Electronic Supplementary Material (ESI) for Inorganic Chemistry Frontiers.

Electronic Supplementary Information

Data-Driven Ligand Field Exploration of Fe(IV)-oxo Sites for C-H Activation

Grier M. Jones,[1a] Brett A. Smith,[1 a] Justin K. Kirkland,[2 a] Konstantinos D. Vogiatzis* [a]

[a] *Department of Chemistry, University of Tennessee, Knoxville, Tennessee 37996, United States*

* Corresponding Author: kvogiatz@utk.edu

[1] These authors contributed equally to this work
[2] Current address: *Department of Chemistry and Biochemistry, Brigham Young University, Provo, Utah 84602, United States*

Table of Contents

S1. Example of Persistence Diagram Used as Molecular Representations

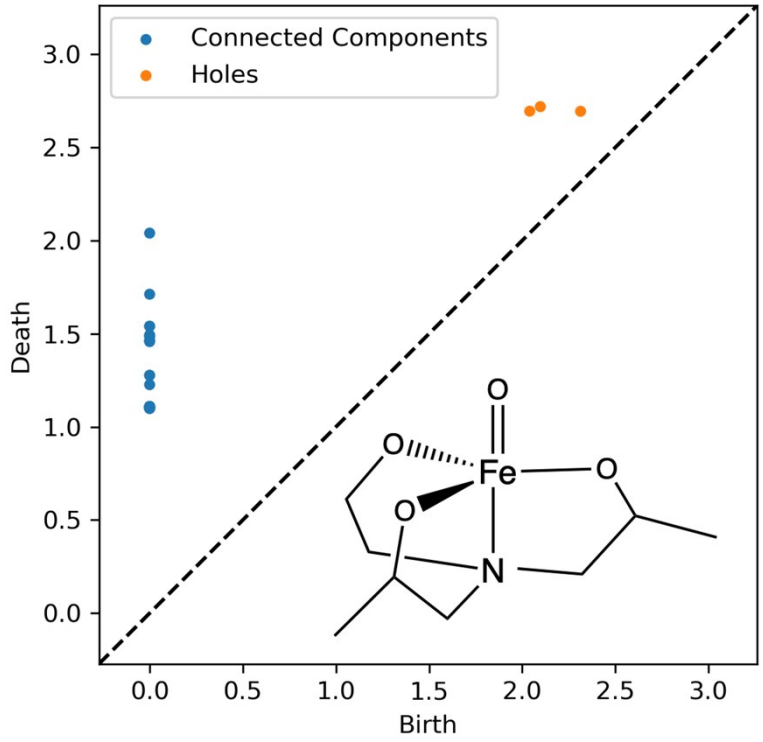


Figure S1: Example of a persistence diagram for an iron(IV)-oxo structure used in this study. The top blue point is a connected component at (0, 2.1) that corresponds to the axial Fe-O bond length of 2.097 Å. The three holes correspond to the three rings formed by the TREN ligand: the two Fe-O-C-C-N rings that include the methylated carbon are located at (2.0400, 2.6939) and (2.3130, 2.6933), while the ring without a methane is located at (2.0971, 2.7182).

Figure S1 shows how persistence diagrams are generated, for a given molecule, by placing a sphere at the center of each atom. As the radius of the sphere increases, connected components, which encode interatomic distances, and holes, which can encode information about functional groups and rings, form. The birth of a connect component occurs at 0 and the spheres are systematically expanded until the spheres intersect and a new connected component is formed. When all spheres that form a hole intersect the death of a hole occurs. Persistence is defined as the difference between birth and death. Figure S1 shows an example of a persistence diagram

from an iron(IV)-oxo species used in this study. For more details, see Ref 1 and the persistence image webpage: https://maroulaslab.github.io/PersistentImages_Chemistry/pages/PI.html.

S2. Database Distribution

Distribution of Structures

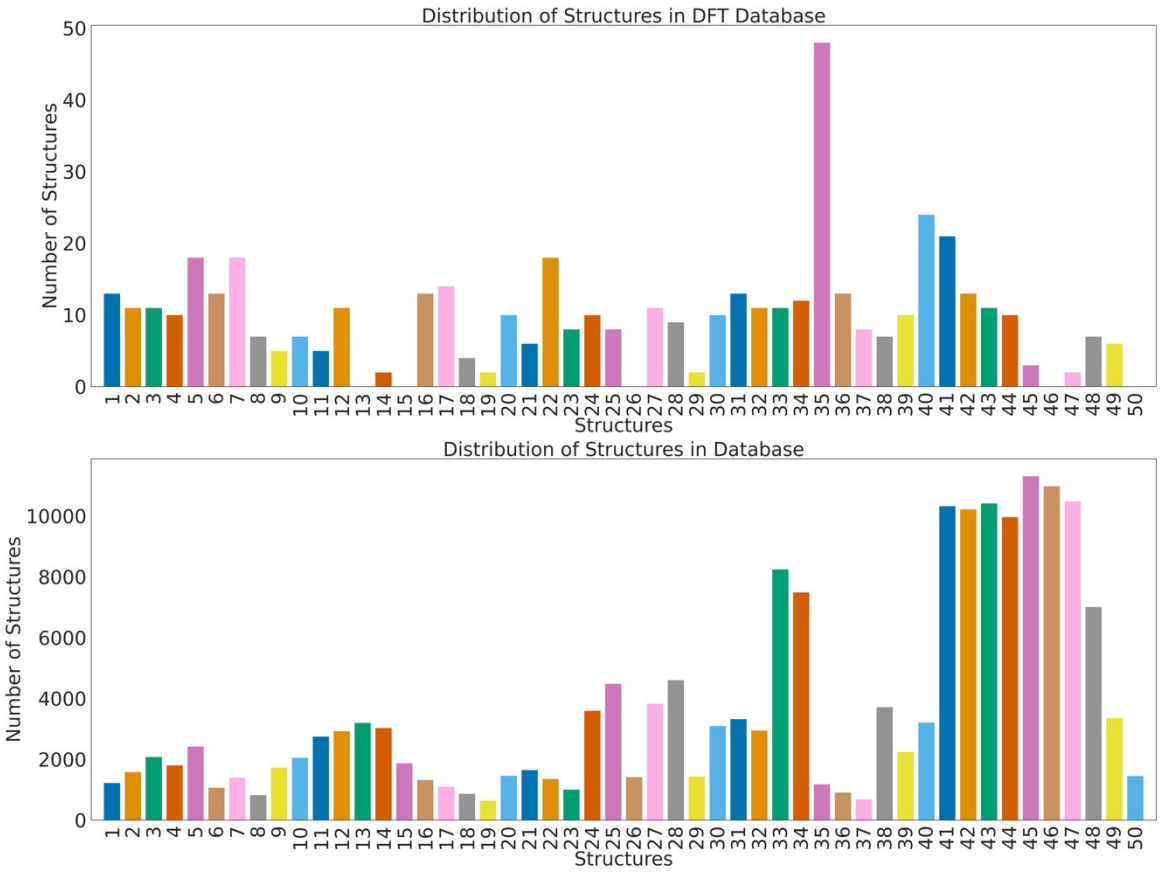


Figure S2: Distribution of structures in the DFT database (top) and the full database (bottom). Base structures that account for most of the structures in the full database correspond to structures that have more hydrogens, like structures **41**-**47**, so there are more places to perform single and double substitutions. While structures were initially added evenly to the run the DFT calculations, the DFT database corresponds to structures that are used in the ML model.

Distribution of Substituents

Functionalization	Average	Min	Max
Br/Br	-1.75	-10.21	4.84
F/Br	-1.45	-11.46	3.41
F/F	-1.45	-10.74	4.90
Cl/CH_3	-1.25	-15.69	9.14
F/CH_3	-0.98	-14.97	9.13
Br	-0.96	-8.98	2.75
Br/CH_3	-0.90	-15.64	8.47
F	-0.86	-10.16	3.66
Br/Cl	-0.77	-10.24	4.77
F/Cl	-0.76	-11.36	4.61
CH_3/CH_3	-0.59	-22.54	11.05
Cl	-0.32	-9.30	2.27
CH_3	-0.10	-12.00	8.57
Cl/Cl	-0.07	-10.24	4.67
F/NH_2	0.09	-12.49	9.13
Br/NH_2	0.12	-13.18	7.90
NH_2	0.52	-9.67	6.71
NH_2/CH_3	0.76	-13.91	11.97
Cl/NH_2	0.94	-13.11	7.97
NH_2/NH_2	1.88	-11.56	10.43

Table S1: Effects of functionalization by type. All values have been obtained by calculating the relative change in activation barrier for a functionalized complex in comparison to its' original base complex (all values are in units of kcal/mol)

The effects of each type of functionalization can also be examined and trends extracted from them. The table above shows the relative change in activation barrier as a function of functionalization type. These values are averaged across all complexes used in the full database. The activation barrier appears to lower when a complex is functionalized with more electronegative groups (halides). Trends between halide also highlight that fluorine, the most electronegative and least polarizable, lowers the barrier more on average than any other functionalization type. On the opposing end of the spectrum, ammine functionalization raises

the barrier on average. The ammine groups are generally going to serve as electron donating groups, which in many cases will lower the ligand field strength and raise the barrier. The halide functionalization groups will serve as electron withdrawing groups, which may polarize the metal-ligand bonds, raising the energy of the iron(IV)-oxo intermediate. Additionally, these functionalizations are heavily dependent on positioning and the overall ligand architecture.

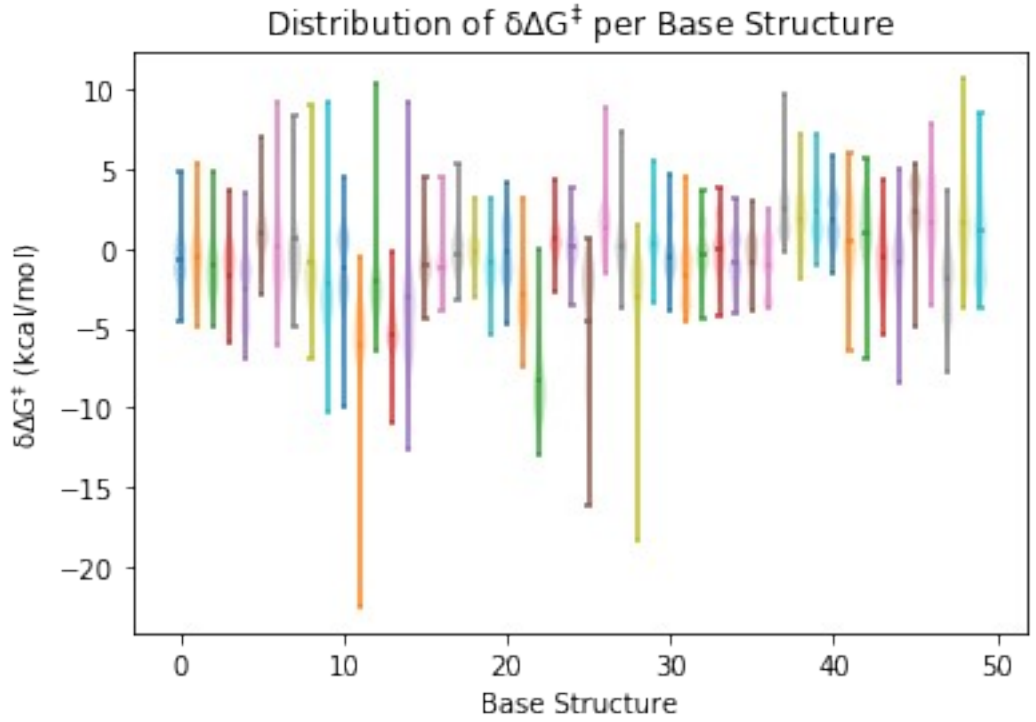
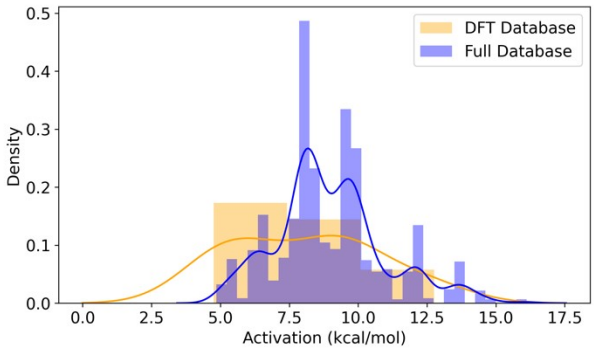


Figure S3: Distribution of the change in C-H activation energy ($\delta\Delta G^{\ddagger}$) in kcal/mol, to highlight the range, spread and average changes in each base structure.

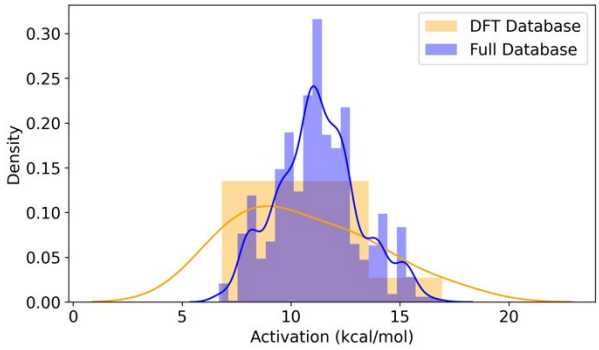
The following plots show a graph containing the kernel density estimate and bar plots of the activation energies, in kcal/mol, for the DFT and full database. The right side of the plot show a table containing the count, mean, standard deviation, minimum, 25% percentile, 50% percentile, 75% percentile, and maximum of the DFT and full database, along with the difference between them. This shows how well the model is predicting the spread of the activation energies for each base structure. We used this model evaluation to select structures from the full database data to perform DFT on. In the manuscript we showed that his is a valid assumption for the validation of the DFT data.

1 (a)



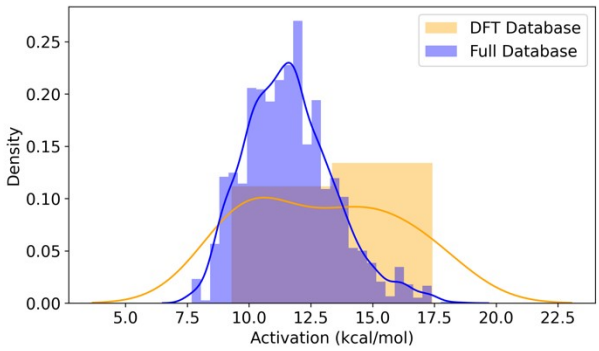
	DFT Database	Full Database	Difference
count	13.0	1226.0	nan
mean	8.0449	8.9879	0.943
std	2.643	1.9753	0.6677
min	4.751	4.8572	0.1063
25%	5.6416	7.9413	2.2996
50%	9.0952	8.5578	0.5373
75%	9.4151	9.9568	0.5417
max	12.7567	16.1191	3.3624

2 (b)



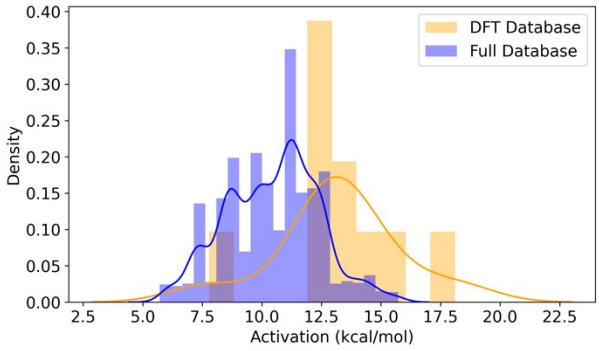
	DFT Database	Full Database	Difference
count	11.0	1583.0	nan
mean	10.5629	11.2245	0.6616
std	3.2007	1.8993	1.3014
min	6.8386	6.6922	0.1463
25%	8.0911	9.8591	1.7681
50%	10.2498	11.1051	0.8553
75%	13.0434	12.4109	0.6325
max	16.9233	17.0038	0.0805

3 (c)



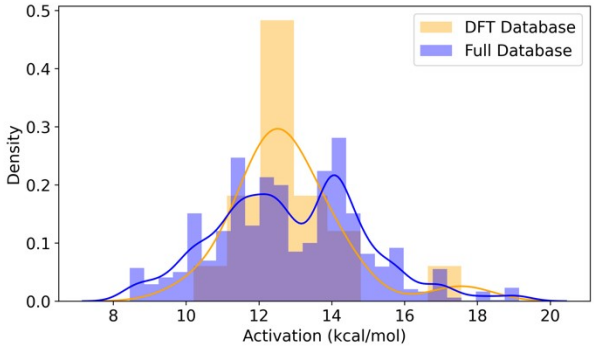
	DFT Database	Full Database	Difference
count	11.0	2083.0	nan
mean	12.9416	11.6358	1.3057
std	3.0265	1.8021	1.2244
min	9.2879	7.6771	1.6108
25%	10.4851	10.3494	0.1356
50%	13.6056	11.4848	2.1209
75%	15.0249	12.7713	2.2536
max	17.4235	18.5078	1.0843

4 (s)



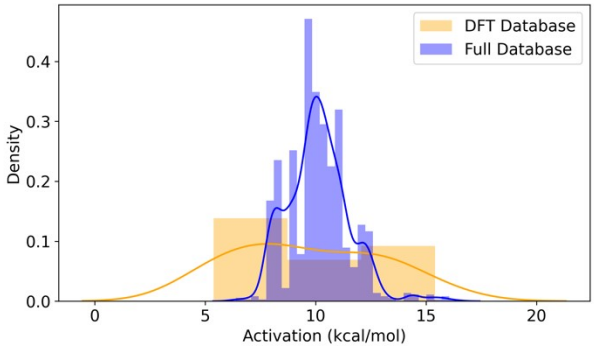
	DFT Database	Full Database	Difference
count	10.0	1807.0	nan
mean	13.2044	10.448	2.7564
std	2.582	1.9288	0.6533
min	7.7818	5.7115	2.0703
25%	12.6883	8.8657	3.8226
50%	12.9474	10.6731	2.2742
75%	13.9081	11.8393	2.0688
max	18.1044	15.7249	2.3795

5 (u)



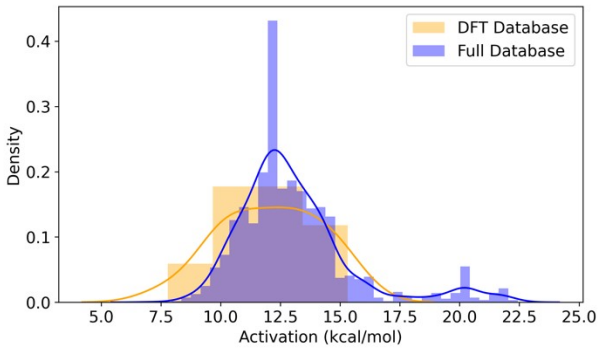
	DFT Database	Full Database	Difference
count	18.0	2424.0	nan
mean	12.8408	12.8917	0.0509
std	1.5435	2.0405	0.4971
min	10.2021	8.4497	1.7524
25%	12.1771	11.4522	0.7249
50%	12.538	12.7594	0.2214
75%	13.3867	14.1489	0.7622
max	17.5573	19.1482	1.5908

6 (af)



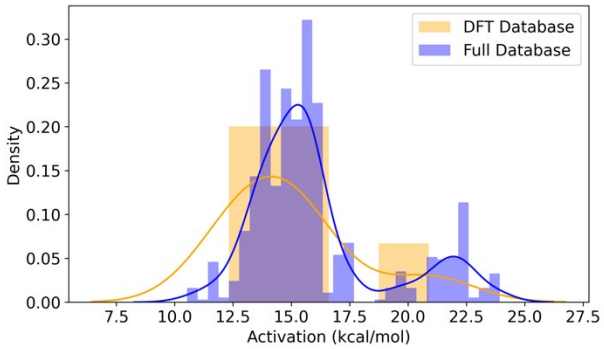
	DFT Database	Full Database	Difference
count	13.0	1071.0	nan
mean	9.6492	10.1245	0.4753
std	3.2995	1.3871	1.9124
min	5.383	6.3894	1.0063
25%	7.8841	9.1545	1.2703
50%	8.7914	10.08	1.2886
75%	12.4042	10.9582	1.4459
max	15.401	16.3904	0.9894

7 (ag)



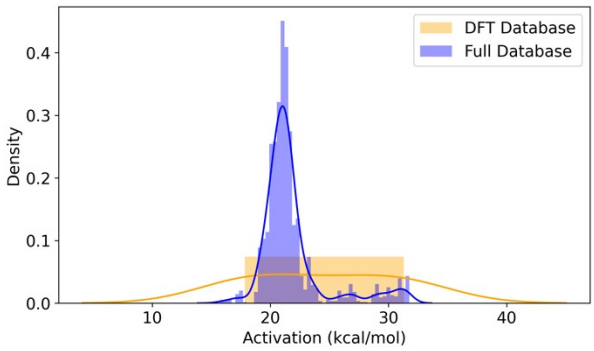
	DFT Database	Full Database	Difference
count	18.0	1399.0	nan
mean	12.0834	13.0555	0.9721
std	2.1243	2.4424	0.318
min	7.7872	7.1449	0.6423
25%	10.2381	11.6662	1.4281
50%	12.2681	12.6197	0.3516
75%	13.428	13.9655	0.5376
max	15.3124	22.4404	7.128

8 (ah)



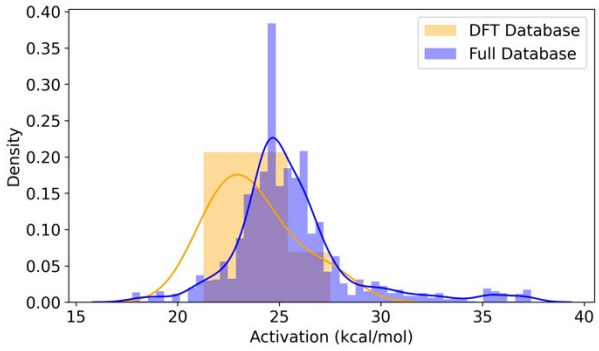
	DFT Database	Full Database	Difference
count	7.0	828.0	nan
mean	15.0235	16.0317	1.0082
std	2.8804	2.8987	0.0183
min	12.3287	10.5389	1.7898
25%	13.2121	14.1096	0.8975
50%	14.5386	15.4035	0.8649
75%	15.4954	16.278	0.7826
max	20.8821	23.9353	3.0532

9 (aq)



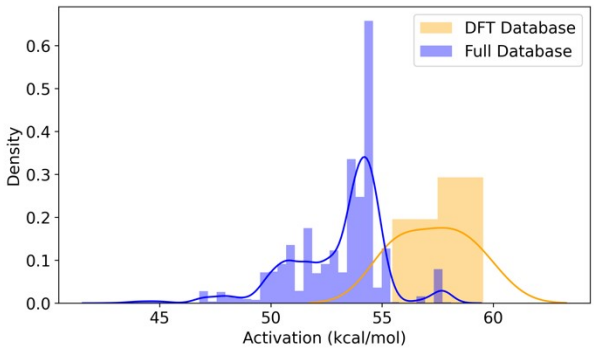
	DFT Database	Full Database	Difference
count	5.0	1731.0	nan
mean	24.4383	21.8109	2.6274
std	6.3077	2.7901	3.5176
min	17.826	15.7441	2.0819
25%	18.5745	20.4558	1.8813
50%	24.231	21.1279	3.1031
75%	30.2486	21.8274	8.4212
max	31.3115	31.7566	0.4451

10 (ar)



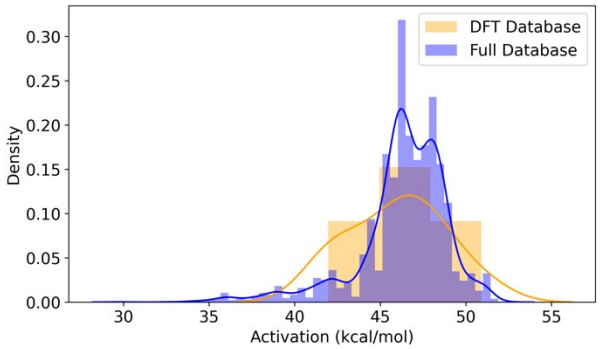
	DFT Database	Full Database	Difference
count	7.0	2056.0	nan
mean	23.7209	25.5416	1.8206
std	2.0929	2.9833	0.8903
min	21.2763	17.7579	3.5184
25%	22.3574	24.0179	1.6606
50%	23.6342	24.9454	1.3112
75%	24.4626	26.2344	1.7719
max	27.4961	37.3893	9.8932

11 (at)



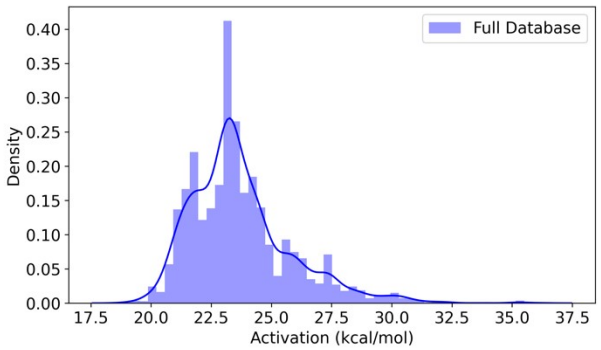
	DFT Database	Full Database	Difference
count	5.0	2753.0	nan
mean	57.3386	53.0402	4.2984
std	1.7121	2.1624	0.4502
min	55.4527	42.8775	12.5751
25%	55.8169	51.6444	4.1724
50%	57.5796	53.6666	3.913
75%	58.2966	54.4162	3.8805
max	59.547	58.1067	1.4403

12 (au)



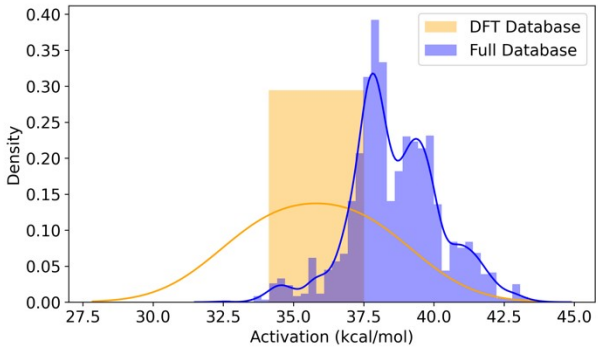
	DFT Database	Full Database	Difference
count	11.0	2930.0	nan
mean	45.9805	46.3162	0.3357
std	2.8295	2.6545	0.175
min	41.9596	29.8067	12.1529
25%	44.0281	45.5195	1.4914
50%	46.4174	46.559	0.1416
75%	47.765	48.0239	0.259
max	50.9103	52.3446	1.4343

13 (av)



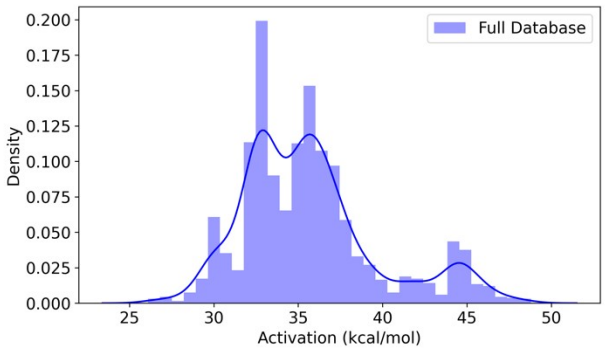
	DFT Database	Full Database	Difference
count	0.0	3200.0	nan
mean	nan	23.7123	nan
std	nan	2.1166	nan
min	nan	18.8374	nan
25%	nan	22.2258	nan
50%	nan	23.3106	nan
75%	nan	24.5725	nan
max	nan	36.1962	nan

14 (aw)



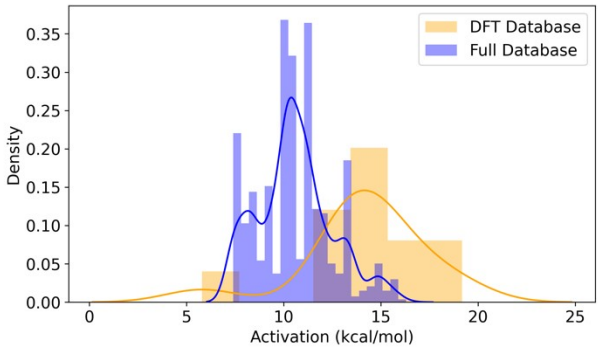
	DFT Database	Full Database	Difference
count	2.0	3033.0	nan
mean	35.8134	38.6212	2.8078
std	2.4015	1.6224	0.7791
min	34.1153	32.4666	1.6487
25%	34.9644	37.5988	2.6344
50%	35.8134	38.42	2.6066
75%	36.6625	39.6339	2.9715
max	37.5115	43.9087	6.3972

15 (as)



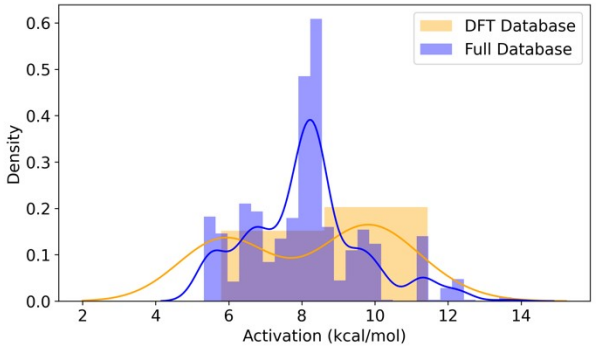
	DFT Database	Full Database	Difference
count	0.0	1875.0	nan
mean	nan	35.6265	nan
std	nan	4.0611	nan
min	nan	26.0977	nan
25%	nan	32.7815	nan
50%	nan	35.169	nan
75%	nan	37.2388	nan
max	nan	48.8056	nan

16 (d)



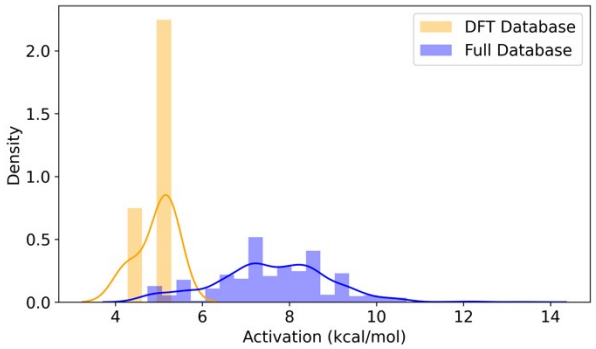
	DFT Database	Full Database	Difference
count	13.0	1323.0	nan
mean	14.172	10.589	3.5831
std	3.1314	1.9059	1.2255
min	5.7958	7.4092	1.6135
25%	13.4351	9.4084	4.0268
50%	13.7506	10.3941	3.3565
75%	15.771	11.698	4.0729
max	19.1792	16.3057	2.8735

17 (e)



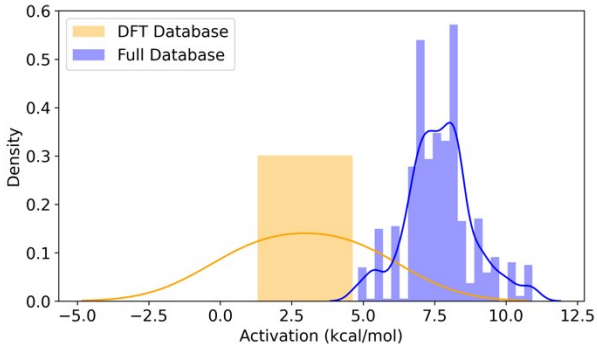
	DFT Database	Full Database	Difference
count	14.0	1102.0	nan
mean	8.2064	8.1053	0.1011
std	2.143	1.5674	0.5756
min	5.7961	5.3159	0.4803
25%	5.9038	6.8878	0.9839
50%	9.468	8.1194	1.3486
75%	9.8487	8.5889	1.2598
max	11.4371	13.7266	2.2895

18 (f)



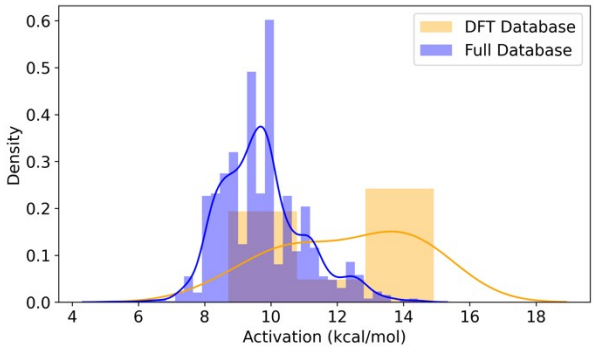
	DFT Database	Full Database	Difference
count	4.0	872.0	nan
mean	4.9473	7.5945	2.6472
std	0.4532	1.3072	0.854
min	4.281	4.7399	0.4589
25%	4.8672	6.8727	2.0055
50%	5.1127	7.6431	2.5304
75%	5.1928	8.4784	3.2856
max	5.2827	13.3435	8.0608

19 (g)



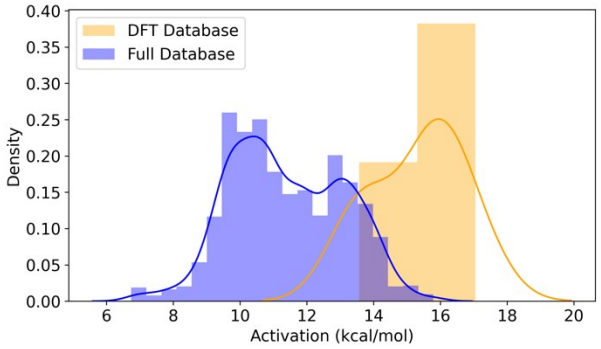
	DFT Database	Full Database	Difference
count	2.0	644.0	nan
mean	2.9709	7.6743	4.7034
std	2.348	1.182	1.166
min	1.3106	4.8241	3.5135
25%	2.1407	6.9887	4.848
50%	2.9709	7.725	4.7542
75%	3.801	8.2553	4.4543
max	4.6312	10.9256	6.2944

20 (k)



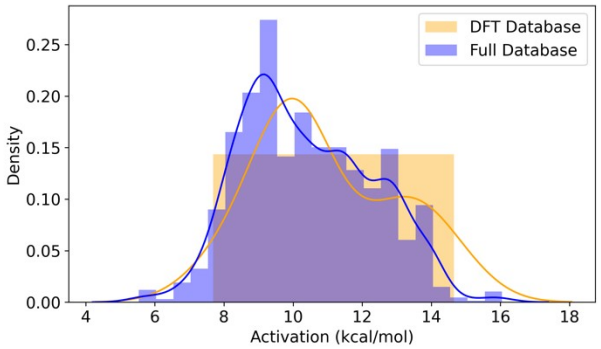
	DFT Database	Full Database	Difference
count	10.0	1463.0	nan
mean	12.2897	9.7341	2.5556
std	2.108	1.2667	0.8413
min	8.7166	5.2051	3.5116
25%	10.5866	8.8317	1.7549
50%	12.7453	9.6364	3.1089
75%	13.8601	10.3907	3.4695
max	14.9161	14.432	0.4841

21 (l)



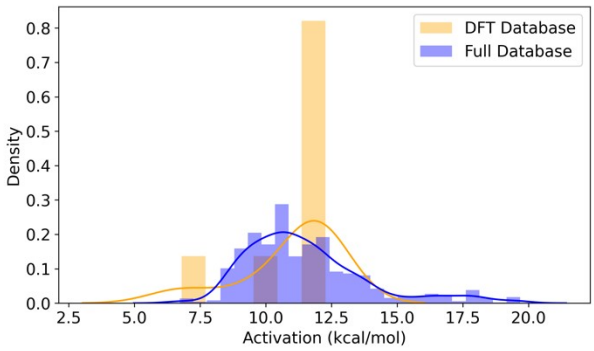
	DFT Database	Full Database	Difference
count	6.0	1650.0	nan
mean	15.3135	11.3945	3.9191
std	1.37	1.6981	0.328
min	13.5633	6.7411	6.8222
25%	14.2569	10.1007	4.1561
50%	15.8268	11.2025	4.6243
75%	15.8677	12.8716	2.996
max	17.0496	15.7889	1.2606

22 (m)



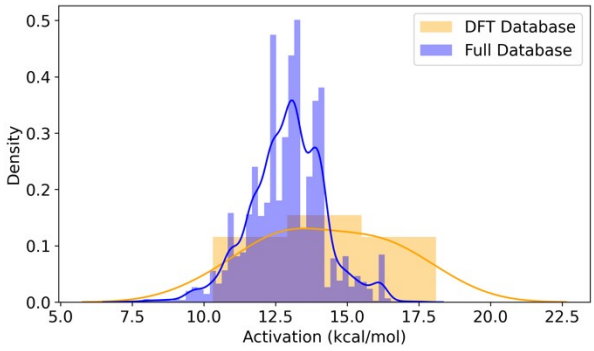
	DFT Database	Full Database	Difference
count	18.0	1356.0	nan
mean	10.9755	10.4091	0.5664
std	2.0179	1.884	0.1339
min	7.6834	5.5347	2.1488
25%	9.6549	8.9359	0.719
50%	10.1988	10.1681	0.0308
75%	12.9218	11.74	1.1817
max	14.656	16.0477	1.3916

23 (n)



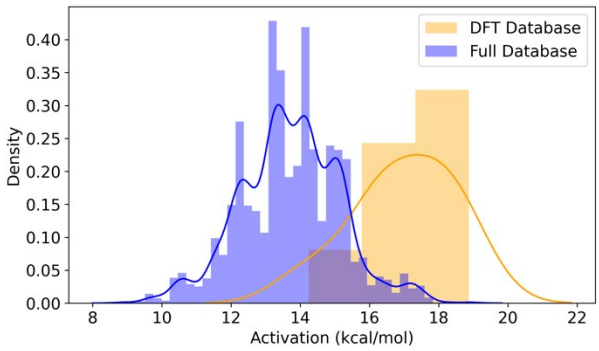
	DFT Database	Full Database	Difference
count	8.0	1006.0	nan
mean	10.9775	11.4885	0.511
std	1.8999	2.3201	0.4202
min	6.7906	6.7218	0.0688
25%	11.0093	9.8242	1.185
50%	11.6954	10.9946	0.7008
75%	12.1045	12.4253	0.3208
max	12.2728	19.6866	7.4138

24 (r)



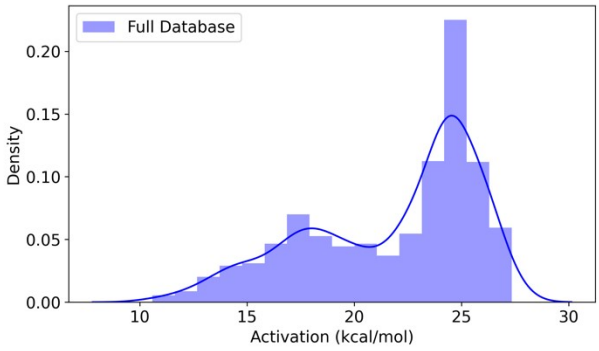
	DFT Database	Full Database	Difference
count	10.0	3602.0	nan
mean	14.2535	12.9012	1.3523
std	2.3981	1.3076	1.0905
min	10.3165	7.2581	3.0583
25%	12.8012	12.1423	0.6589
50%	14.0222	12.966	1.0562
75%	16.1881	13.7889	2.3993
max	18.094	17.574	0.52

25 (o)



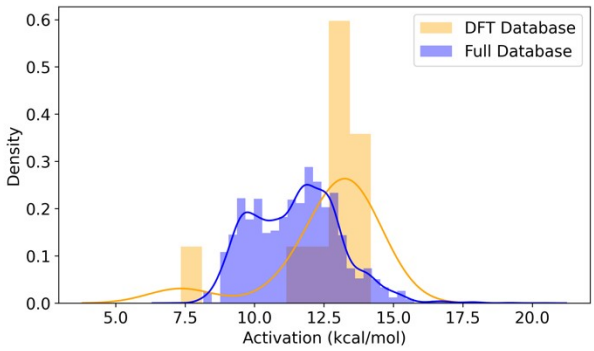
	DFT Database	Full Database	Difference
count	8.0	4487.0	nan
mean	16.9934	13.7182	3.2752
std	1.4973	1.4397	0.0576
min	14.2393	8.8003	5.439
25%	16.2865	12.7783	3.5082
50%	17.0325	13.7069	3.3256
75%	18.1541	14.741	3.4131
max	18.8725	19.0222	0.1497

26 (bb)



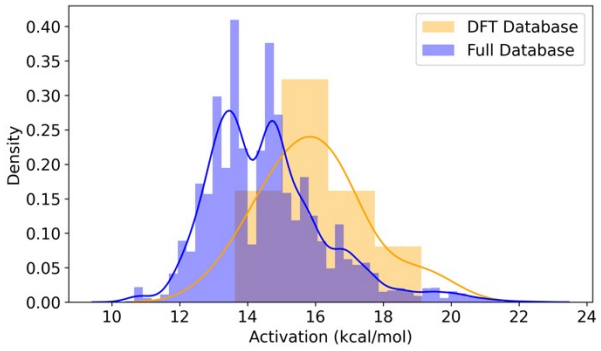
	DFT Database	Full Database	Difference
count	0.0	1417.0	nan
mean	nan	21.7686	nan
std	nan	3.9631	nan
min	nan	10.5842	nan
25%	nan	18.6338	nan
50%	nan	23.3221	nan
75%	nan	24.663	nan
max	nan	27.3423	nan

27 (t)



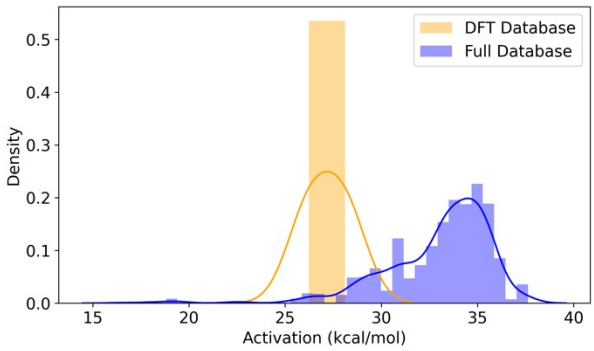
	DFT Database	Full Database	Difference
count	11.0	3831.0	nan
mean	12.6116	11.494	1.1176
std	1.8999	1.5873	0.3127
min	7.3454	7.2438	0.1016
25%	12.6567	10.1599	2.4968
50%	13.1895	11.5724	1.6172
75%	13.5266	12.5585	0.9681
max	14.1926	20.3022	6.1097

28 (p)



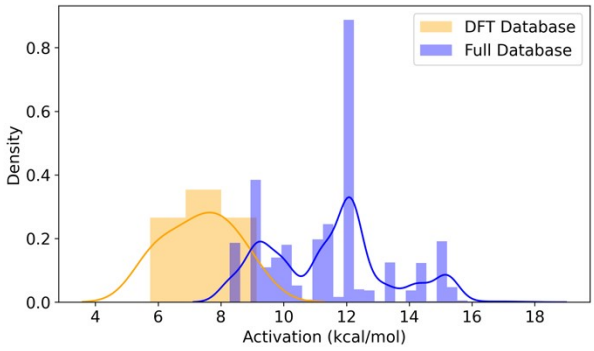
	DFT Database	Full Database	Difference
count	9.0	4609.0	nan
mean	15.9723	14.5894	1.3829
std	1.5594	1.7437	0.1843
min	13.6259	10.3957	3.2302
25%	15.0371	13.3355	1.7016
50%	16.1445	14.4271	1.7173
75%	16.6913	15.4836	1.2076
max	19.1227	22.4954	3.3727

29 (bc)



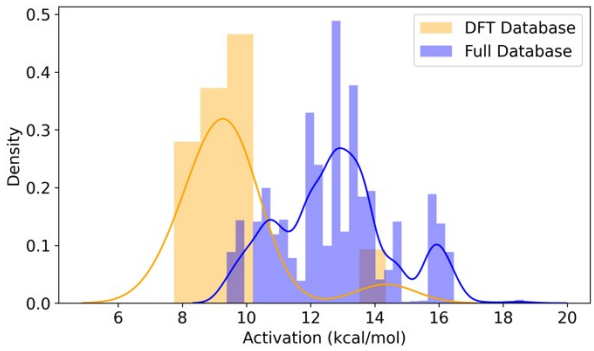
	DFT Database	Full Database	Difference
count	2.0	1442.0	nan
mean	27.1762	32.97	5.7938
std	1.3211	2.8304	1.5093
min	26.2421	16.4568	9.7853
25%	26.7092	31.5317	4.8225
50%	27.1762	33.5963	6.42
75%	27.6433	34.8898	7.2464
max	28.1104	37.6253	9.515

30 (h)



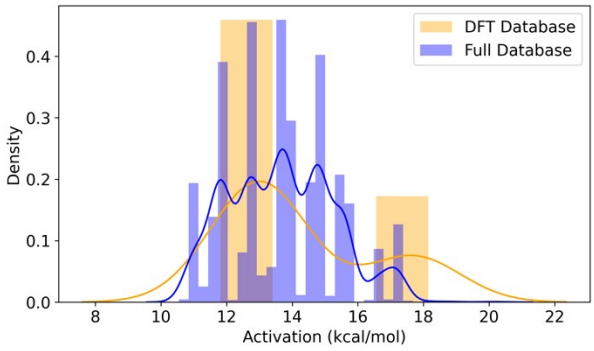
	DFT Database	Full Database	Difference
count	10.0	3101.0	nan
mean	7.3478	11.4471	4.0993
std	1.135	1.9184	0.7834
min	5.7453	8.282	2.5367
25%	6.5391	9.7913	3.2522
50%	7.3725	11.854	4.4815
75%	8.1096	12.2026	4.0929
max	9.137	17.84	8.703

31 (i)



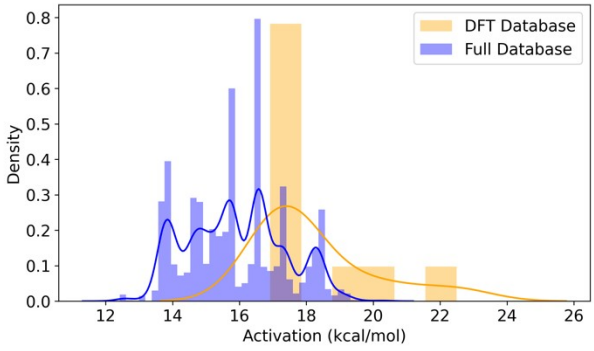
	DFT Database	Full Database	Difference
count	13.0	3327.0	nan
mean	9.5431	12.7648	3.2217
std	1.5822	1.7605	0.1783
min	7.7357	9.3848	1.6491
25%	8.7099	11.6497	2.9398
50%	9.3319	12.7354	3.4035
75%	9.7445	13.7412	3.9966
max	14.3441	18.9207	4.5766

32 (j)



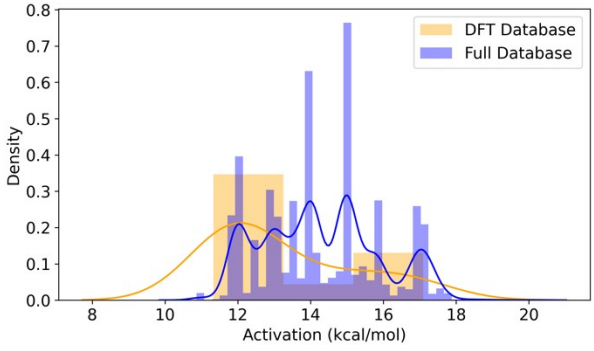
	DFT Database	Full Database	Difference
count	11.0	2954.0	nan
mean	14.2383	13.6859	0.5524
std	2.2613	1.6035	0.6578
min	11.8122	10.5454	1.2668
25%	13.0903	12.625	0.4652
50%	13.2357	13.6402	0.4045
75%	15.2993	14.8029	0.4964
max	18.1455	20.347	2.2015

33 (ac)



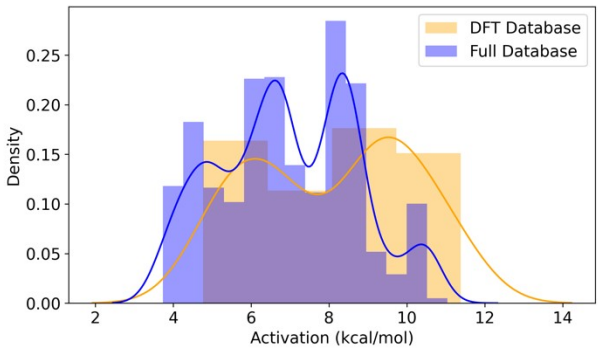
	DFT Database	Full Database	Difference
count	11.0	8251.0	nan
mean	18.2436	15.8942	2.3495
std	1.7626	1.4585	0.3041
min	16.9182	12.0356	4.8826
25%	17.2734	14.7849	2.4885
50%	17.449	15.7736	1.6753
75%	18.4612	16.7509	1.7103
max	22.4937	20.4723	2.0215

34 (ab)



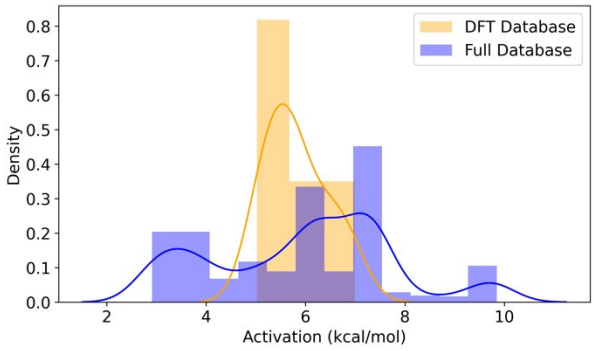
	DFT Database	Full Database	Difference
count	12.0	7493.0	nan
mean	13.2221	14.2685	1.0464
std	1.9673	1.6015	0.3658
min	11.3287	10.6531	0.6755
25%	11.8363	12.9797	1.1434
50%	12.3904	14.0415	1.6511
75%	14.4322	15.0923	0.6601
max	17.0968	20.2138	3.117

35 (v)



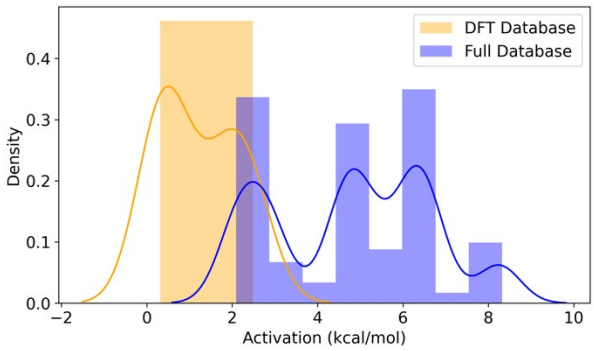
	DFT Database	Full Database	Difference
count	48.0	1185.0	nan
mean	8.0913	6.9559	1.1354
std	2.0534	1.7621	0.2913
min	4.7674	3.7324	1.035
25%	6.191	5.561	0.6301
50%	8.7782	6.7935	1.9847
75%	9.6501	8.3907	1.2594
max	11.3835	11.0401	0.3434

36 (x)



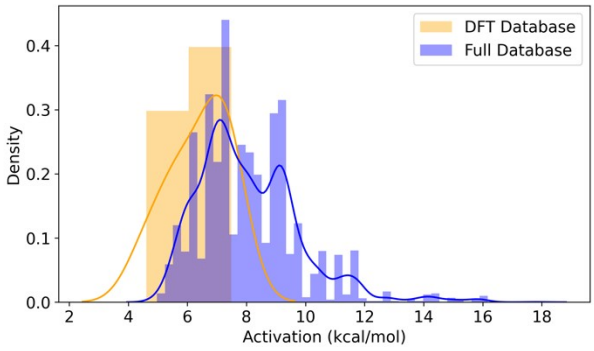
	DFT Database	Full Database	Difference
count	13.0	914.0	nan
mean	5.8185	5.9356	0.1171
std	0.6226	1.8171	1.1945
min	5.0177	2.9124	2.1053
25%	5.3834	4.4084	0.9749
50%	5.6087	6.2463	0.6377
75%	6.1675	7.2174	1.0499
max	6.9913	9.8481	2.8568

37 (y)



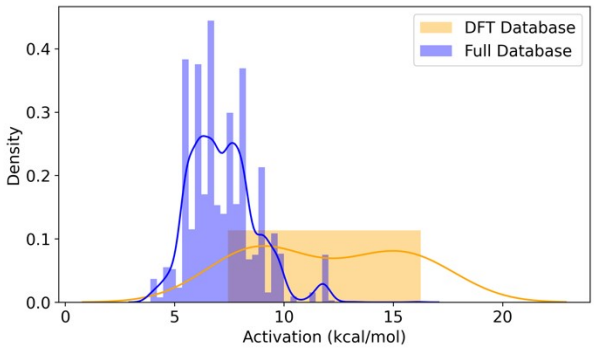
	DFT Database	Full Database	Difference
count	8.0	687.0	nan
mean	1.2099	4.8252	3.6152
std	0.917	1.8448	0.9278
min	0.3137	2.0915	1.7779
25%	0.4387	2.8251	2.3864
50%	0.9474	4.892	3.9446
75%	2.0901	6.2738	4.1836
max	2.4809	8.321	5.8401

38 (aa)



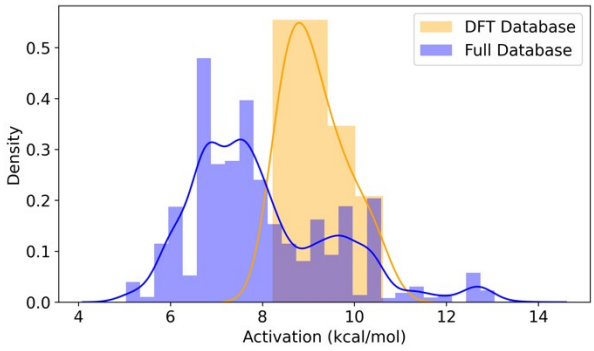
	DFT Database	Full Database	Difference
count	7.0	3717.0	nan
mean	6.3824	8.0855	1.7032
std	1.0591	1.7434	0.6843
min	4.6085	4.9661	0.3576
25%	5.7291	6.8689	1.1399
50%	6.7376	7.7908	1.0532
75%	7.1949	9.024	1.8291
max	7.4827	17.8076	10.3249

39 (ae)



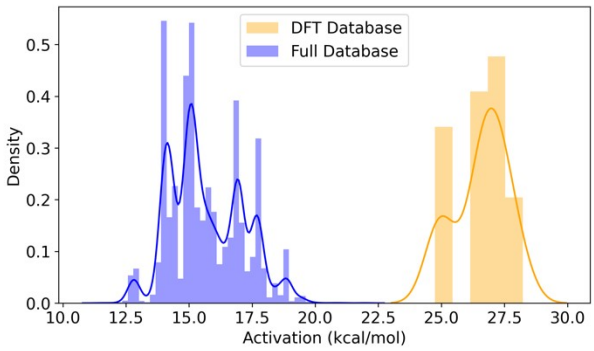
	DFT Database	Full Database	Difference
count	10.0	2246.0	nan
mean	11.955	7.2085	4.7465
std	3.5104	1.5109	1.9995
min	7.441	3.8956	3.5454
25%	8.9955	6.1588	2.8367
50%	11.6145	7.0889	4.5256
75%	15.3085	8.0962	7.2123
max	16.2494	16.1006	0.1488

40 (ad)



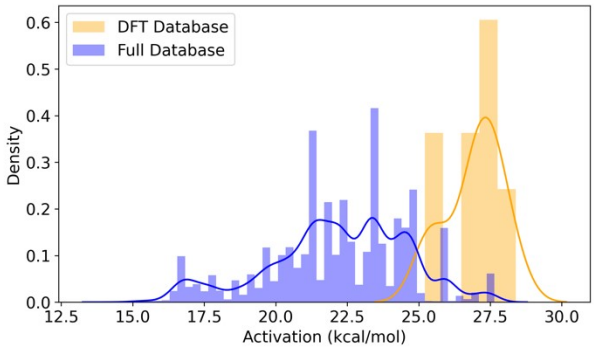
	DFT Database	Full Database	Difference
count	24.0	3216.0	nan
mean	9.149	7.9935	1.1555
std	0.6767	1.5724	0.8957
min	8.2196	5.0327	3.1869
25%	8.6627	6.7933	1.8693
50%	8.9607	7.6168	1.3439
75%	9.5615	9.1411	0.4204
max	10.624	13.6638	3.0398

41 (ai)



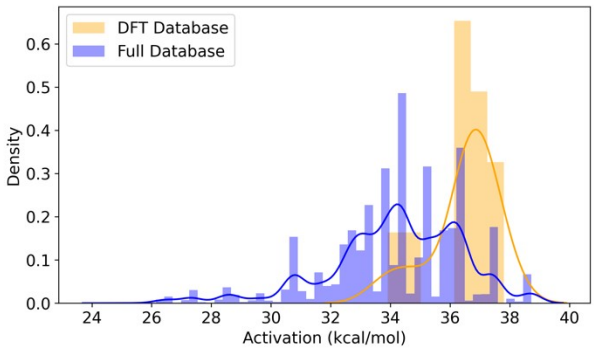
	DFT Database	Full Database	Difference
count	21.0	10329.0	nan
mean	26.5717	15.6777	10.8941
std	1.0479	1.446	0.3981
min	24.7407	11.4699	13.2708
25%	26.2043	14.5029	11.7014
50%	26.7131	15.2823	11.4308
75%	27.2477	16.9295	10.3182
max	28.2323	22.0425	6.1898

42 (aj)



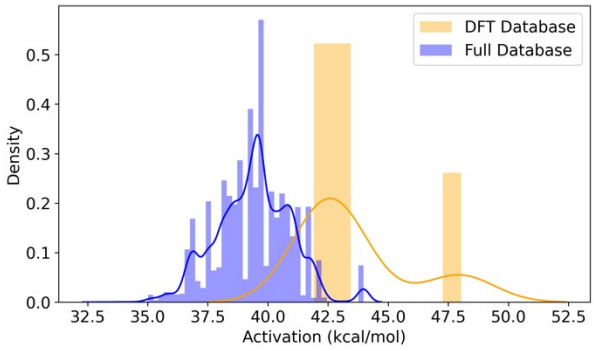
	DFT Database	Full Database	Difference
count	13.0	10222.0	nan
mean	26.9388	22.1803	4.7585
std	0.9781	2.4231	1.4451
min	25.2171	14.4008	10.8162
25%	26.6052	20.8349	5.7703
50%	27.1531	22.336	4.8171
75%	27.4643	23.7924	3.6719
max	28.393	27.6419	0.7512

43 (ak)



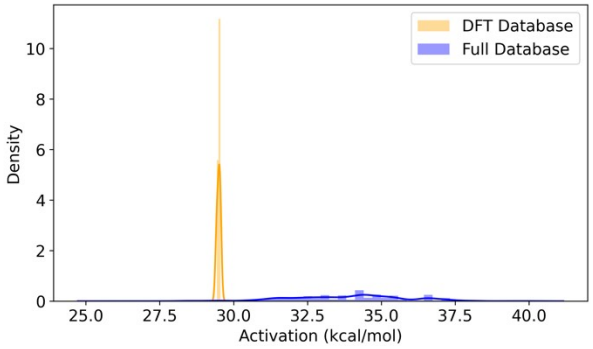
	DFT Database	Full Database	Difference
count	11.0	10419.0	nan
mean	36.4688	34.167	2.3019
std	1.1212	2.2398	1.1185
min	33.9081	24.7483	9.1598
25%	36.4971	32.9229	3.5743
50%	36.6545	34.3628	2.2917
75%	37.1285	35.9028	1.2258
max	37.8054	38.7248	0.9194

44 (am)



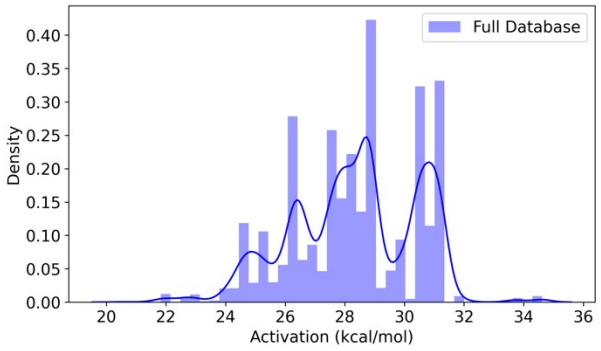
	DFT Database	Full Database	Difference
count	10.0	9966.0	nan
mean	43.6564	39.408	4.2483
std	2.3031	1.5528	0.7504
min	41.9144	33.0377	8.8767
25%	42.289	38.4171	3.8719
50%	42.8156	39.5609	3.2547
75%	43.1622	40.4078	2.7544
max	48.0341	43.9792	4.0549

45 (az)



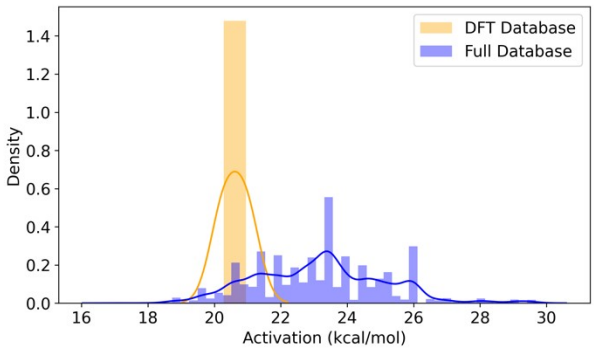
	DFT Database	Full Database	Difference
count	3.0	11312.0	nan
mean	29.4967	33.888	4.3913
std	0.0603	1.9941	1.9338
min	29.432	25.6513	3.7806
25%	29.4693	32.5607	3.0913
50%	29.5067	34.2482	4.7416
75%	29.529	35.2247	5.6957
max	29.5513	40.2408	10.6894

46 (ay)



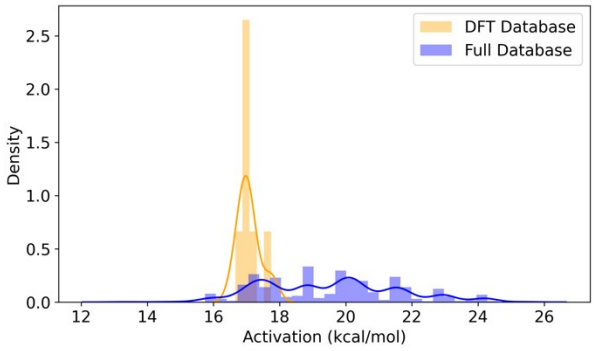
	DFT Database	Full Database	Difference
count	0.0	10983.0	nan
mean	nan	28.3052	nan
std	nan	2.085	nan
min	nan	20.512	nan
25%	nan	26.7649	nan
50%	nan	28.3005	nan
75%	nan	30.4288	nan
max	nan	34.6183	nan

47 (ax)



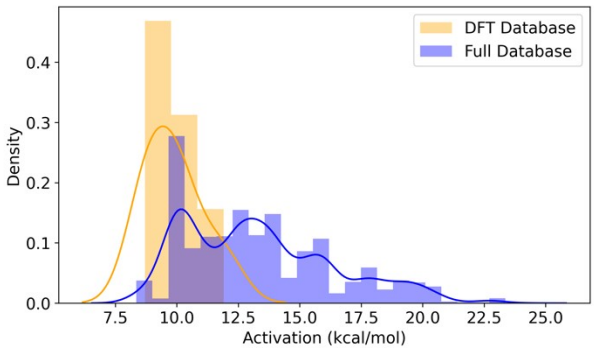
	DFT Database	Full Database	Difference
count	2.0	10481.0	nan
mean	20.6159	23.1721	2.5563
std	0.4781	1.9351	1.457
min	20.2778	16.9477	3.33
25%	20.4468	21.894	1.4471
50%	20.6159	23.3313	2.7154
75%	20.7849	24.513	3.7281
max	20.9539	29.6774	8.7234

48 (ba)



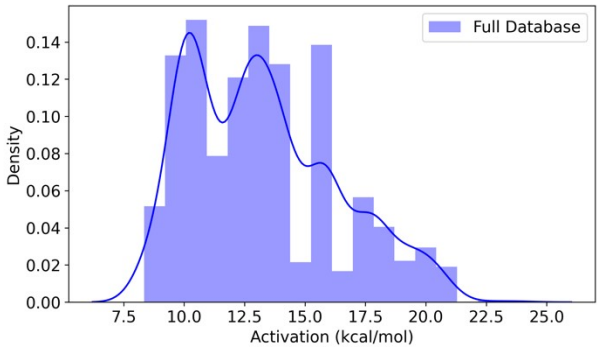
	DFT Database	Full Database	Difference
count	7.0	7012.0	nan
mean	17.0723	19.5311	2.4588
std	0.3407	2.0852	1.7446
min	16.6622	13.1118	3.5503
25%	16.9091	17.7723	0.8632
50%	16.9947	19.7776	2.7829
75%	17.1451	20.9245	3.7793
max	17.7405	25.5968	7.8562

49 (bd)



	DFT Database	Full Database	Difference
count	6.0	3361.0	nan
mean	9.8666	13.3711	3.5045
std	1.2054	3.0594	1.854
min	8.7008	8.3564	0.3445
25%	8.9167	10.3329	1.4162
50%	9.7196	12.8369	3.1173
75%	10.3117	15.4259	5.1141
max	11.9012	24.0176	12.1165

50 (be)



	DFT Database	Full Database	Difference
count	0.0	1457.0	nan
mean	nan	13.2776	nan
std	nan	3.0555	nan
min	nan	8.349	nan
25%	nan	10.3169	nan
50%	nan	12.8247	nan
75%	nan	15.42	nan
max	nan	23.8801	nan

S3. Classification Model Metrics

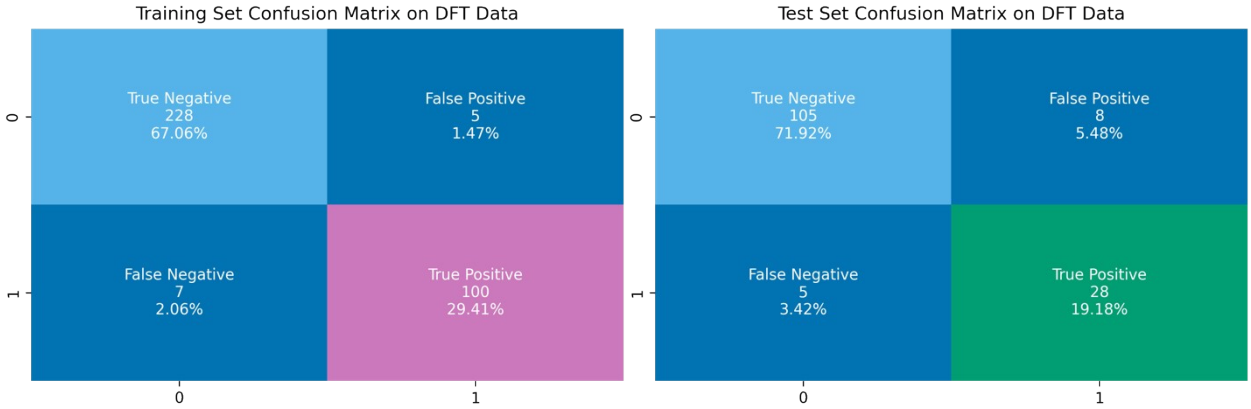


Figure S4: A true negative is a quintet that was predicted correctly, a false positive is a quintet predicted to be a triplet, a false negative is a triplet that was predicted to be a quintet, and a true positive was a triplet that was predicted correctly.

The classification model was evaluated in several ways. First, we looked at the confusion matrices (Figure S4) which show the percentage and number of values correctly or incorrectly predicted for quintets (negative values) and triplets (positive values). Second, we looked at the classification reports (Table S2) that include information about the precision, recall, F_1-score, accuracy, macro-average, and weighed average of each class.

Train	precision	recall	f1-score	support
quintet	0.97	0.98	0.97	233
triplet	0.95	0.93	0.94	107
accuracy			0.96	
macro avg	0.96	0.96	0.96	340
weighted avg	0.96	0.96	0.96	340
Test	**precision**	**recall**	**f1-score**	**support**
quintet	0.95	0.93	0.94	113
triplet	0.78	0.85	0.81	33
accuracy			0.91	
macro avg	0.87	0.89	0.88	146
weighted avg	0.91	0.91	0.91	146

Table S2: Classification reports for the training and test set.

The accuracy is defined as:

$$accuracy = \frac{true\ positive + true\ negative}{positive + \ positive}$$

The precision is the accuracy of positive predictions:

$$precision = \frac{true\ positive}{true\ positive + false\ positive}$$

The recall is the ability of a classifier to predict positive outcomes:

$$recall = \frac{true\ positive}{true\ positive + false\ negative}$$

The F_1-score is the weighted harmonic mean of precision and recall:

$$F_1 = 2 * \left(\frac{precision * recall}{precision + recall}\right)$$

The support column is the amount of data in each class. The macro average is the average without considering the proportion of each label in the dataset. The weighted average considers the proportion of each label in the dataset.

S4. Regression Model Metrics

To find the best possible model in a systematic fashion, we used GridSearchCV in Scikit-Learn to perform a three-fold cross-validation over a set of parameters defined below. For the sake of reproducibility, the three lines of code are listed verbatim below:

```
parameters={'kernel': ['laplacian','rbf'], 'alpha':np.logspace(-3,3,7),
'gamma':np.logspace(-3,3,7)}

GridSearch=GridSearchCV(KernelRidge(), param_grid=parameters, cv=3,
verbose=0, scoring='r2')

model=GridSearch.fit(X_train, reg_y_train).best_estimator_
```

To evaluate our regression model, we train and tested our model using a 70%/30% split along with 10-fold cross-validation (CV). For our 70%/30% split, we use the coefficient of determination (R^2), root-mean-squared error (RMSE), and mean absolute error (MAE) evaluation metrics. To get the average RMSE and R^2 over the complete dataset, we used 10-fold CV.

S5. Model Comparison with Common Molecular Representations

We benchmarked the PIs method against two common molecular representations, Coulomb matrices (CMs) and smooth overlap of atomic positions (SOAP), that were generated using DScribe.[2] We generated the CMs using the default parameters that use the L2-norm for producing sorted CMs and n_atoms_max equal to 55. To generate SOAPs we used a species list containing [O, Br, F, P, Cl, H, Fe, C, N], a cutoff value for the local region (rcut) of 8.0 Å, the number of radial basis functions (nmax) of 4, the maximum degree of the spherical harmonics (lmax) of 4, and a standard deviation of the Gaussians used to expand the atomic densities (sigma) of 1.5. All other parameters were set to the default parameters. We found that normalizing the SOAPs, using the normalize function in Scikit-Learn, before passing the SOAPs to the regularized entropy match (REMatch) kernel provided improved performance of the method. For the REMatch kernel, we used the Gaussian ("rbf") metric with a gamma of 2, alpha of 1.2, convergence threshold of 1e-8, and kernel normalization set to false.

The classification models for CMs and SOAPs were performed using linear ridge classification. We used five-fold cross-validation using GridSearchCV in Sci-kit Learn[3] to find the optimal alpha parameters of 1 and 0.001 for CMs and SOAPs, respectively. For the regression step using kernel ridge regression, we also used three-fold cross-validation using GridSearchCV to find the optimal kernel, alpha, and gamma parameters for CMs and SOAPs. Both CMs and SOAPs use linear kernels and gammas of 1e-3, whereas CMs has an alpha of 1e3 and SOAPs has an alpha of 1e-3. Like we did with PIs in the main text, we performed 10-fold cross-validation get average RMSEs and R^2 setting the parameters: n_splits=10, random_state=12, and shuffle=True.

Representation	Average train RMSE (kcal/mol)	Average test RMSE (kcal/mol)	Average train R^2	Average test R^2
CMs	1.77 ± 0.05	3.85 ± 1.39	0.97 ± 0.00	0.86 ± 0.05
SOAPs	1.83 ± 0.03	2.09 ± 0.30	0.97 ± 0.00	0.96 ± 0.01

PIs	1.07 ± 0.04	2.24 ± 0.40	0.99 ± 0.00	0.95 ± 0.02

Table S3: Results from the 10-fold cross-validation of the regression data using the CMs, SOAPs, and PIs molecular representations. PIs offer a competitive representation to SOAPs with respect to accuracy. Despite overfitting in the method, PIs are less computationally expensive.

Representation	Size	Timing (s)
CMs	3025	0.59
SOAPs	486	106.00
PIs	400	7.82

Table S4: Size and timings of each molecular representation over the full DFT dataset. The increased cost of SOAPs is due to the computation of the REKernel.

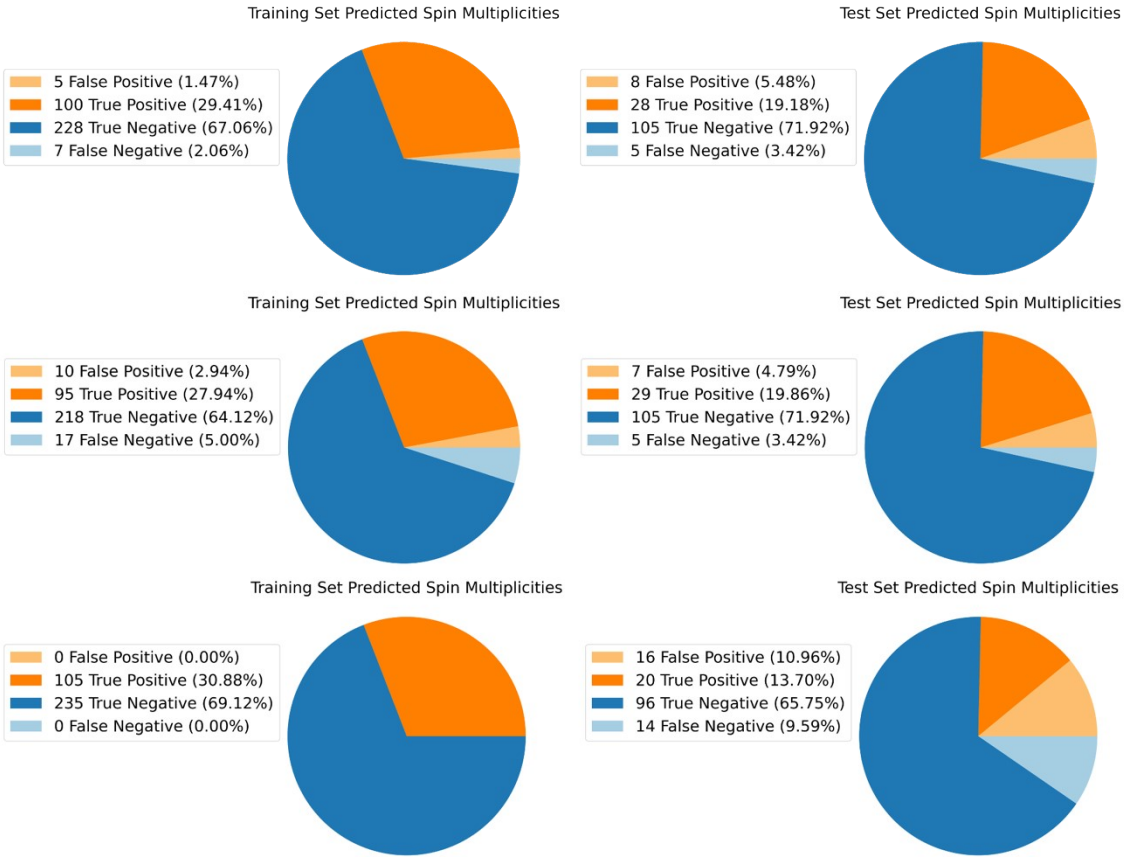
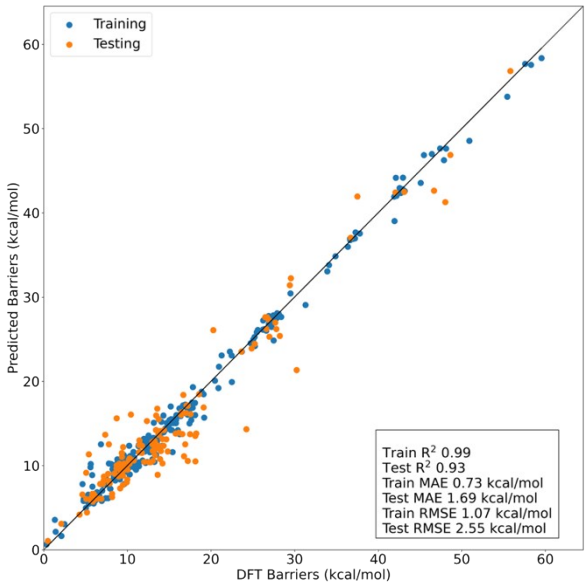
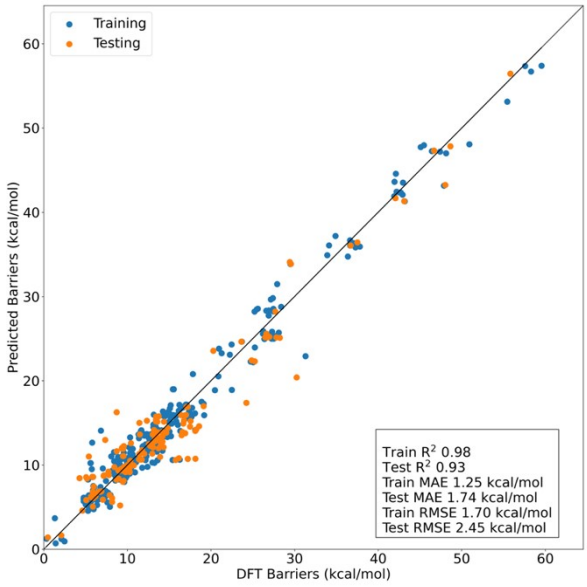


Figure S5: The PI (top), SOAP (middle), and CM (bottom) molecular representations were used to predict the spin-states of each molecule in the DFT data. The CM model performs the worst since it overfits data severely. SOAPs and PIs show very similar performance, but PIs outperform SOAPs.


Training
Testing
Predicted Barriers (kcal/mol)
DFT Barriers (kcal/mol)
Train R^2 0.99
Test R^2 0.93
Train MAE 0.73 kcal/mol
Test MAE 1.69 kcal/mol
Train RMSE 1.07 kcal/mol
Test RMSE 2.55 kcal/mol


Training
Testing
Predicted Barriers (kcal/mol)
DFT Barriers (kcal/mol)
Train R^2 0.98
Test R^2 0.93
Train MAE 1.25 kcal/mol
Test MAE 1.74 kcal/mol
Train RMSE 1.70 kcal/mol
Test RMSE 2.45 kcal/mol

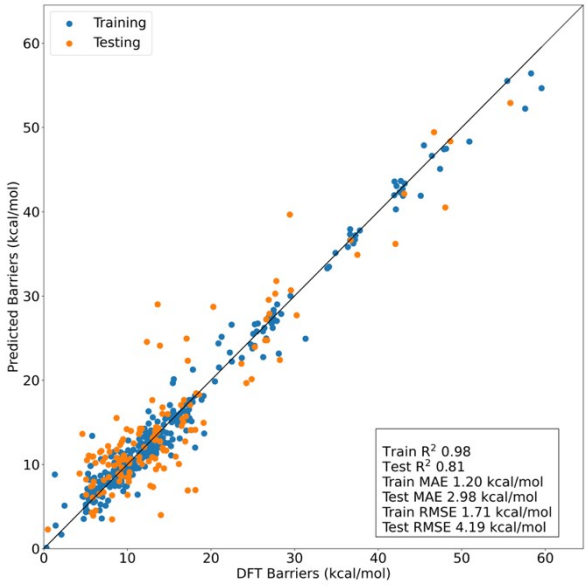

Training
Testing
Predicted Barriers (kcal/mol)
DFT Barriers (kcal/mol)
Train R^2 0.98
Test R^2 0.81
Train MAE 1.20 kcal/mol
Test MAE 2.98 kcal/mol
Train RMSE 1.71 kcal/mol
Test RMSE 4.19 kcal/mol

Figure S6: The prediction of the activation barriers using PIs (top), SOAPs (middle), and CMs (bottom) shows that CMs is the least accurate method for predicting barriers, whereas the SOAPs and PIs method perform very similarly.

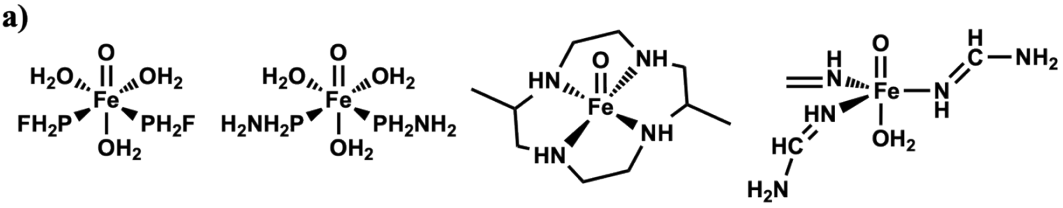
TRAIN					TEST				
PIs	precision	recall	f1-score	support	PIs	precision	recall	f1-score	support
quintet	0.97	0.98	0.97	233	quintet	0.95	0.93	0.94	113
triplet	0.95	0.93	0.94	107	triplet	0.78	0.85	0.81	33
accuracy			0.96		accuracy			0.91	
macro avg	0.96	0.96	0.96	340	macro avg	0.87	0.89	0.88	146
weighted avg	0.96	0.96	0.96	340	weighted avg	0.91	0.91	0.91	146
SOAPs	precision	recall	f1-score	support	SOAPs	precision	recall	f1-score	support
quintet	0.93	0.96	0.94	228	quintet	0.95	0.94	0.95	112
triplet	0.90	0.85	0.88	112	triplet	0.81	0.85	0.83	34
accuracy			0.92		accuracy			0.92	
macro avg	0.92	0.90	0.91	340	macro avg	0.88	0.90	0.89	146
weighted avg	0.92	0.92	0.92	340	weighted avg	0.92	0.92	0.92	146
CMs	precision	recall	f1-score	support	CMs	precision	recall	f1-score	support
quintet	1.00	1.00	1.00	235	quintet	0.87	0.86	0.86	112
triplet	1.00	1.00	1.00	105	triplet	0.56	0.59	0.57	34
accuracy			1.00		accuracy			0.79	
macro avg	1.00	1.00	1.00	340	macro avg	0.71	0.72	0.72	146
weighted avg	1.00	1.00	1.00	340	weighted avg	0.80	0.79	0.80	146

Table S5*:* A classification report for PIs, SOAPs, and CMs where the evaluations metrics include those mentioned above.

Table S5 shows the accuracy, macro and weighted average of the precision, recall, and F1-scores for the classification of the spin states using PIs, SOAPs, and CMs. As expected, CMs is more prone to overfitting than SOAPs and PIs and exhibits lower accuracies on the test set. While similar results can be achieved with SOAPs, the parameters we use for the PIs offer a faster method than SOAPs as seen in Table S5. Overall, while the SOAP molecular representation offers a competitive representation with respect to accuracy, due to the speed of which a PI can be generated, PIs offer a favorable representation for high throughput ML studies, as we have demonstrated herein.

S6. Density Functional Theory Validation of Machine Learning Model

a)



b)

Structure	DFT (kcal/mol)	Predicted (kcal/mol)	Predicted Spin	DFT Spin
706-aq	24.23	14.32	Triplet	Triplet
884-aq	30.25	21.35	Quintet	Triplet
155-j	18.15	10.50	Quintet	Triplet
554-k	8.72	15.62	Quintet	Quintet

Figure S7: (a) Structures of the four outliers. (b) Values of the true and predicted barriers, in kcal/mol.

Label	DFT (kcal/mol)	ML (kcal/mol)	Absolute Error (kcal/mol)
985-(**1**)	7.5	8.1	0.6
2768-(**1**)	9.6	10.8	1.2
338-(**1**)	6.6	8.1	1.5
515-(**2**)	11.3	11.1	0.2
1075-(**2**)	10.8	11.1	0.3
1178-(**2**)	11.4	11.1	0.3
530-(**2**)	11.5	11.1	0.4
2003-(**2**)	11.6	11.1	0.5
525-(**2**)	12	11.1	0.9
732-(**4**)	8.6	9.1	0.5
447-(**4**)	8.6	11.2	2.6
727-(**4**)	8.6	11.2	2.6
110-(**27**)	12	16.7	4.7
3494-(**30**)	7.3	15.2	7.9
5234-(**30**)	6.6	15.2	8.6

Table S6: Data used in validation of the machine learning model. Labels indicate a unique structure given name within the database. All reaction barriers in kcal/mol.

References

1. J. Townsend, C. P. Micucci, J. H. Hymel, V. Maroulas and K. D. Vogiatzis, Representation of molecular structures with persistent homology for machine learning applications in chemistry, *Nature Communications*, 2020, **11**, 3230.
2. L. Himanen, M. O. J. Jager, E. V. Morooka, F. F. Canova, Y. S. Ranawat, D. Z. Gao, P. Rinke and A. S. Foster, DScribe: Library of descriptors for machine learning in materials science, *Computer Physics Communications*, 2020, **247**, 106949-106949.
3. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, Scikit-learn: Machine Learning in Python, *Journal of Machine Learning Research*, 2011, **12**, 2825-2830.