This version of the ESI published 07/08/2020 replaces the previous version published 13/07/2018.

Supplementary Information

Machine learning meets volcano plots: Computational discovery of cross-coupling catalysts†

Benjamin Meyer,[a,c] Boodsarin Sawatlon,[a,c] Stefan Heinen,[b,c] O. Anatole von Lilienfeld,*[b,c] and Clémence Corminboeuf*[a,c]

[a] *Laboratory for Computational Molecular Design, Institute of Chemical Sciences and Engineering, École Polytechnique Fédérale de Lausanne (EPFL), CH-1015 Lausanne, Switzerland* [b] *Institute of Physical Chemistry, Department of Chemistry, University of Basel, Klingelbergstrasse 80, CH-4056 Basel, Switzerland* [c] *National Center for Computational Design and Discovery of Novel Materials (MARVEL), École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland*

Contents

1 Ligands dataset **3**

2 Computational Details **7**
2.1 Complexes generation and optimization 7
2.2 Construction of the volcano plot 8

3 Details on machine learning curves **10**

4 Machine learning predictions **14**
4.1 Coulomb Matrix representation 14
4.1.1 Distribution of the metals 14
4.1.2 Distribution of the ligands 15
4.2 Bag of Bonds representation 18
4.2.1 Distribution of the ligands 18
4.3 SLATM representation 21
4.3.1 Distribution of the metals 21
4.3.2 Distribution of the ligands 22

5 Price of metals and ligands **25**

6 Generated XYZ structures for database construction **28**

7 DFT Optimized XYZ structures and associated binding energies **28**

8 Out-of-sample machine learning predicted binding energies **28**

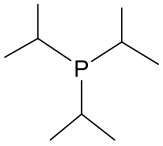
1 Ligands dataset

Chemical structures of 91 ligands in database

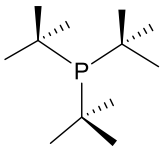
0

1

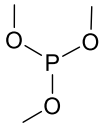
2



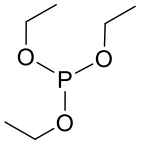
3



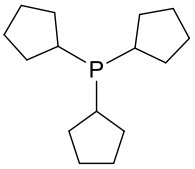
4



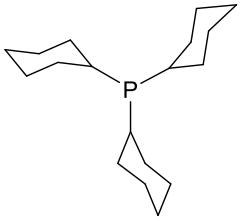
5



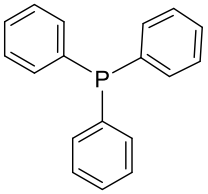
6



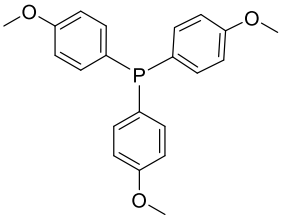
7



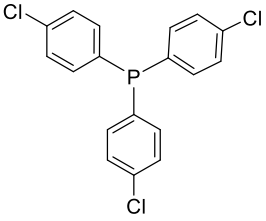
8



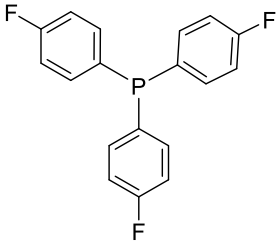
9



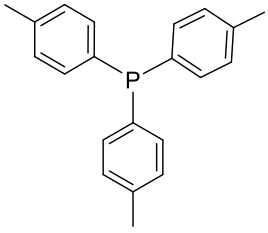
10



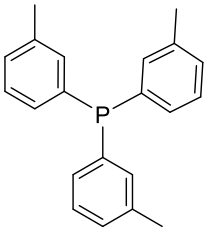
11



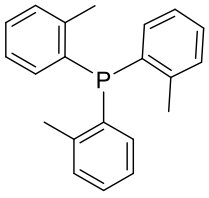
12



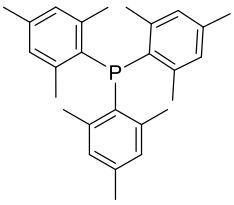
13



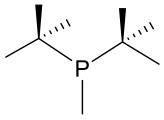
14



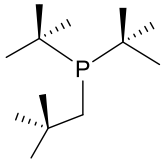
15



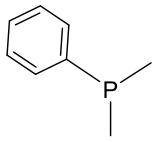
16



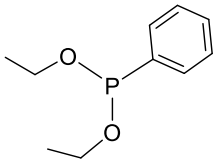
17



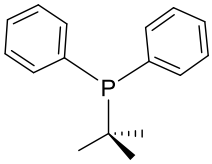
18



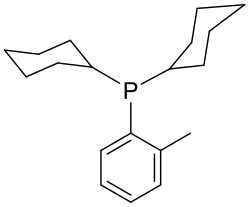
19



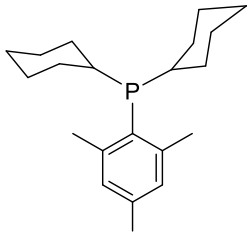
20



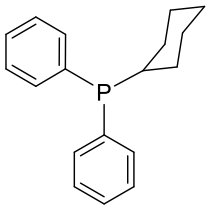
21



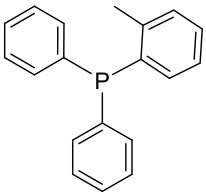
22



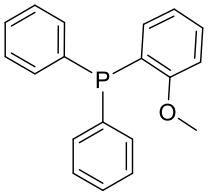
23



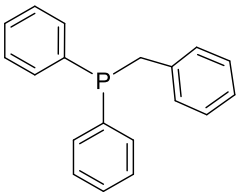
24



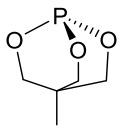
25



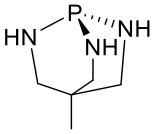
26



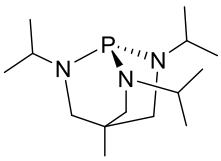
27



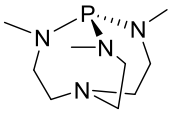
28



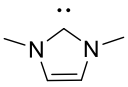
29



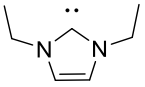
30



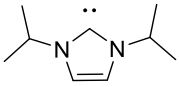
31



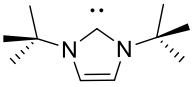
32



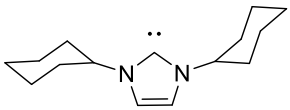
33



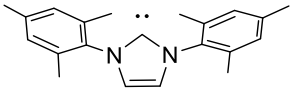
34



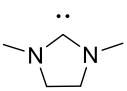
35



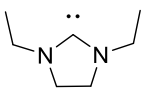
36



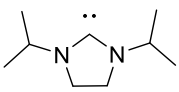
37



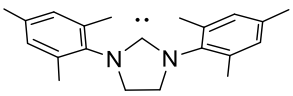
38



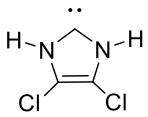
39



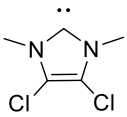
40



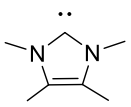
41



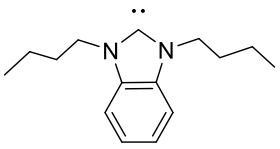
42



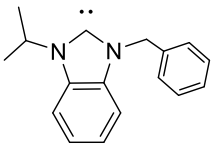
43



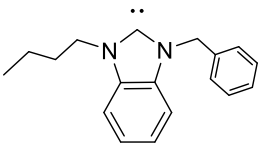
44



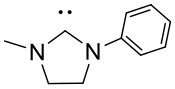
45



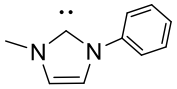
46



47

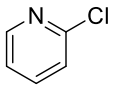


48

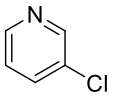


49

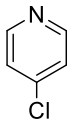
50



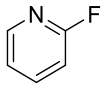
51



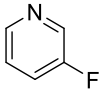
52



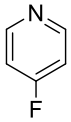
53



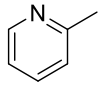
54



55

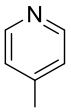


56

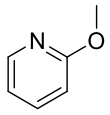


57

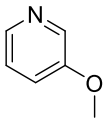
58



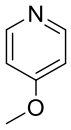
59



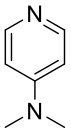
60



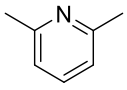
61



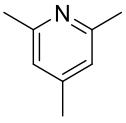
62



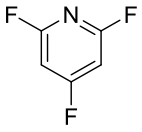
63



64



65

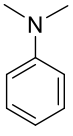


66

67

68

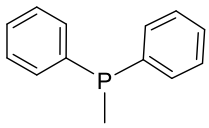
69



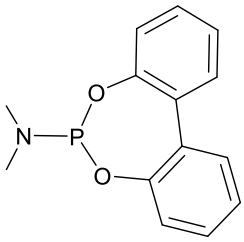
70

C=O

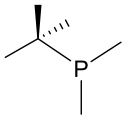
71



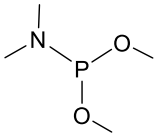
72



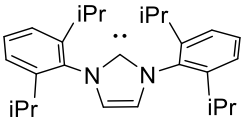
73



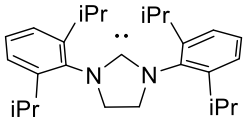
74



75

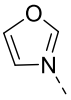


76

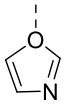


77

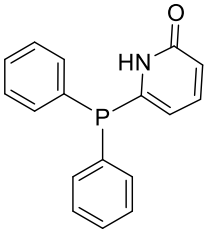
78



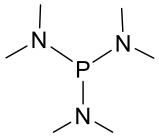
79



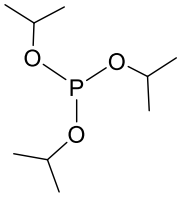
80



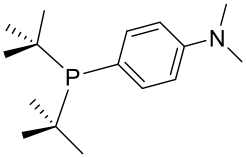
81



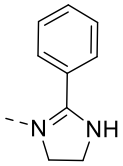
82



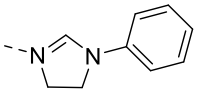
83



84

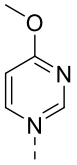


85

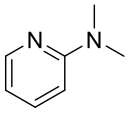


86

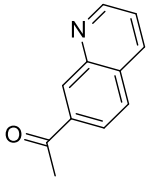
87



88



89



90

2 Computational Details

2.1 Complexes generation and optimization

In the simplified cross-coupling reaction presented in Fig. 1, the catalysts proceed through two 16 electron square planar intermediates (**2-3**) in trans conformation. The two isomerization steps (*cis* → *trans* and *trans* → *cis*) are not considered. To obtain the square planar intermediates in the *trans* conformation, we specifically used the square-planar class symbol SP of the SMILE notation (see top right-hand corner of Fig. 3). SMILES were then converted into Cartesian coordinates through the 3D structure generator of the OpenBabel software using the following series of steps:

- The 3D structure is constructed instantaneously using a combination of simple rules (e.g. sp3 atoms should have four bonds arranged in a tetrahedron) and ring templates (e.g. cyclohexane is shaped like a chair)
- Perform 250 steps of a steepest descent geometry optimization with the MMFF94 force field
- Do 200 iterations of a Weighted Rotor conformational search (optimizing each conformer with 25 steps of a steepest descent)
- Do 250 steps of a conjugate gradient geometry optimization

Each of these steps ensures that the generated 3D structure is likely to be the global minimum energy conformer at the force field level. Starting from this structure, we then further optimized the geometry at B3LYP/3-21G and B3LYP/def2-SVP level for for Ni, Pd, Cu, Ag and for the Pt and Au complexes, respectively.

For seven catalysts falling in the ideal thermodynamic window we then verified that the level of the basis set exploited during the geometrical optimization does not play a prominent role on the final computed binding energies.

Table S1: Comparison between electronic binding energies obtained from different level of theory (def2-TZVP//3-21G,def2-SVP and def2-TZVP) and the energies predicted by the machine learning model (BoB representation)

Metal	Ligand 1 no.	Ligand 2 no.	def2-TZVP//3-21G,def2-SVP	def2-TZVP	ML predictions
Pd	0	74	-26.52	-24.38	-26.56
Pd	15	74	-27.28	-26.82	-31.33
Pd	1	89	-22.64	-20.03	-25.46
Pd	4	72	-27.35	-29.35	-26.64
Pt	52	71	-29.01	-29.28	-28.50
Pt	71	84	-25.41	-25.68	-28.08
Pt	1	84	-27.14	-27.22	-29.02
Pt	17	23	-31.52	-31.67	-28.80

2.2 Construction of the volcano plot

The construction of the molecular volcano plots associated with the catalytic cycle in Figure 1 is based on linear scaling relationships between the relative stability of the catalytic species intermediates and the descriptor variable. In line with our previous work, the energy profiles of 30 catalysts [*i.e.*, six metals (Ni, Pd, Pt, Cu, Ag, and Au) combined with five ligands sets (CO ($\times$ 2), NH_3 ($\times$ 2), PMe_3 ($\times$ 2), acetone ($\times$ 2) and an N-heterocyclic carbene ($\times$ 2)] were determined at the same level of theory as used to compute the descriptor ΔE(Rxn A) in the machine learning training set. Choosing the electronic energy of the oxidative addition step [ΔE(Rxn A)] as the descriptor (see Equation 1), the energies of the other intermediates are then plotted against the descriptor, which gives the linear scaling relationships in Figure S1. The mathematical equations defining the linear scaling relationships are then exploited to establish theoretical reaction energies for each step of the catalytic cycle as a function of the descriptor. The final volcano plot is then obtained by selecting the lowest -ΔE(pds) values amongst all reactions (A-C) for each $\Delta E_{RRS}(3)$ where

$$\Delta E(pds) = max[\Delta E_{Rxn}(A), \Delta E_{Rxn}(B), \Delta E_{Rxn}(C)] \quad (1)$$

The volcano plot (Figure 2) for the chosen cross-coupling reaction is divided into three thermodynamic regions. The left slope, also known as the "strong binding slope" [(I), Figure 2] contains catalysts that bind the reactant too strongly relative to a hypothetical ideal catalyst. Since these catalysts tend to bind intermediates too strongly, reductive elimination (Rxn C) is the potential determining (most difficult) reaction step. Conversely, the right "weak binding" slope [(III), Figure

2] consists of catalysts than bind the reactant too weakly, which makes oxidative addition (Rxn A) potential determining.

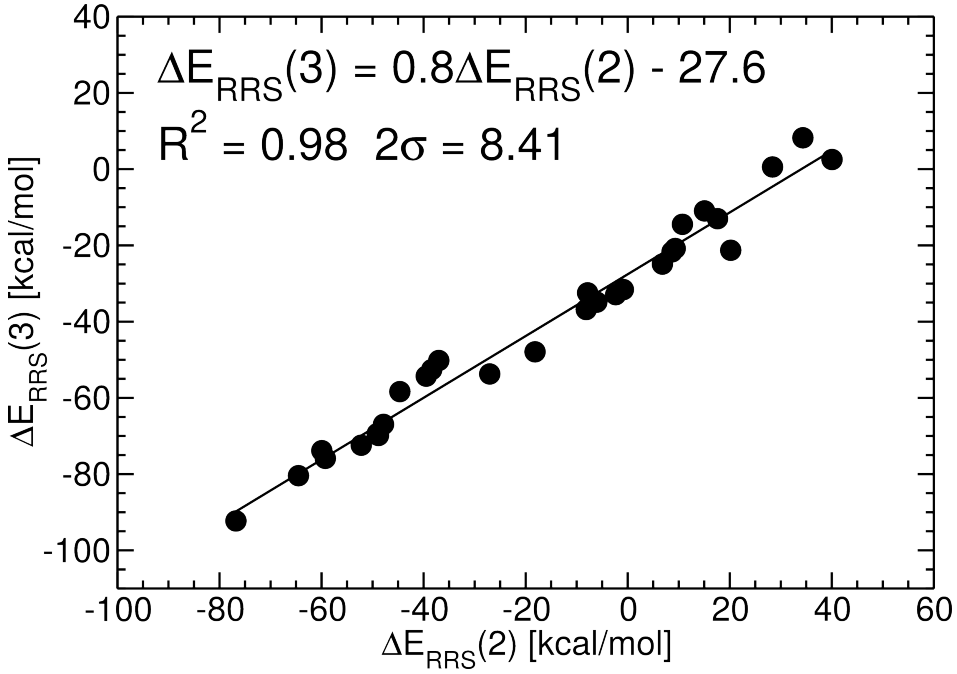


Figure S1: Linear scaling relationship between the [$\Delta E_{RRS}(2)$] and [$\Delta E_{RRS}(3)$] for 30 catalysts.

Catalysts having the most appealing "thermodynamic" profiles should lie on the plateau of the volcano ((II), Figure 2), region where the energies associated with the oxidative addition and reductive elimination are roughly balanced. This targeted plateau region encompasses a narrow binding energy range of only 9.1 kcal/mol between -23.0 and -32.1 kcal/mol.

3 Details on machine learning curves

Table S2: Learning curves detailed data (in terms of Mean Absolute Error ± Standard Deviation for different training set size) for different molecular representations. For the cross validation the data set was split into a training (6354 complexes) and a test set (700 complexes). The hyperparameters (σ's in the gaussian, laplacian kernels) where obtained using a 10-fold cross validation and screening over a range of $\sigma = 0.1 - 104857.6$ using the training set. With the obtained sigmas we used random subsampling cross validation to generate the learning curves using the test set for validation.

Representation	Kernel	Sigma	Cross Validation	Training Set Size	MAE [kcal/mol]
CM (Coulomb Matrix)	Laplacian	13107.2	10	100	11.97 ± 0.40
				200	9.54 ± 0.12
				400	6.49 ± 0.11
				800	5.12 ± 0.09
				1600	4.10 ± 0.06
				3200	3.55 ± 0.01
				6354	3.05 ± 0.00
BoB (Bag of Bonds)	Laplacian	26214.4	10	100	5.58 ± 0.12
				200	4.69 ± 0.08
				400	4.10 ± 0.05
				800	3.62 ± 0.04
				1600	3.33 ± 0.03
				3200	2.96 ± 0.02
				6354	2.73 ± 0.00
SLATM (global)	Gaussian	819.2	10	100	6.22 ± 0.22
				200	5.01 ± 0.11
				400	4.53 ± 0.11
				800	3.81 ± 0.06
				1600	3.38 ± 0.04
				3200	2.96 ± 0.03
				6354	2.61 ± 0.00

Table S3: Learning curves detailed data (in terms of Mean Absolute Error ± Standard Deviation for different training set size) for different type of complexes using the BoB representation.

Metal	Kernel	Sigma	Cross Validation	Training Set Size	MAE [kcal/mol]
Pd	Laplacian	26214.4	10	100	4.65 ± 0.05
				200	4.42 ± 0.14
				400	3.94 ± 0.13
				800	3.45 ± 0.05
				1600	3.15 ± 0.03
				3200	3.00 ± 0.02
				6354	2.81 ± 0.00
Pt	Laplacian	26214.4	10	100	6.49 ± 0.82
				200	4.55 ± 0.23
				400	3.88 ± 0.30
				800	3.24 ± 0.06
				1600	2.53 ± 0.10
				3200	2.11 ± 0.06
				6354	1.81 ± 0.00
Ni	Laplacian	26214.4	10	100	10.37 ± 1.09
				200	6.68 ± 0.40
				400	5.96 ± 0.24
				800	5.54 ± 0.30
				1600	5.40 ± 0.18
				3200	4.35 ± 0.14
				6354	3.74 ± 0.00
Cu	Laplacian	26214.4	10	100	6.97 ± 0.25
				200	5.98 ± 0.20
				400	5.52 ± 0.16
				800	5.00 ± 0.10
				1600	4.90 ± 0.10
				3200	4.24 ± 0.04
				6354	4.04 ± 0.00
Ag	Laplacian	26214.4	10	100	5.65 ± 0.29
				200	4.35 ± 0.13
				400	3.43 ± 0.06
				800	2.90 ± 0.11
				1600	2.70 ± 0.07
				3200	2.30 ± 0.06
				6354	2.08 ± 0.00
Au	Laplacian	26214.4	10	100	3.78 ± 0.10
				200	3.37 ± 0.09
				400	2.75 ± 0.05
				800	2.34 ± 0.05
				1600	1.98 ± 0.05
				3200	1.76 ± 0.03
				6354	1.60 ± 0.00

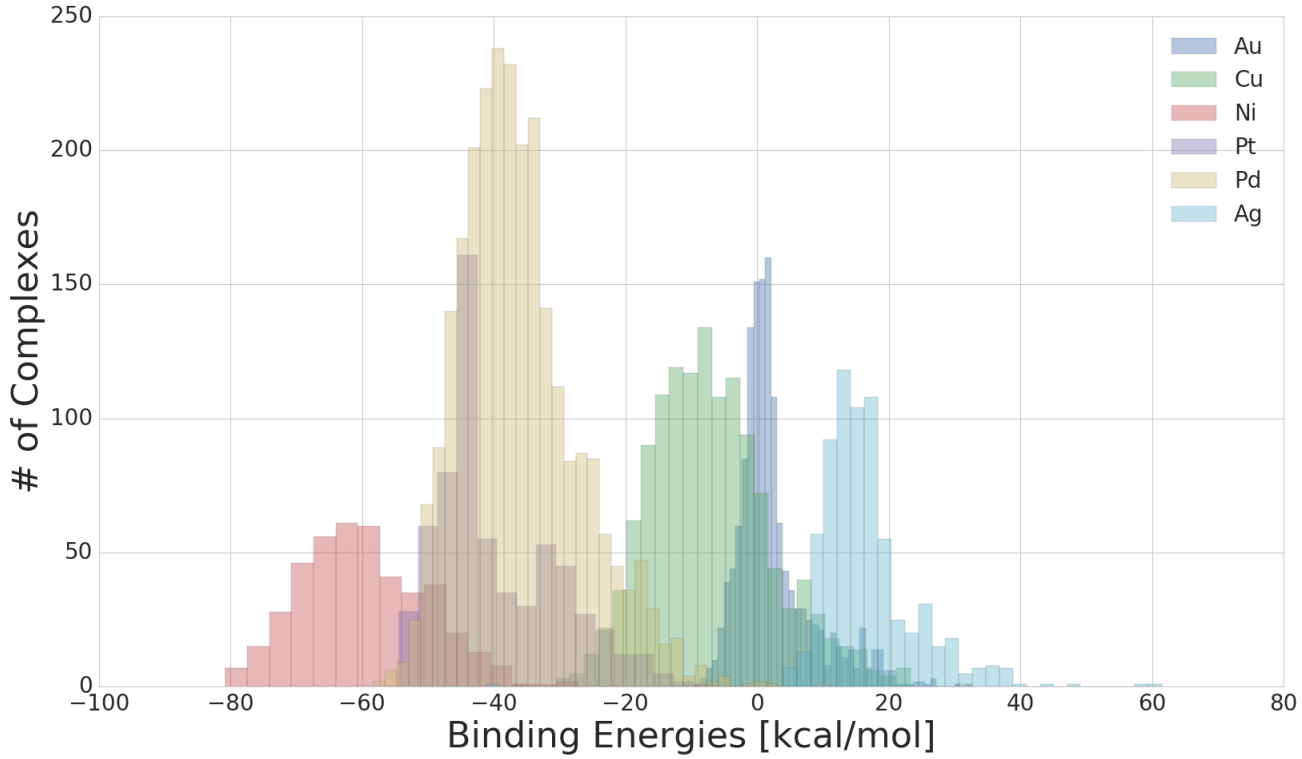


Figure S2: Histogram representing the occurrence of each metal complexes in the training set. The size of the beans are selected following the Freedman–Diaconis rule.

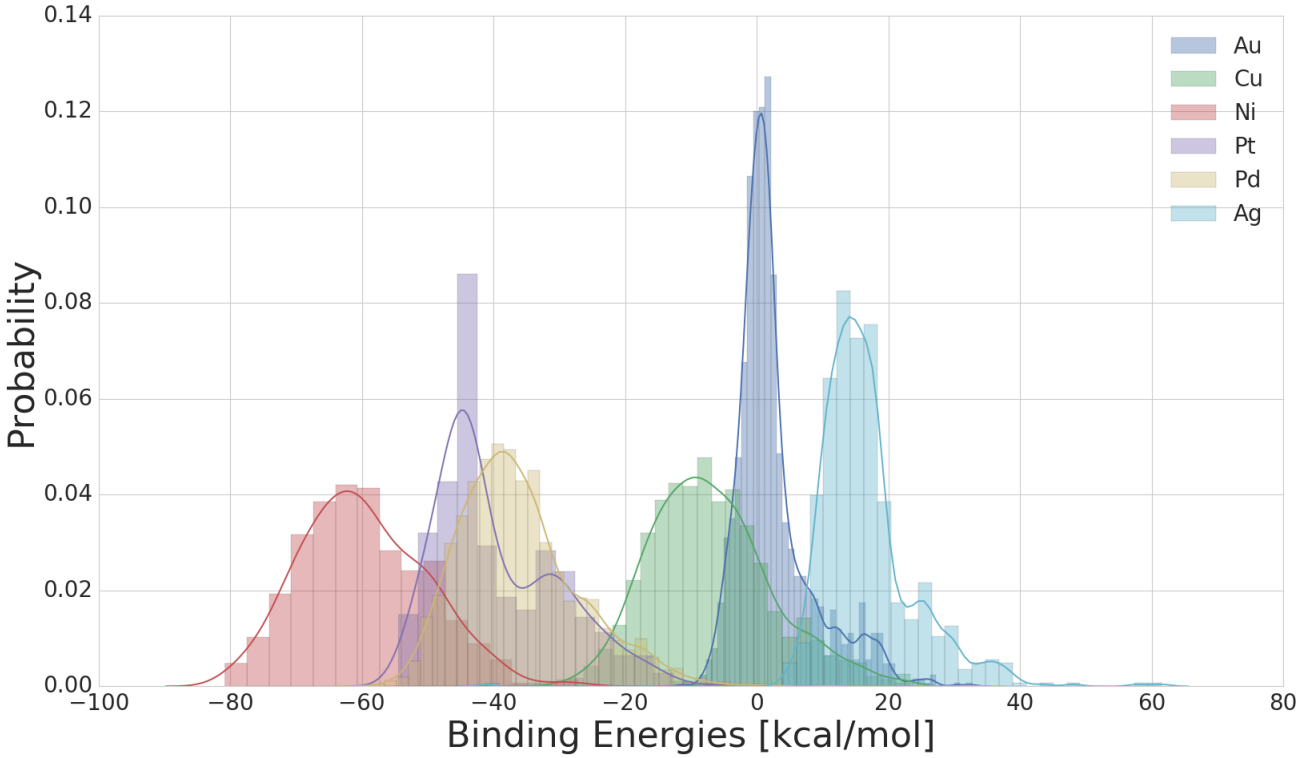


Figure S3: Linear free energy scaling relationships between the binding energies

4 Machine learning predictions

4.1 Coulomb Matrix representation

4.1.1 Distribution of the metals

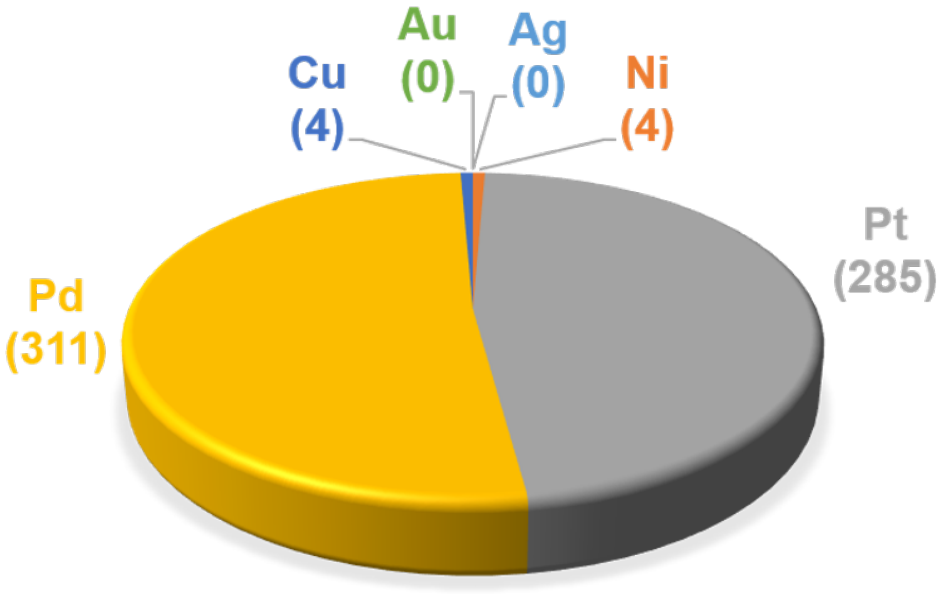


Figure S4: Occurrence of the 6 metal complexes in the selected range of -32.1/-23.0 kcal/mol predicted by the machine learning model using the Coulomb Matrix representation.

4.1.2 Distribution of the ligands

Table S4: Occurrence of the different ligands (ranging from number 0 to number 30) in the selected range of -32.1/-23.0 kcal/mol predicted by the machine learning model using the Coulomb Matrix representation

Ligand no.	Ag	Au	Ni	Cu	Pt	Pd
0	0	0	1	0	31	9
1	0	0	1	0	16	12
2	0	0	0	0	2	0
3	0	0	0	0	9	4
4	0	0	0	0	36	14
5	0	0	0	0	5	10
6	0	0	0	0	3	8
7	0	0	0	0	11	5
8	0	0	0	0	26	11
9	0	0	0	0	1	3
10	0	0	0	0	7	4
11	0	0	0	1	7	5
12	0	0	0	0	5	5
13	0	0	0	0	5	5
14	0	0	0	0	4	6
15	0	0	0	0	10	8
16	0	0	3	0	23	11
17	0	0	0	0	5	2
18	0	0	0	0	27	10
19	0	0	0	0	2	1
20	0	0	0	0	1	4
21	0	0	0	0	6	7
22	0	0	0	1	20	8
23	0	0	0	0	22	10
24	0	0	0	0	4	9
25	0	0	0	0	2	4
26	0	0	0	0	4	6
27	0	0	0	0	4	6
28	0	0	0	0	4	18
29	0	0	0	0	3	0
30	0	0	0	0	30	11

Table S5: Occurrence of the different ligands (ranging from number 31 to number 60) in the selected range of -32.1/-23.0 kcal/mol predicted by the machine learning model using the Coulomb Matrix representation

Ligand no.	Ag	Au	Ni	Cu	Pt	Pd
31	0	0	0	0	8	5
32	0	0	0	0	0	0
33	0	0	0	0	0	1
34	0	0	0	0	0	0
35	0	0	0	0	2	1
36	0	0	0	0	0	2
37	0	0	0	2	2	2
38	0	0	0	0	0	1
39	0	0	0	0	0	1
40	0	0	0	0	0	1
41	0	0	0	1	4	4
42	0	0	0	0	0	0
43	0	0	0	0	1	1
44	0	0	0	2	0	1
45	0	0	0	0	1	0
46	0	0	0	0	0	0
47	0	0	0	0	1	1
48	0	0	0	0	0	1
49	0	0	0	0	0	0
50	0	0	0	0	4	2
51	0	0	0	0	4	0
52	0	0	0	0	5	1
53	0	0	0	0	6	2
54	0	0	0	0	6	4
55	0	0	0	0	5	2
56	0	0	0	0	5	0
57	0	0	0	0	2	2
58	0	0	0	0	5	1
59	0	0	0	0	5	2
60	0	0	0	0	3	4

Table S6: Occurrence of the different ligands (ranging from number 61 to number 90) in the selected range of -32.1/-23.0 kcal/mol predicted by the machine learning model using the Coulomb Matrix representation

Ligand no.	Ag	Au	Ni	Cu	Pt	Pd
61	0	0	0	0	3	3
62	0	0	0	0	4	1
63	0	0	0	0	4	2
64	0	0	0	0	2	0
65	0	0	0	0	1	0
66	0	0	0	0	3	4
67	0	0	0	0	5	1
68	0	0	0	0	6	3
69	0	0	0	0	4	2
70	0	0	0	0	2	1
71	0	0	2	0	20	8
72	0	0	0	0	2	14
73	0	0	0	0	7	23
74	0	0	0	0	4	8
75	0	0	0	0	4	19
76	0	0	0	1	13	27
77	0	0	1	0	19	32
78	0	0	0	0	5	26
79	0	0	0	0	3	15
80	0	0	0	0	4	5
81	0	0	0	0	7	18
82	0	0	0	0	5	15
83	0	0	0	0	12	47
84	0	0	0	0	19	46
85	0	0	0	0	0	3
86	0	0	0	0	1	6
87	0	0	0	0	3	10
88	0	0	0	0	3	11
89	0	0	0	0	1	8
90	0	0	0	0	0	6

4.2 Bag of Bonds representation

4.2.1 Distribution of the ligands

Table S7: Occurrence of the different ligands (ranging from number 0 to number 30) in the selected range of -32.1/-23.0 kcal/mol predicted by the machine learning model using the Bag of Bonds representation

Ligand no.	Ag	Au	Ni	Cu	Pt	Pd
0	0	0	4	0	18	8
1	0	0	19	0	20	14
2	0	0	0	1	3	0
3	0	0	0	0	12	6
4	0	0	1	0	18	9
5	0	0	1	2	8	3
6	0	0	1	2	5	7
7	0	0	0	0	16	1
8	0	0	1	0	22	9
9	0	0	1	0	4	3
10	0	0	0	0	3	2
11	0	0	0	1	6	6
12	0	0	0	0	4	5
13	0	0	0	1	2	4
14	0	0	0	1	2	4
15	0	0	1	0	7	8
16	0	0	2	0	22	13
17	0	0	0	0	6	0
18	0	0	2	0	19	13
19	0	0	0	1	4	1
20	0	0	1	0	5	6
21	0	0	0	0	5	6
22	0	0	3	0	21	7
23	0	0	4	0	27	8
24	0	0	1	0	6	7
25	0	0	1	0	3	3
26	0	0	1	1	3	3
27	0	0	0	0	4	4
28	0	0	2	0	9	10
29	0	0	0	1	7	0
30	0	0	4	0	24	10

Table S8: Occurrence of the different ligands (ranging from number 31 to number 60) in the selected range of -32.1/-23.0 kcal/mol predicted by the machine learning model using the Bag of Bonds representation

Ligand no.	Ag	Au	Ni	Cu	Pt	Pd
31	0	0	1	1	12	6
32	0	0	0	0	1	0
33	0	0	0	0	0	0
34	0	0	0	1	0	1
35	0	0	0	0	1	2
36	0	0	0	1	0	1
37	0	0	0	4	0	1
38	0	0	0	0	0	1
39	0	0	0	1	0	0
40	0	0	0	0	0	0
41	0	0	0	2	0	1
42	0	0	0	1	2	0
43	0	0	0	0	2	0
44	0	0	0	0	0	0
45	0	0	0	1	0	0
46	0	0	0	1	2	0
47	0	0	1	0	2	1
48	0	0	0	0	0	1
49	0	0	0	1	0	0
50	0	0	0	0	0	2
51	0	0	0	0	1	3
52	0	0	0	0	2	3
53	0	0	0	0	1	3
54	0	0	1	0	5	5
55	0	0	0	0	1	4
56	0	0	0	0	1	4
57	0	0	0	0	0	0
58	0	0	0	0	2	1
59	0	0	0	0	2	1
60	0	0	0	0	3	1

Table S9: Occurrence of the different ligands (ranging from number 61 to number 90) in the selected range of -32.1/-23.0 kcal/mol predicted by the machine learning model using the Bag of Bonds representation

Ligand no.	Ag	Au	Ni	Cu	Pt	Pd
61	0	0	0	0	2	0
62	0	0	0	0	2	0
63	0	0	0	0	1	0
64	0	0	0	0	1	2
65	0	0	0	0	2	0
66	0	0	0	0	0	5
67	0	0	1	0	1	5
68	0	0	2	0	0	2
69	0	0	0	0	1	2
70	0	0	0	0	1	1
71	0	0	1	1	20	9
72	0	0	1	0	7	14
73	0	0	4	1	9	19
74	0	0	0	0	6	12
75	0	0	0	3	9	19
76	0	0	0	2	2	11
77	0	0	0	0	2	17
78	0	0	1	0	8	38
79	0	0	0	0	4	31
80	0	0	1	0	1	7
81	0	0	0	1	6	13
82	0	0	0	6	8	11
83	0	0	1	0	11	30
84	0	0	0	0	18	35
85	0	0	0	0	0	2
86	0	0	0	0	0	4
87	0	0	0	0	0	2
88	0	0	0	1	0	4
89	0	0	0	0	0	4
90	0	0	1	0	1	9

4.3 SLATM representation

4.3.1 Distribution of the metals

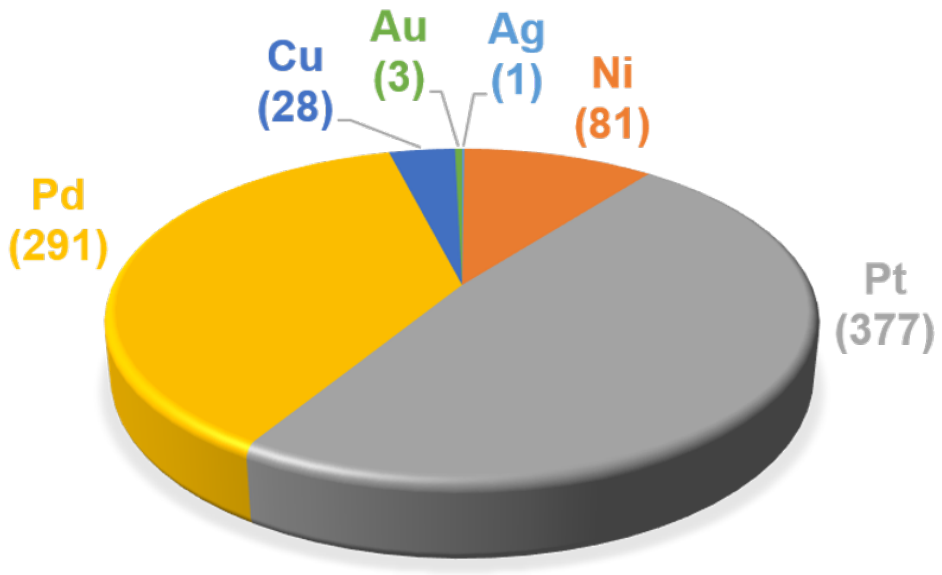


Figure S5: Occurrence of the 6 metal complexes in the selected range of -32.1/-23.0 kcal/mol predicted by the machine learning model using the SLATM representation.

4.3.2 Distribution of the ligands

Table S10: Occurrence of the different ligands (ranging from number 0 to number 30) in the selected range of -32.1/-23.0 kcal/mol predicted by the machine learning model using the SLTAM representation

Ligand no.	Ni	Cu	Ag	Au	Pt	Pd
0	0	0	13	0	32	6
1	0	0	20	0	13	12
2	0	0	1	0	4	1
3	0	0	1	0	10	1
4	0	0	1	0	30	7
5	0	0	3	0	8	5
6	0	0	1	0	6	3
7	0	0	1	0	7	1
8	0	0	7	0	11	7
9	0	0	1	1	7	3
10	0	0	1	0	5	0
11	0	0	0	0	7	3
12	0	0	0	0	5	3
13	0	0	1	0	3	4
14	0	0	1	0	8	1
15	0	0	2	0	7	5
16	0	0	4	0	17	6
17	0	0	1	0	6	1
18	0	0	1	0	10	7
19	0	0	1	1	5	1
20	0	0	2	0	3	5
21	0	0	1	0	4	5
22	0	0	4	0	13	9
23	0	0	5	0	11	6
24	0	0	1	0	5	7
25	0	0	1	0	4	7
26	0	0	1	0	4	3
27	0	0	1	0	6	2
28	0	0	13	0	11	11
29	0	0	0	1	5	1
30	0	0	3	0	16	8

Table S11: Occurrence of the different ligands (ranging from number 31 to number 60) in the selected range of -32.1/-23.0 kcal/mol predicted by the machine learning model using the SLATM representation

Ligand no.	Ni	Cu	Ag	Au	Pt	Pd
31	0	0	0	0	9	6
32	0	0	0	0	3	3
33	0	0	0	1	0	1
34	0	0	0	0	0	1
35	0	0	0	0	3	2
36	0	0	0	0	3	3
37	0	0	0	4	6	3
38	0	0	0	0	3	3
39	0	0	0	1	2	2
40	0	0	0	1	1	2
41	0	0	0	0	5	3
42	0	0	0	1	4	3
43	0	1	0	0	5	1
44	0	0	0	0	4	1
45	0	0	0	1	4	2
46	0	0	0	0	4	1
47	0	0	0	0	7	3
48	0	0	0	3	4	2
49	0	0	0	0	2	3
50	0	0	0	0	5	3
51	0	0	0	1	4	3
52	0	0	0	0	6	4
53	0	0	1	1	6	6
54	0	0	0	0	4	3
55	0	0	1	0	6	3
56	0	0	0	1	5	6
57	0	0	0	1	5	2
58	0	0	0	0	7	5
59	0	0	0	1	6	6
60	0	0	0	0	4	1

Table S12: Occurrence of the different ligands (ranging from number 61 to number 90) in the selected range of -32.1/-23.0 kcal/mol predicted by the machine learning model using the SLATM representation

Ligand no.	Ni	Cu	Ag	Au	Pt	Pd
61	0	0	0	0	6	4
62	0	0	0	1	8	2
63	0	0	0	1	4	4
64	0	0	0	0	4	2
65	0	0	0	0	5	1
66	0	0	2	0	5	2
67	0	0	4	1	6	3
68	0	0	2	0	6	2
69	0	0	0	0	5	1
70	0	0	2	1	5	2
71	0	0	6	0	18	9
72	0	0	0	1	5	23
73	0	0	11	0	28	27
74	0	0	1	0	7	19
75	0	0	1	7	4	22
76	0	2	0	0	8	10
77	2	3	1	0	11	14
78	0	0	8	0	40	5
79	0	0	12	0	54	20
80	0	0	6	0	6	37
81	0	0	2	0	10	50
82	0	0	0	19	7	15
83	0	0	1	0	9	20
84	0	0	1	1	18	20
85	0	0	0	1	1	8
86	0	0	0	1	0	7
87	0	0	0	0	2	6
88	0	0	0	1	2	7
89	0	0	0	1	3	6
90	0	0	7	0	57	6

5 Price of metals and ligands

Table S13: Estimated commercial price of the ligands (for one mmol) in US dollars

Ligand no.	Price	Ligand no.	Price	Ligand no.	Price
0	1000.00	31	500.00	61	2.95
1	1000.00	32	3.31	62	1.05
2	1.37	33	0.55	63	0.14
3	17.17	34	9.55	64	0.02
4	5.58	35	8.50	65	0.06
5	0.12	36	17.35	66	68.53
6	0.01	37	18.14	67	0.01
7	12.63	38	20.56	68	0.00
8	3.10	39	500.00	69	0.02
9	0.05	40	32.53	70	0.00
10	9.83	41	17.59	71	0.00
11	26.80	42	12.56	72	1.42
12	8.76	43	500.00	73	12.18
13	3.46	44	51.29	74	18.15
14	747.65	45	500.00	75	9.67
15	4.20	46	1000.00	76	52.27
16	18.49	47	1000.00	77	84.68
17	4.44	48	500.00	78	4.03
18	26.50	49	500.00	79	4.03
19	2.74	50	0.01	80	4.03
20	2.50	51	0.02	81	500.00
21	9.91	52	0.17	82	1.25
22	13.64	53	0.41	83	0.08
23	11.98	54	0.20	84	8.20
24	3.97	55	1.48	85	0.09
25	9.42	56	13.07	86	8.91
26	14.53	57	0.01	87	1.05
27	6.99	58	0.01	88	5.46
28	500.00	59	0.07	89	0.42
29	500.00	60	0.11	90	352.67
30	500.00				

Table S14: Estimated commercial price of the metals (for one mmol) in US dollars

Ni	Cu	Ag	Au	Pt	Pd
0.05	0.03	0.50	66.37	32.49	5.74

Table S15: Estimated commercial price (for one mmol) in US dollars of the 37 promising complexes that have a price less than 10 US$/mmol

Metal	Ligand 1 no.	Ligand 2 no.	Price
Cu	13	82	4.74
Cu	19	82	4.02
Pd	13	83	9.28
Pd	19	83	8.56
Pd	20	83	8.32
Pd	5	78	9.89
Pd	5	79	9.89
Pd	50	78	9.78
Pd	51	78	9.79
Pd	51	79	9.79
Pd	52	78	9.93
Pd	52	79	9.93
Pd	54	78	9.97
Pd	54	79	9.97
Pd	54	82	7.19
Pd	6	72	7.17
Pd	6	78	9.77
Pd	6	79	9.77
Pd	6	83	5.82
Pd	64	83	5.83
Pd	67	67	5.76
Pd	67	78	9.77
Pd	67	79	9.77
Pd	68	78	9.77
Pd	68	83	5.82
Pd	69	83	5.84
Pd	70	78	9.77
Pd	71	82	6.99
Pd	71	85	5.83
Pd	71	87	6.79
Pd	71	89	6.16
Pd	72	83	7.24
Pd	78	83	9.84
Pd	79	83	9.84
Pd	83	83	5.89
Pd	9	78	9.82
Pd	9	79	9.82

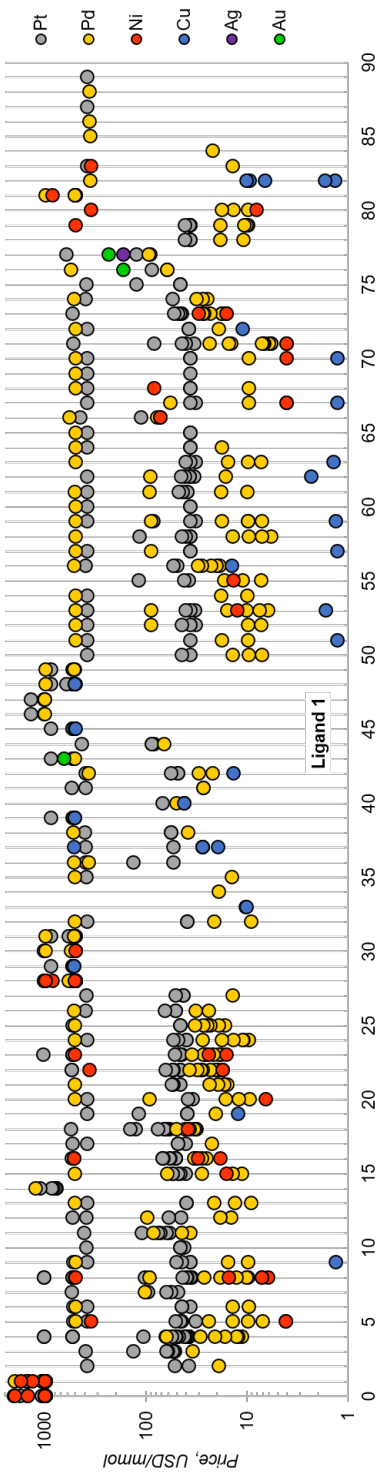


Figure S6: Estimated price (for one mmol in US dollars) of all the 557 catalysts in the selected range of -32.1/-23.0 kcal/mol.

6 Generated XYZ structures for database construction

Note that all our data (optimized structures, energies, ML predictions) are located on the Material Cloud (doi: 10.24435/materialscloud:2018.0014/v1 and 10.24435/materialscloud:2019.0007/v3).

The overall 25,116 generated structures of each catalytic intermediates **1** and **2** are provided in the folders *All2Lig* and *All4Lig*, respectively.

7 DFT Optimized XYZ structures and associated binding energies

The overall 7,054 optimized geometries at the B3LYP-D3/3-21G level of each catalytic intermediates **1** and **2** are provided in the folders *DFTgeom2Lig* and *DFTgeom4Lig*, respectively. The 700 complexes exploited in the test set are given in the file *test.txt*.

The single point energies computed at the B3LYP-D3/def2-TZVP level and the corresponding binding energies are given in the file *CompBindEn.txt*.

8 Out-of-sample machine learning predicted binding energies

The 18,062 out-of-sample machine learning predicted binding energies using the Coulomb Matrix, Bag of Bonds and SLTAM representations are given in the files *CMpredictions.out*, *BoBpredictions.out* and *SLATMpredictions.out*, respectively.