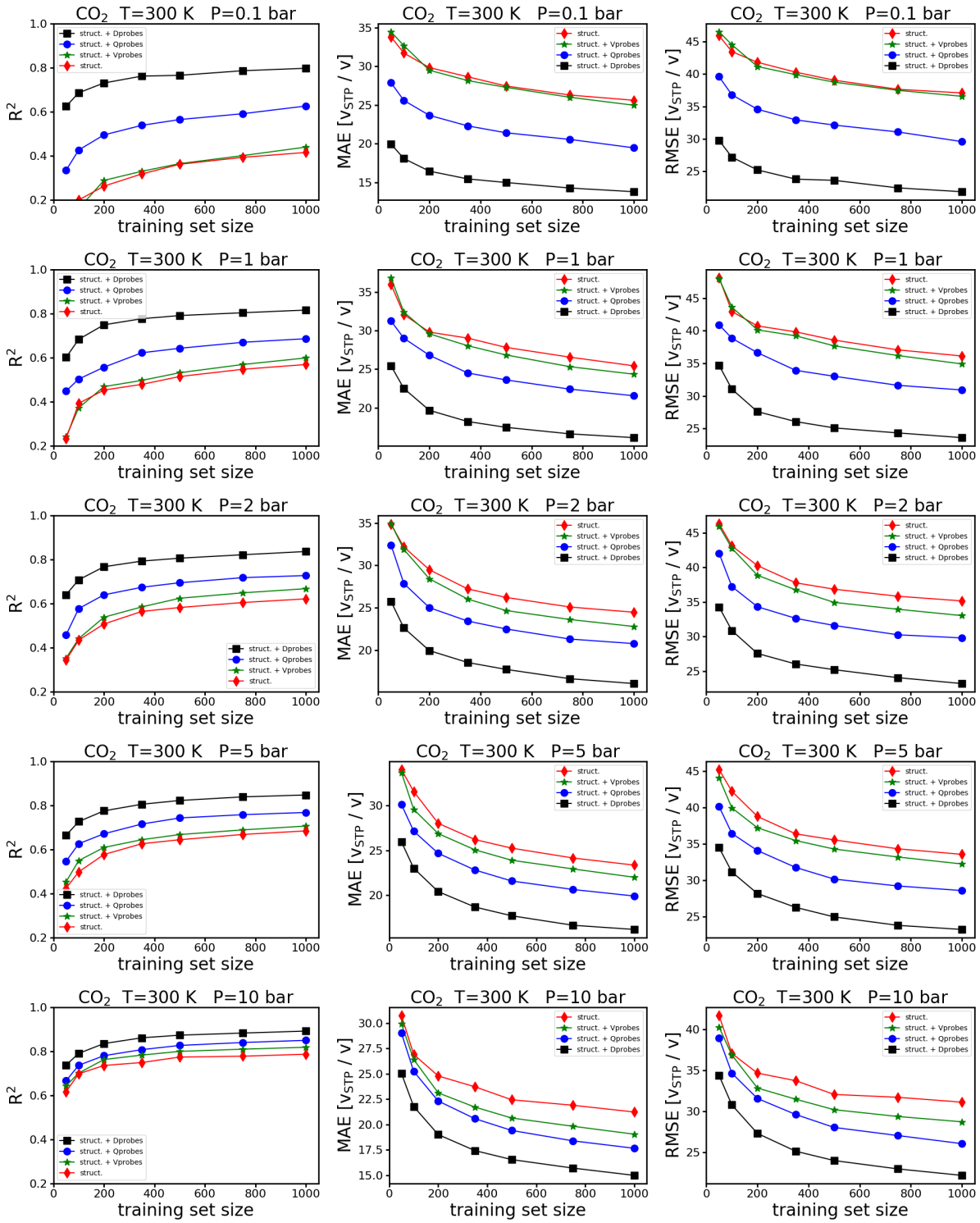

Figure S5: Variation of the R^2, MAE and RMSE, as a function of the training set size for the carbon dioxide. In each graph with different colors are shown different sets of descriptors (see text for details).

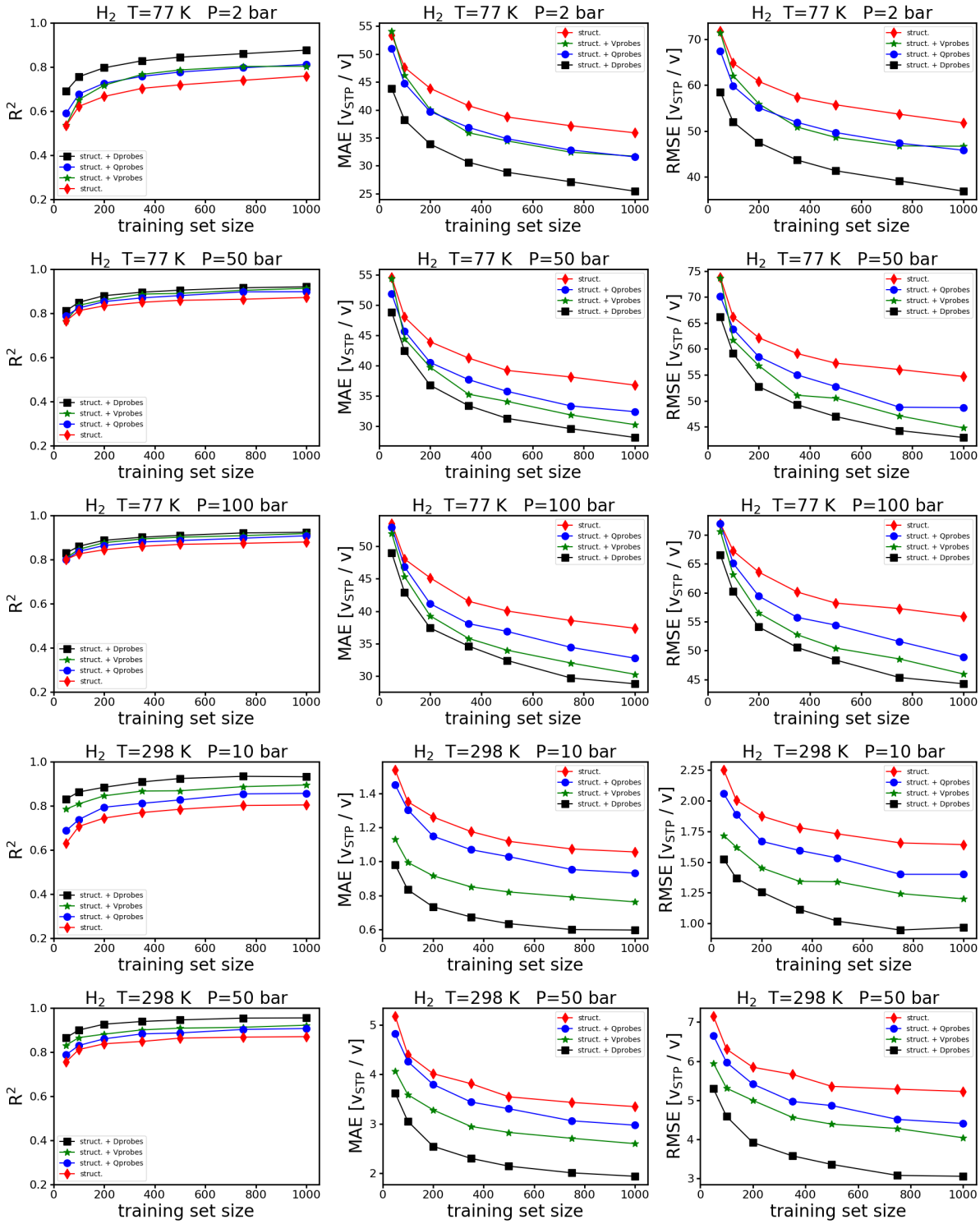

Figure S6: Same as in figure S5 for molecular hydrogen.

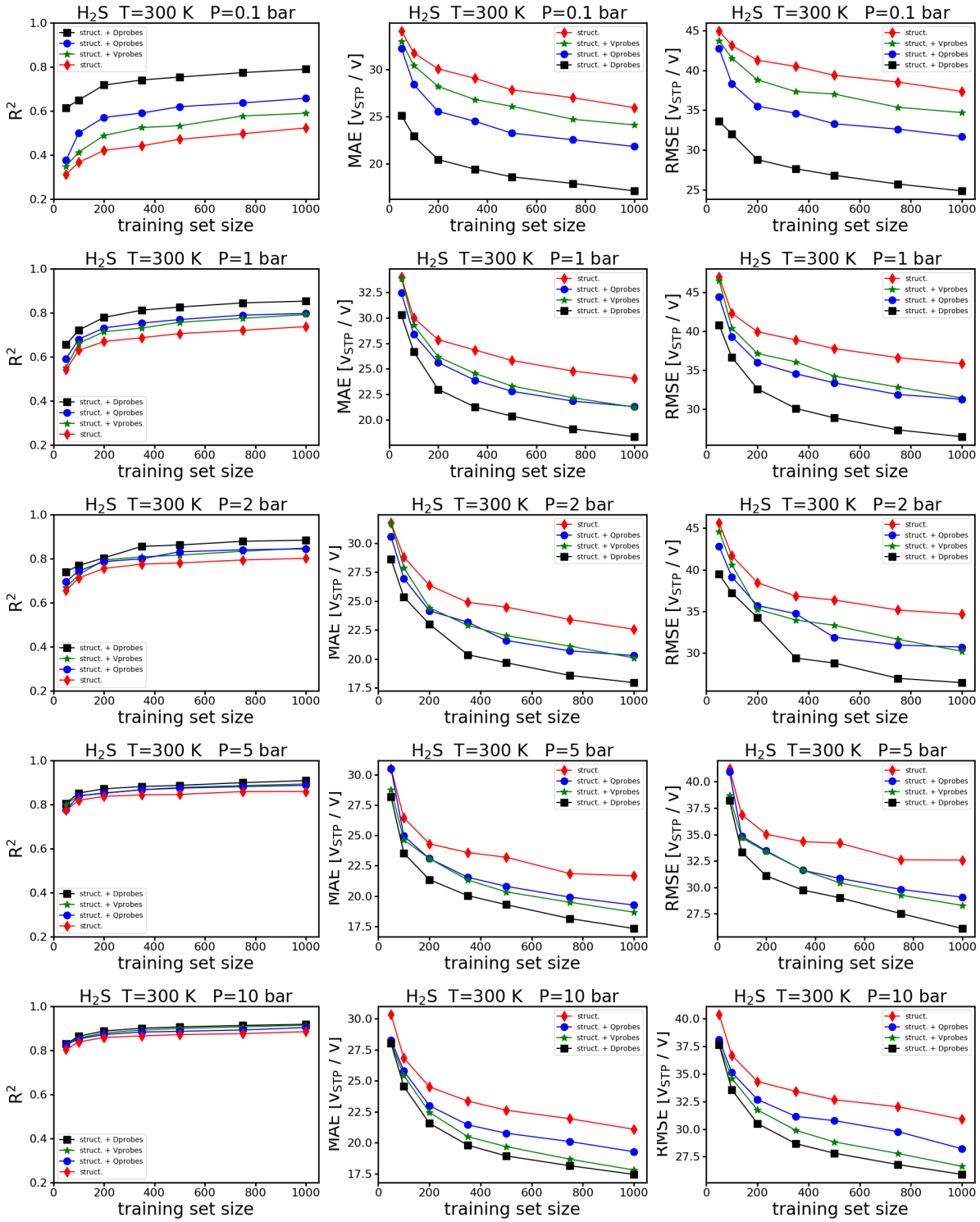

Figure S7: Same as in figure S5 for hydrogen sulfide.

Sensitivity to the probe particles parameters

In analogy to the sensitivity of the ML results to the Dprobe parameters shown in Table 3 using the R^2, the corresponding MAE and RMSE values are tabulated in Tables S4 and S5, respectively.

Table S4: The MAE of the ML predictions for the 2932 CoRE MOFs computed using different sets of the Dprobe parameters are compared to the reference run.

gas	T [K]	P [bar]	Ref.	T [K] 250	T [K] 350	q^P [eu] 0.08	q^P [eu] 0.16	d [Å] 0.3	d [Å] 0.9	ε/k_B [K] 30	ε/k_B [K] 70
CO_2	300	0.1	12.1	12.2	12.3	15.3	11.9	18.7	12	12.3	12.7
CO_2	300	1	14.2	14.2	14.7	16.5	14	18.9	14.3	14.1	14.6
CO_2	300	2	14.1	13.9	14.4	16	14	17.9	14.1	13.8	14.7
CO_2	300	5	14.3	14.3	14.4	15.7	14.3	17.3	14.4	14.1	14.7
CO_2	300	10	13	13.1	13.1	14.1	13	15.4	13.1	13	13.5
H_2S	300	0.1	15.3	15.2	15.4	17.1	14.8	18.8	14.7	15.2	15.8
H_2S	300	1	16.3	16.2	16.3	16.8	16.3	17.7	16.4	16.1	16.2
H_2S	300	2	15.7	15.6	15.8	16.2	15.8	16.9	15.9	16	15.7
H_2S	300	5	15.2	15.1	15.2	15.4	15.4	16.1	15.4	15.3	15.1
H_2S	300	10	15.2	15.1	15.2	15.4	15.3	15.9	15.4	15.4	15.1
H_2	77	2	21.9	22.5	22.4	22	23.4	22.4	23.6	22.7	22.2
H_2	77	50	23.6	24.3	23.8	24.1	24.6	25	24.9	24.3	23.8
H_2	77	100	24.5	24.9	24.2	24.5	25.3	25.3	25.5	25	24.4
H_2	298	10	0.541	0.553	0.53	0.511	0.604	0.527	0.627	0.601	0.524
H_2	298	50	1.73	1.78	1.69	1.63	1.95	1.7	2.02	1.9	1.7

Table S5: Same as Table S4 for the RMSE.

gas	T [K]	P [bar]	Ref.	T [K] 250	T [K] 350	q^P [eu] 0.08	q^P [eu] 0.16	d [Å] 0.3	d [Å] 0.9	ε/k_B [K] 30	ε/k_B [K] 70
CO_2	300	0.1	19.9	20.2	19.7	23.1	19.6	28	19.9	20.1	20.6
CO_2	300	1	21.3	21.7	22	24.3	21.2	27.7	21.7	21	22
CO_2	300	2	21.2	20.8	21	23.7	20.9	26.1	21.4	20.5	21.8
CO_2	300	5	21.1	21.1	21	23.1	21.2	25.4	21.4	20.5	21.6
CO_2	300	10	19.7	19.9	19.6	21.1	19.6	23.2	19.9	19.3	20.3
H_2S	300	0.1	22.8	23	22.6	25.2	22.3	27.3	22.4	22.8	23.4
H_2S	300	1	24.1	24	24	25.2	24	26.5	24.4	23.6	24.3
H_2S	300	2	23.8	23.4	23.6	24.5	23.6	25.6	23.9	24	23.7
H_2S	300	5	23.2	23.3	23.1	23.6	23.6	24.5	23.5	23.1	23.2
H_2S	300	10	22.7	22.9	22.8	23	23	23.6	23.1	23	23
H_2	77	2	32.4	33.5	33	32.6	34.7	33.3	35.3	33.4	33.7
H_2	77	50	36.8	37.7	37.1	37.2	38.6	37.7	38.6	37.3	36.8
H_2	77	100	38.3	39.2	38	38.1	39.6	38.8	40	39.1	38.2
H_2	298	10	0.839	0.888	0.823	0.829	0.953	0.81	0.971	0.943	0.821
H_2	298	50	2.66	2.75	2.52	2.59	2.95	2.64	3.03	2.81	2.73