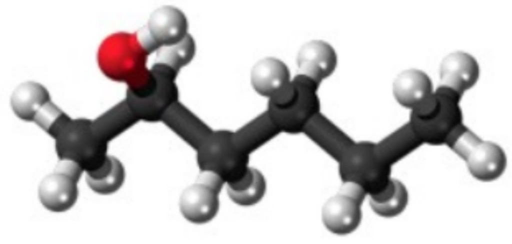


Lab ML for Data Science: Project II
Getting Insights into Quantum-Chemical Relations

Jan Jascha Jestel (5547158)

Mustafa Suman (5564676)

Gabriele Inciuraite (5208806)

Introduction

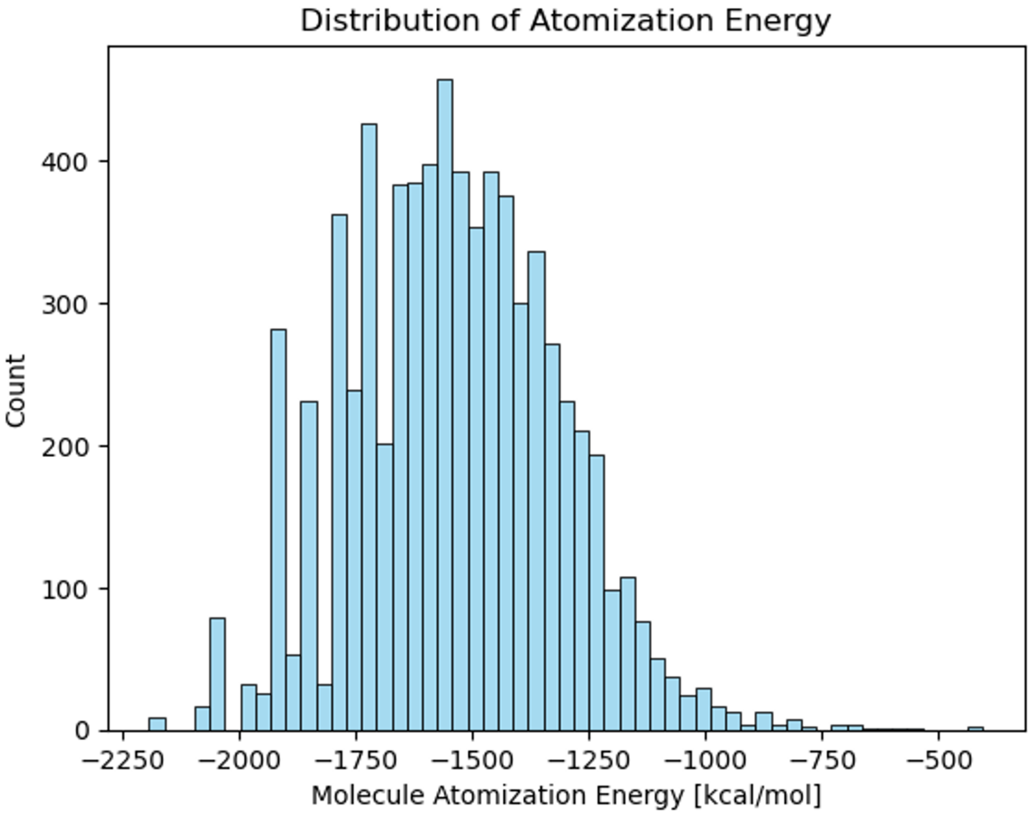
Task: investigate the relation between a molecule's geometry and its atomization energy

Dataset: 7165 molecules, each consisting of up to 23 atoms [H, C, N, O, S]

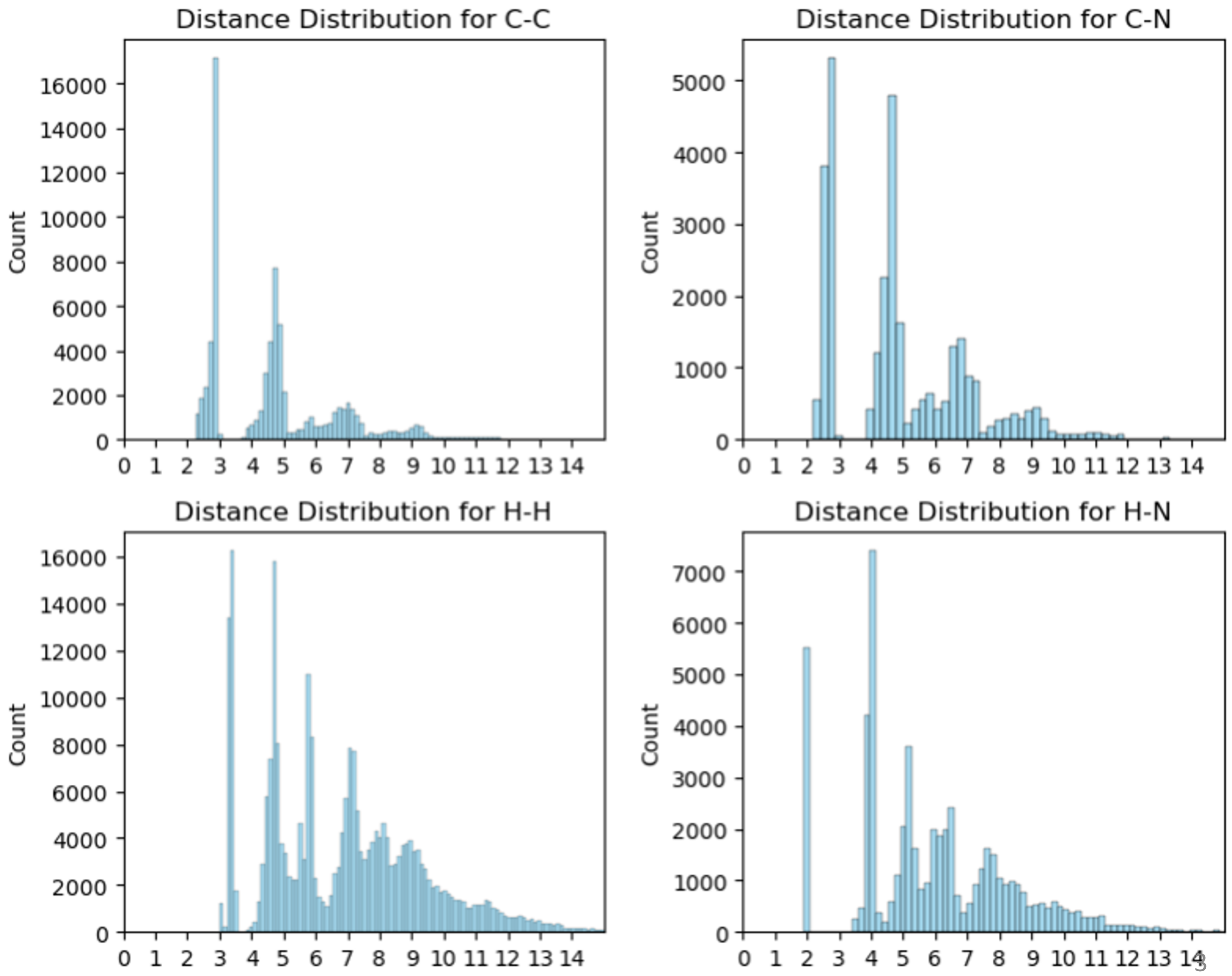
with atom coordinates and atom types as features, atomisation energy as target

Purpose of ML approach: computationally effective, identify what is needed in the molecular structure to be able to predict, and, using Explainable AI, what exact substructures are being used for prediction

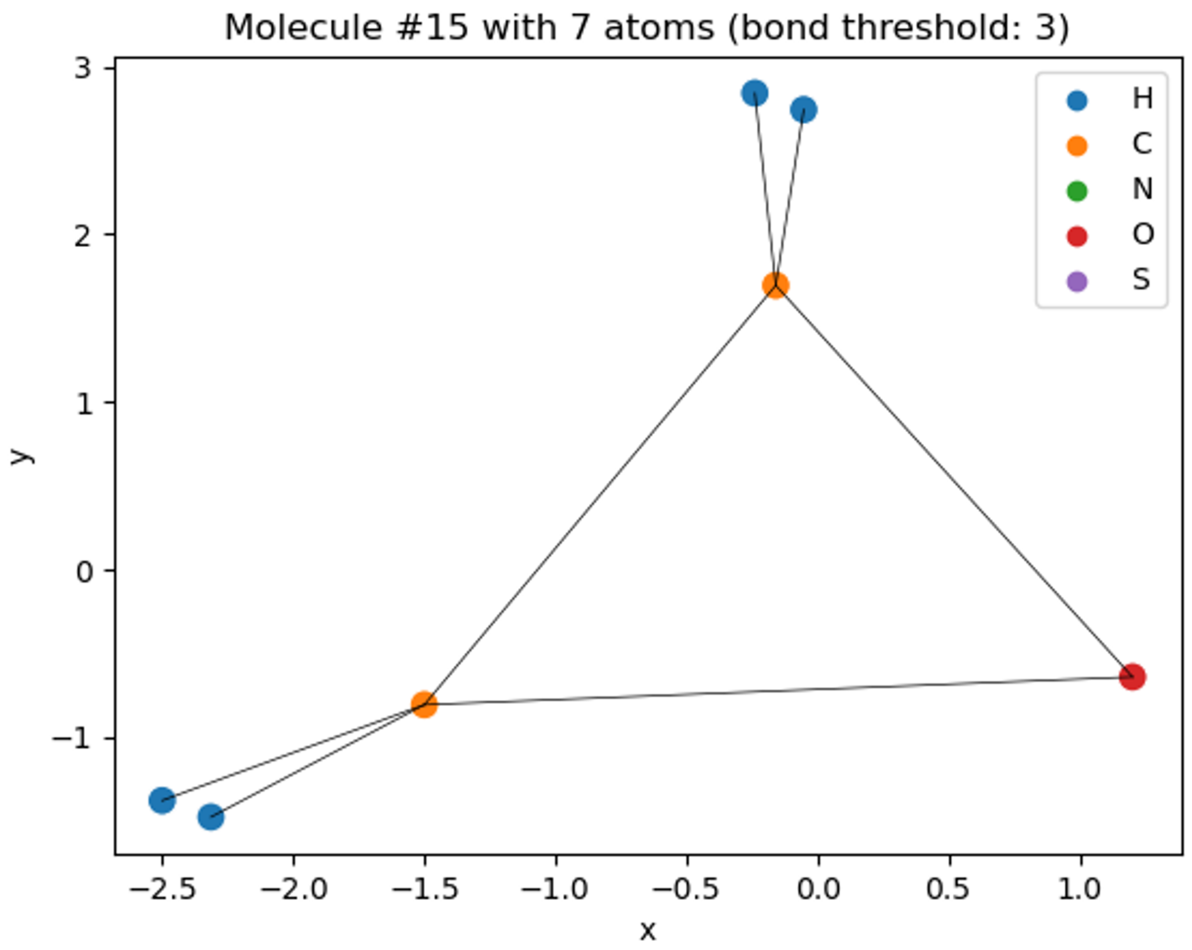
First look at the data

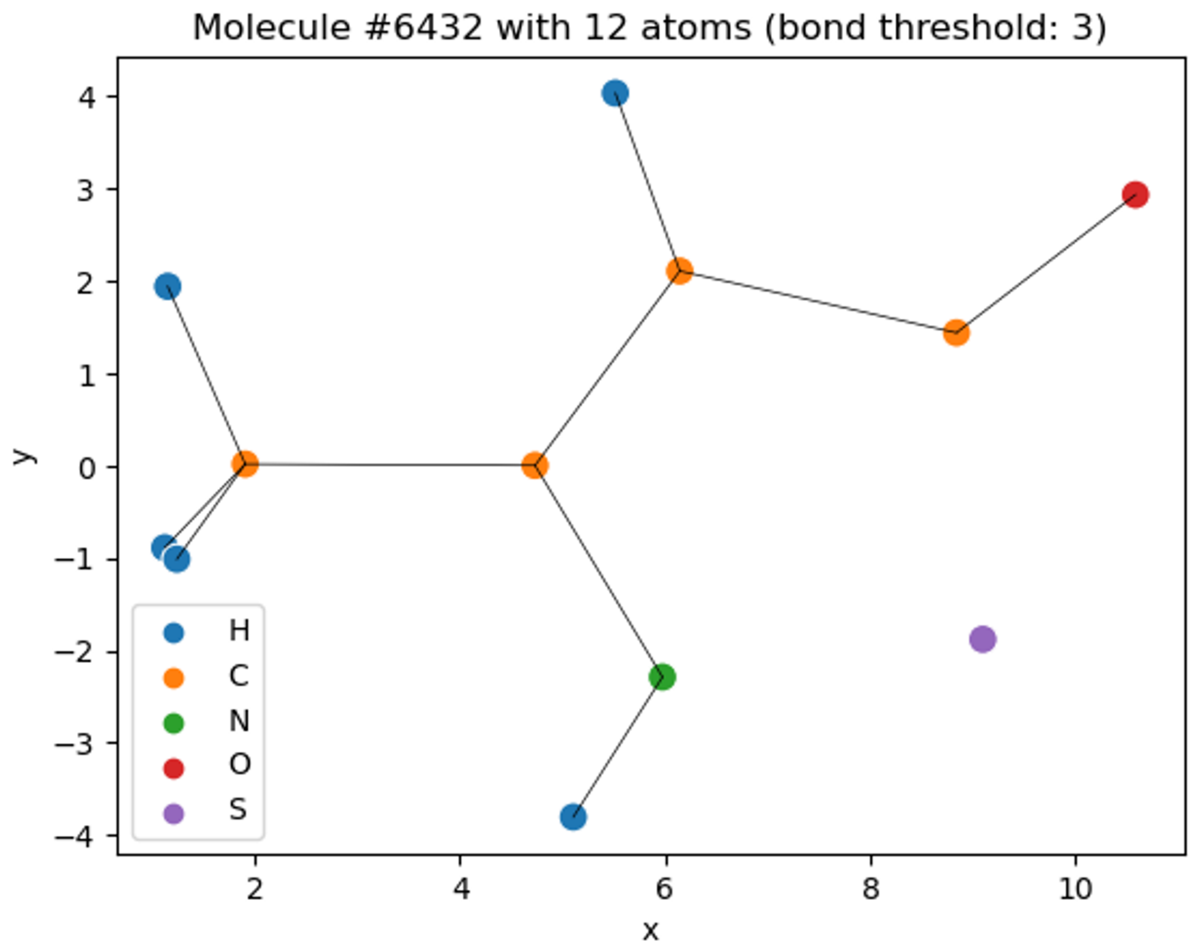


Distribution of intra-molecule euclidean distances of atoms by atom types

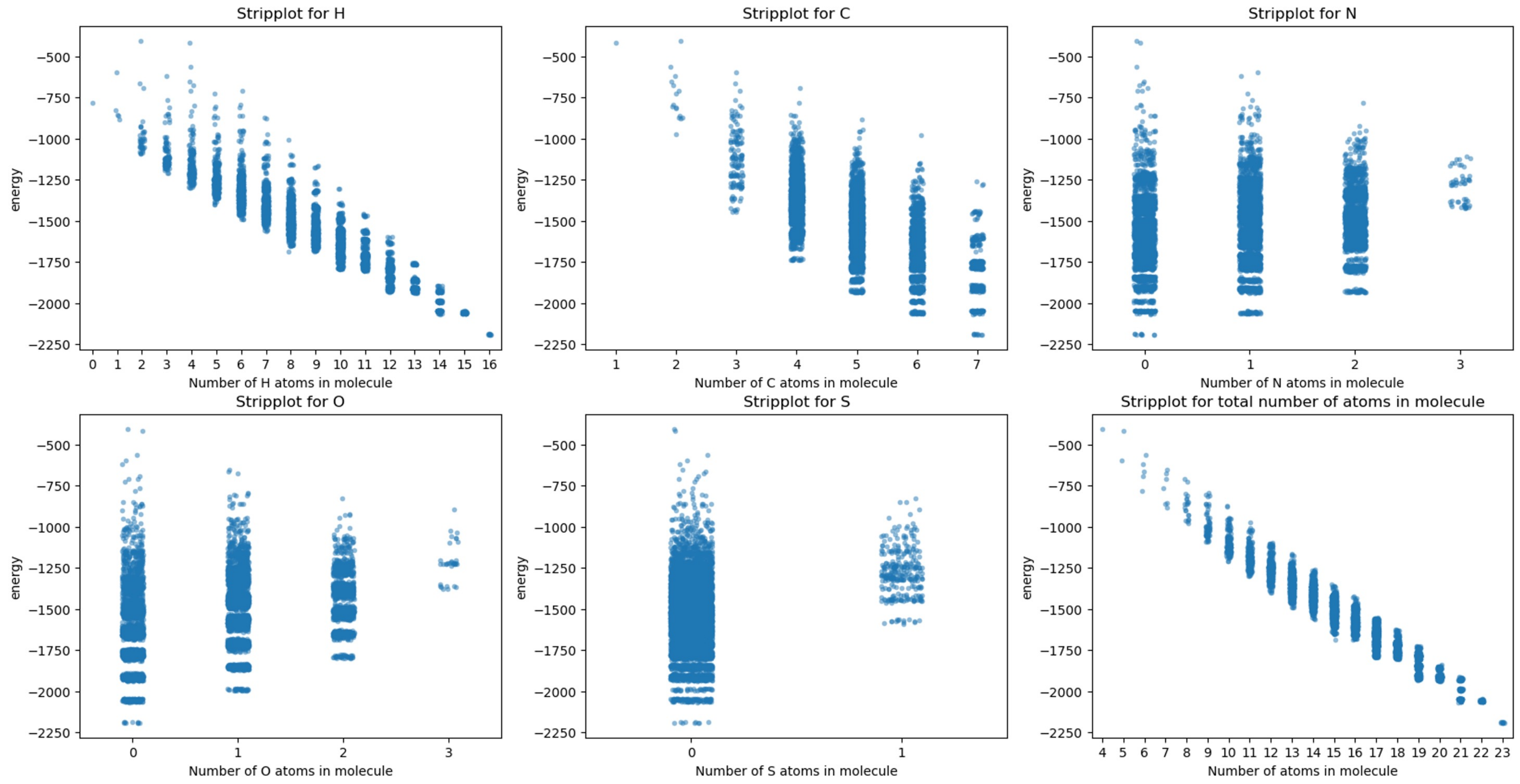


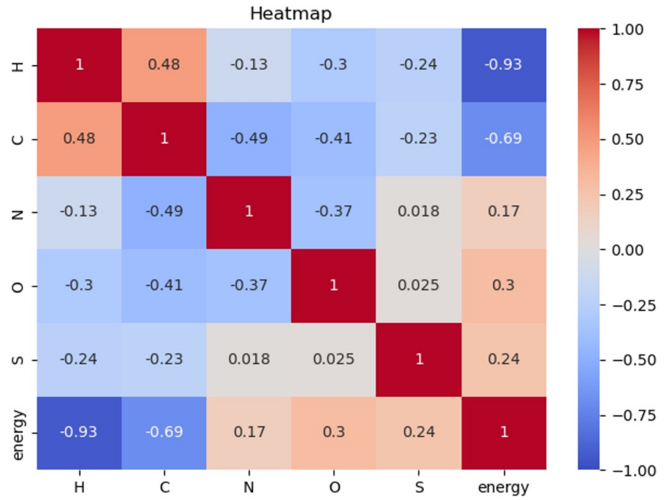
Molecule visualisation

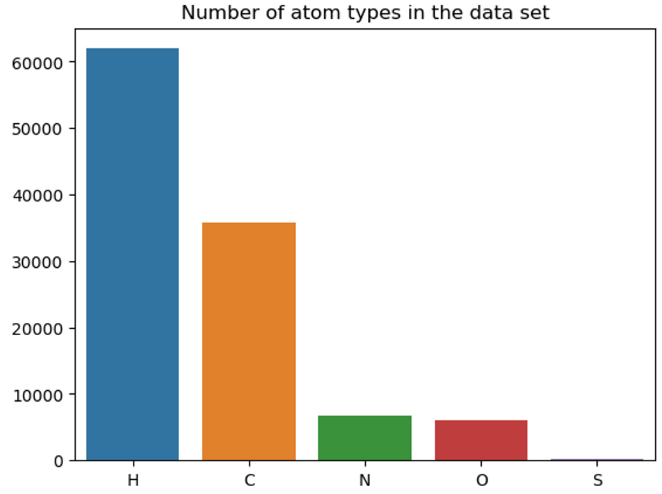




Atom-type-count molecule representation [#H, #C, #N, #O, #S]

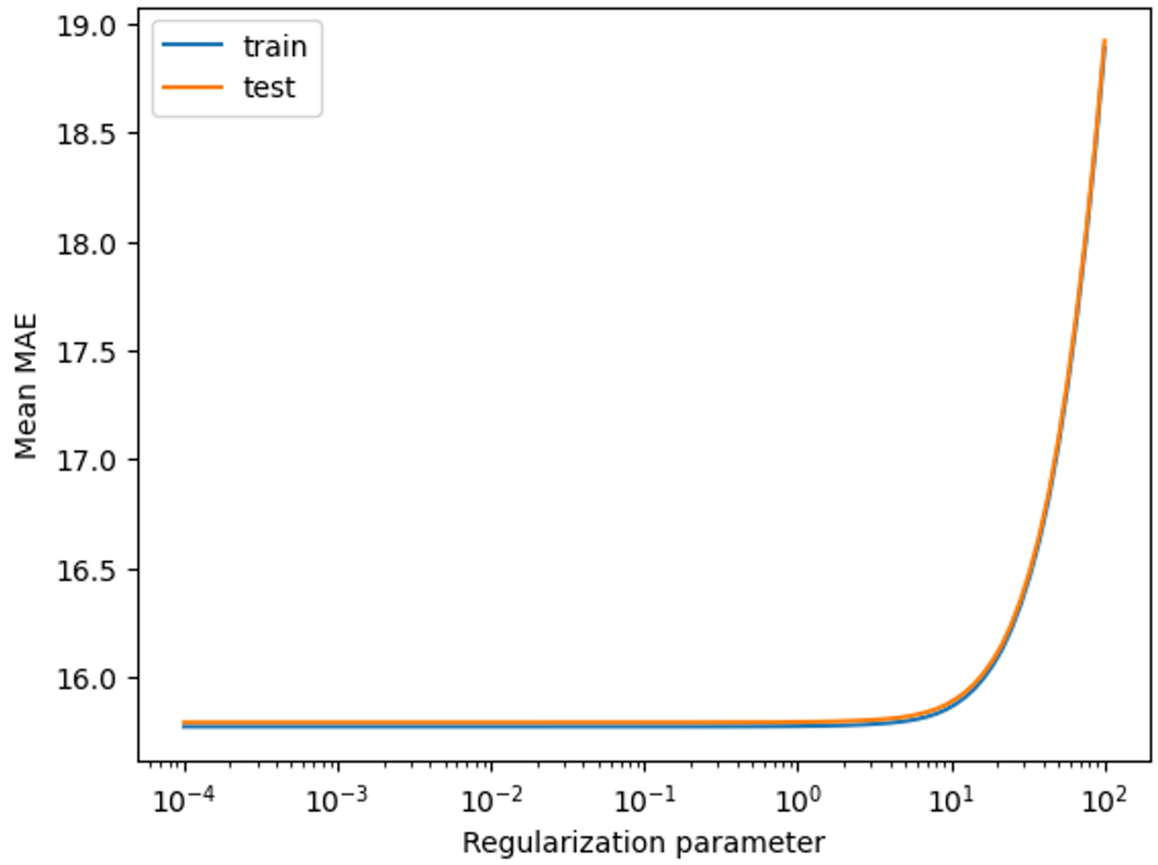




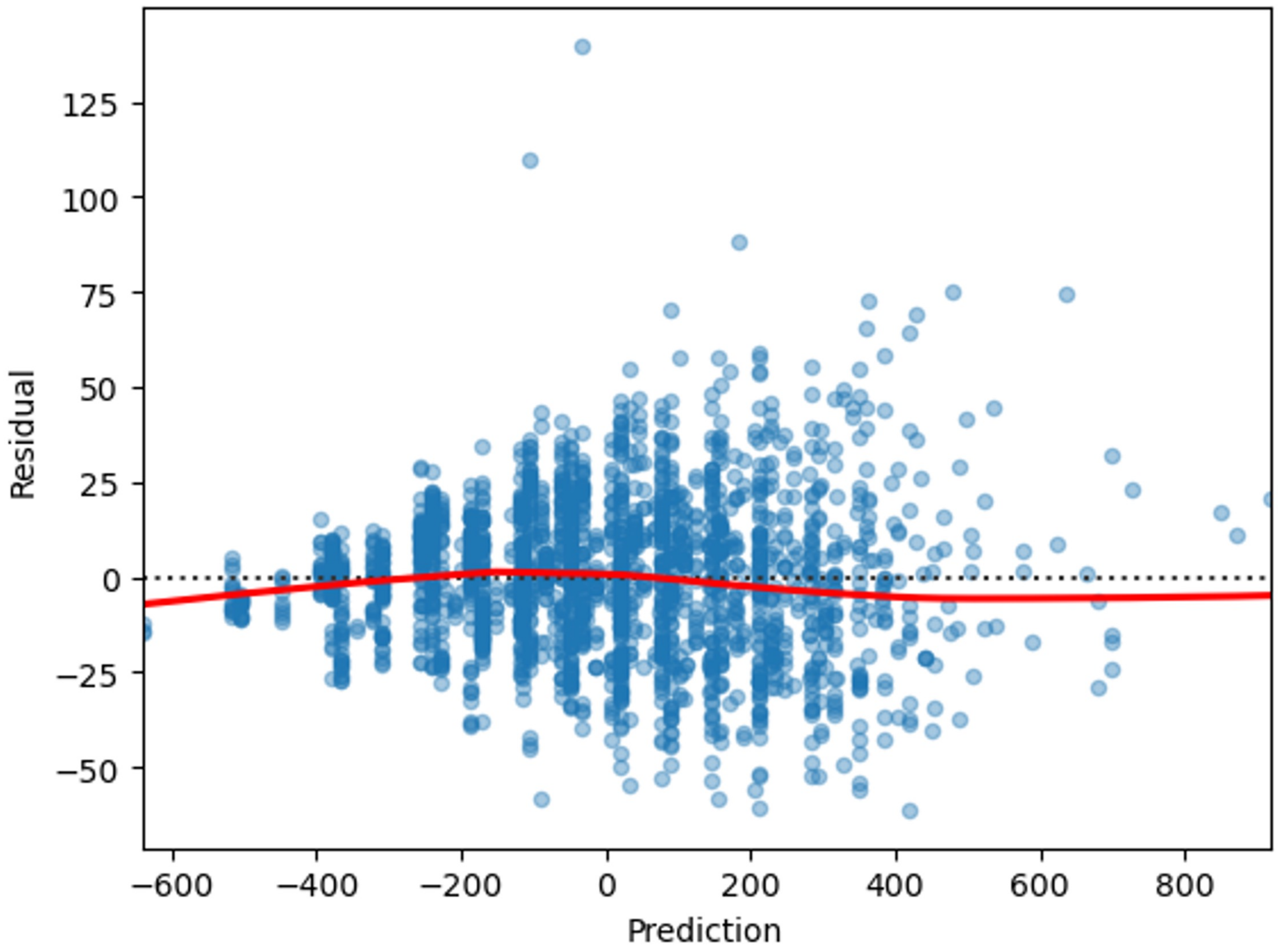


Ridge regression on atom-type-count representation

- split simple representation data in train and test
- center data
- apply grid search with 10-fold cross validation to tune regularization parameter



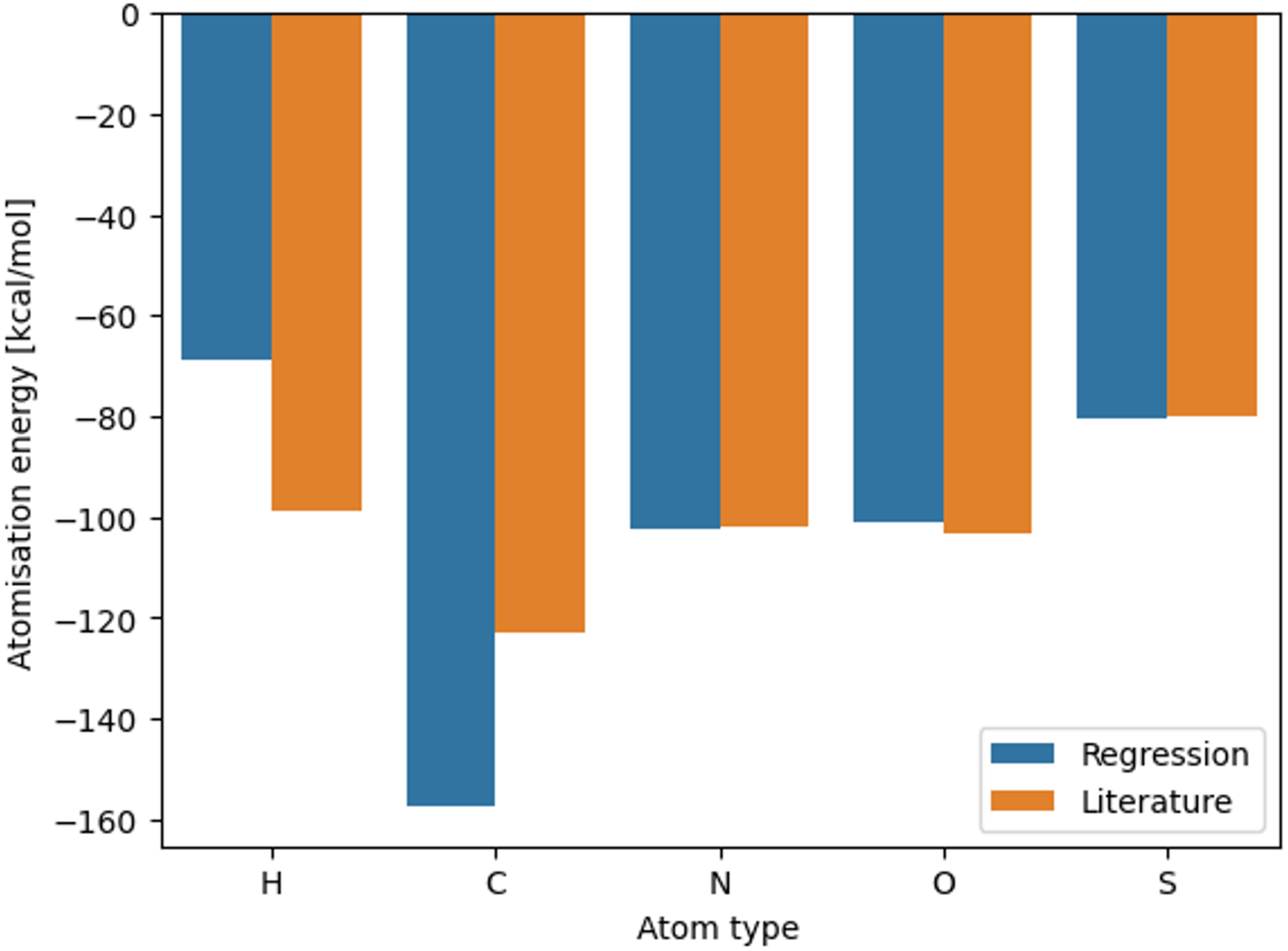
Analyse best model's performance




```
Best parameter: {'alpha': 0.0001}
R2: 0.992
MAE: 15.461
MSE: 403.582
```

Insights compared to literature

Mean bonding energy according to literature vs. ridge regression weights



Abhängigkeit der mittleren Bindungsenergie von der Bindungslänge[4]

Bindungslänge d in pm, Bindungsenthalpie ΔH in kJ/mol

Halogene untereinander			mit Wasserstoff			mit Kohlenstoff			mit Sauerstoff			gleiches Element		
Bindung	ΔH	d	Bindung	ΔH	d	Bindung	ΔH	d	Bindung	ΔH	d	Bindung	ΔH	d
F–F	159	142	H–C	413	108	C–H	413	108	O=N	607		C–C	348	154
Cl–Cl	242	199	H–O	463	97	C–O	358	143	O–N	201	136	C=C	614	134
Br–Br	193	228	H–N	391	101	C=O	745	122	O–P	335	154	C≡C	839	120
I–I	151	267	H–P	322	142	C–N	305	147	O=S	420	143	H–H	436	74
Br–Cl	219	214	H–S	367	134	C=N	615	130	O–F	193	142	N–N	163	146
Br–F	249	176	H–F	567	92	C≡N	891	116	O–Cl	208	170	N=N	418	125
Br–I	178		H–Cl	431	128	C–P	264	184	O–Br	234		N≡N	945	110
Cl–F	253	163	H–Br	366	141	C–S	272	182	O–I	234		O–O	146	148
Cl–I	211	232	H–I	298	160	C=S	536	189				O=O	498	121
						C–F	489	138				P–P	172	221
						C–Cl	339	177				S–S	255	205
						C–Br	285	194						
						C–I	218	214						

Pairs-of-atoms-based representation

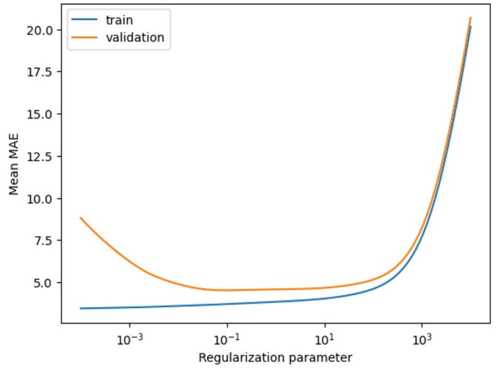
- pairs encoded as type x distance matrix
- prepare array of all pairs to generate representations
- from literature min and max bond length
- grid search with 5-fold CV to tune alpha for ridge regression
- wrapped in grid search to tune the no. distance intervals and the standard deviation for the soft encoding

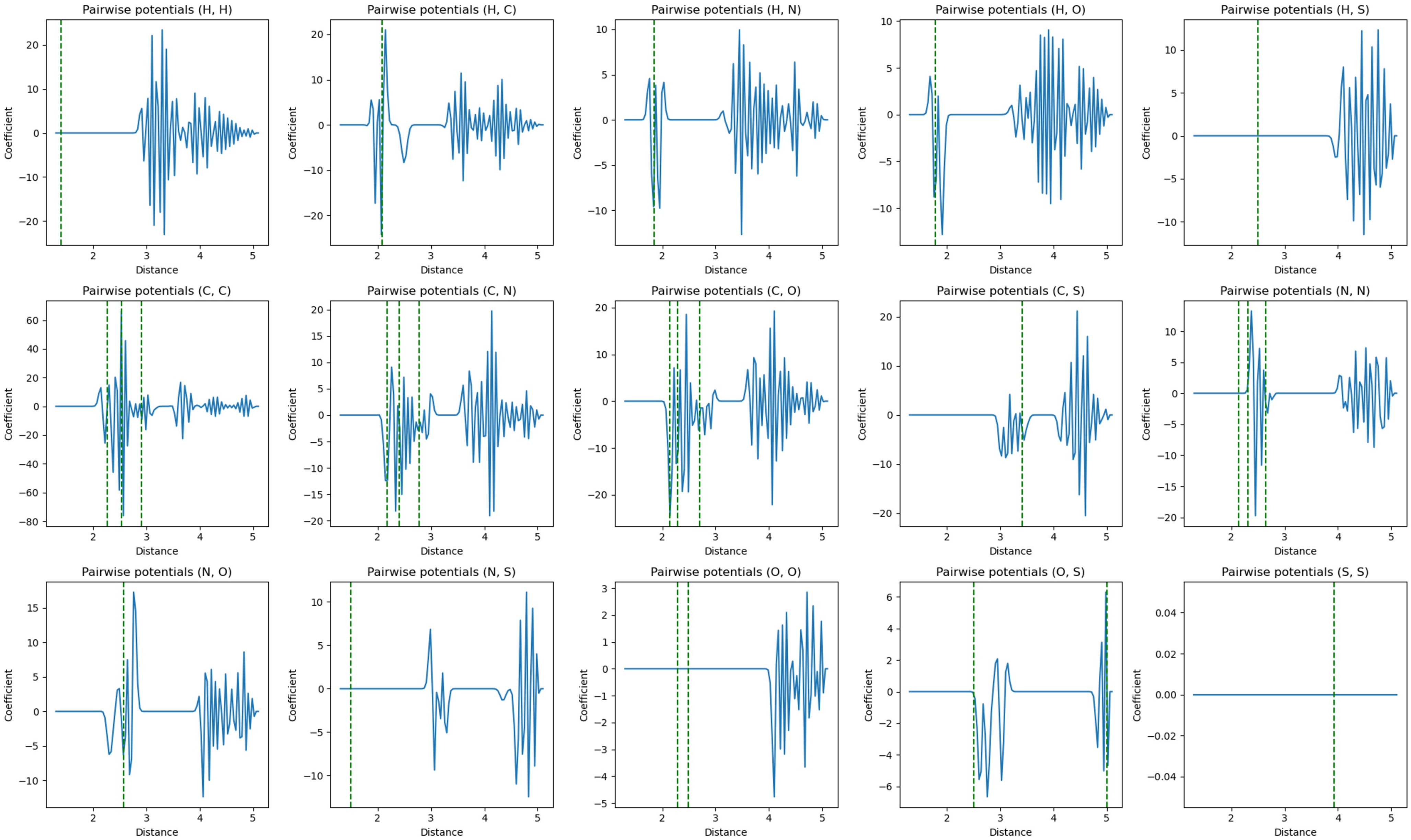
```
param_grid = {
    "M": [10, 40, 70, 100],
    "STD": np.linspace(0.05, 0.15, 10),
    "THETA_1": [1.3],
    "THETA_M": [5.1],
}
```

regularization parameter α between 0.0001 and 1000 (log-space)

	M	STD	THETA_1	THETA_M	model_idx	best_score	best_alpha	n_coef_smaller_1e-4
30	100	0.050000	1.3	5.1	30	-4.526262	0.117681	1131
31	100	0.061111	1.3	5.1	31	-4.669757	0.010476	1060
21	70	0.061111	1.3	5.1	21	-4.796374	0.117681	774
20	70	0.050000	1.3	5.1	20	-4.802526	0.247708	824
32	100	0.072222	1.3	5.1	32	-4.853629	0.038535	1045
22	70	0.072222	1.3	5.1	22	-4.887790	0.022051	734
33	100	0.083333	1.3	5.1	33	-4.939198	0.018307	1010
23	70	0.083333	1.3	5.1	23	-4.962279	0.012619	715
34	100	0.094444	1.3	5.1	34	-5.024969	0.004132	963
24	70	0.094444	1.3	5.1	24	-5.034173	0.002848	682
25	70	0.105556	1.3	5.1	25	-5.104097	0.000443	638
35	100	0.105556	1.3	5.1	35	-5.114303	0.000643	922
36	100	0.116667	1.3	5.1	36	-5.168880	0.007221	985
26	70	0.116667	1.3	5.1	26	-5.178351	0.004977	690
37	100	0.127778	1.3	5.1	37	-5.185714	0.002848	980
27	70	0.127778	1.3	5.1	27	-5.187397	0.001963	684
28	70	0.138889	1.3	5.1	28	-5.206526	0.000933	659
38	100	0.138889	1.3	5.1	38	-5.209823	0.001630	937
29	70	0.150000	1.3	5.1	29	-5.242591	0.000254	645
39	100	0.150000	1.3	5.1	39	-5.245416	0.000443	924
15	40	0.105556	1.3	5.1	15	-5.248717	0.015199	419
16	40	0.116667	1.3	5.1	16	-5.250616	0.004977	406
14	40	0.094444	1.3	5.1	14	-5.252231	0.081113	444
17	40	0.127778	1.3	5.1	17	-5.255090	0.001963	407
13	40	0.083333	1.3	5.1	13	-5.257568	0.247708	466
19	40	0.150000	1.3	5.1	19	-5.258480	0.000120	380
18	40	0.138889	1.3	5.1	18	-5.261892	0.000443	391
12	40	0.072222	1.3	5.1	12	-5.265598	0.628029	487
11	40	0.061111	1.3	5.1	11	-5.277614	1.592283	501
10	40	0.050000	1.3	5.1	10	-5.474154	4.862602	499
9	10	0.150000	1.3	5.1	9	-10.059917	0.000100	116
8	10	0.138889	1.3	5.1	8	-14.591495	0.000100	120
7	10	0.127778	1.3	5.1	7	-20.071752	0.000100	120
6	10	0.116667	1.3	5.1	6	-26.477178	0.055908	120
5	10	0.105556	1.3	5.1	5	-31.582763	0.038535	120
4	10	0.094444	1.3	5.1	4	-35.941278	0.012619	118
3	10	0.083333	1.3	5.1	3	-40.676922	0.002364	116
2	10	0.072222	1.3	5.1	2	-46.813600	0.000100	118
1	10	0.061111	1.3	5.1	1	-56.015233	0.038535	117
0	10	0.050000	1.3	5.1	0	-63.536613	0.000305	117

Best model in terms of MAE (soft encoding)



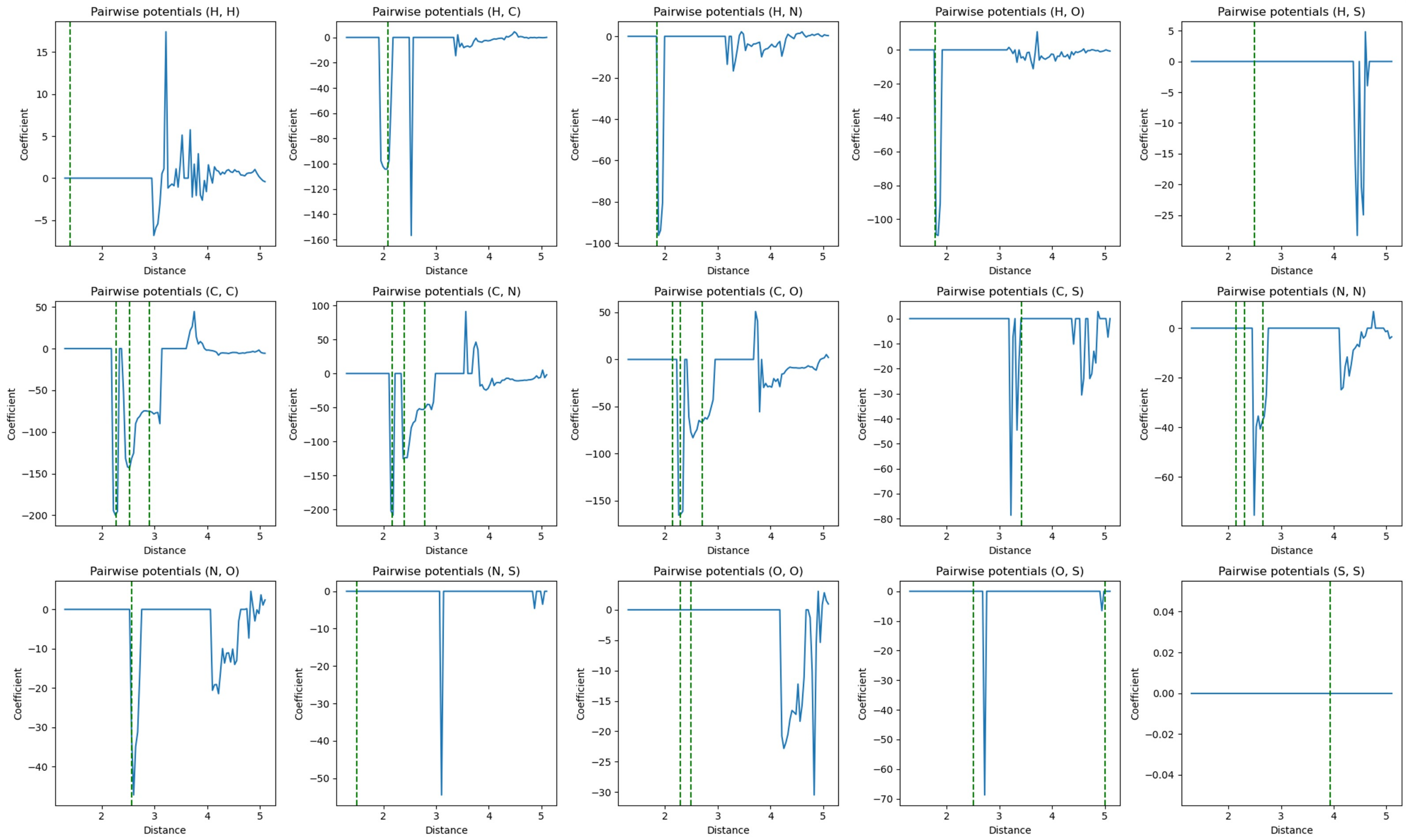


Parameters:

- $M = 100$
- $\theta_1 = 1.3$
- $\theta_M = 5.1$
- $\sigma = 0.05$
- $\alpha = 0.117$

```
R2: 0.997
MAE: 4.446
MSE: 128.539
#coef < 1e-10: 1069
#coef < 1e-50: 921
#coef == 0: 173
```

Same parameters with hard encoding

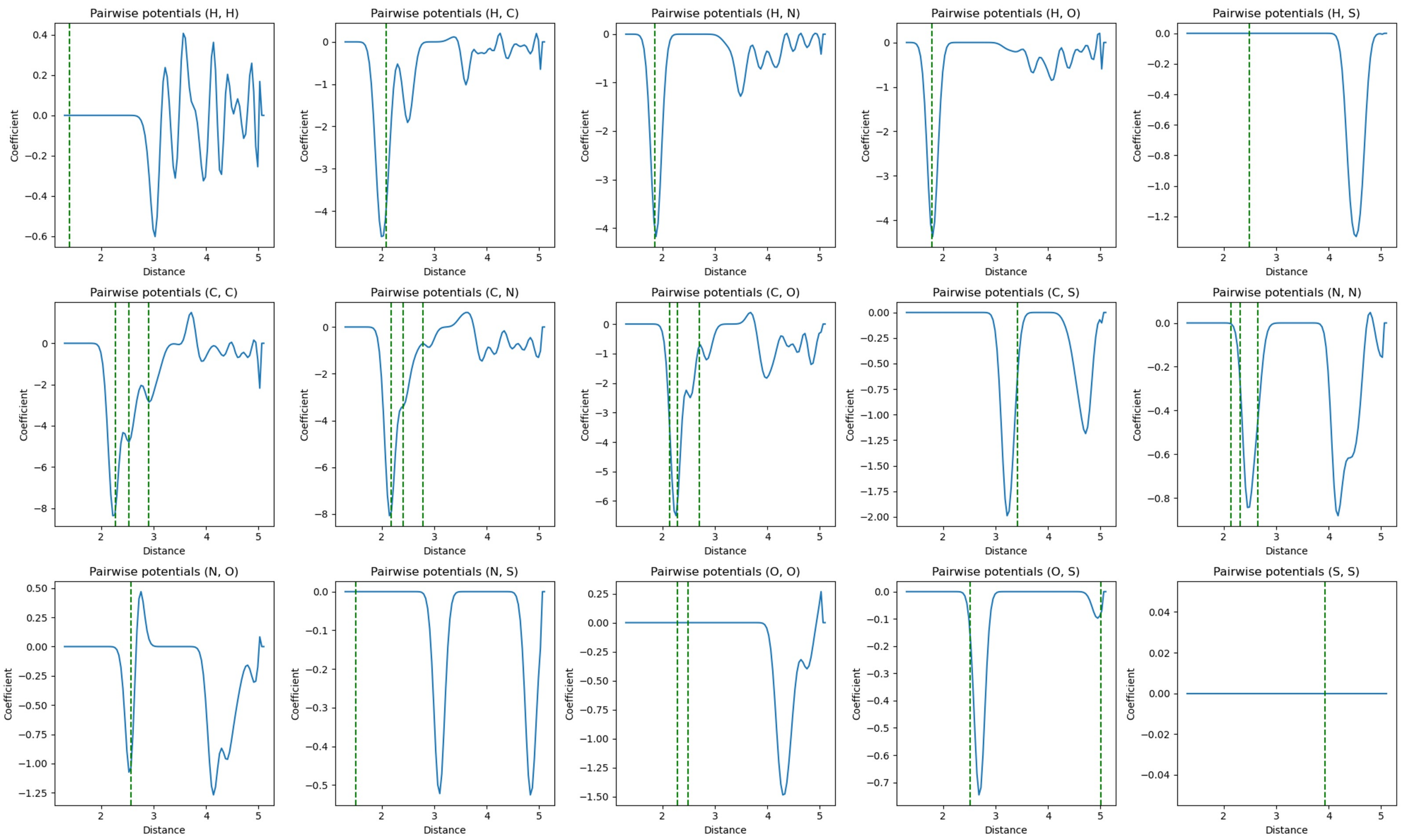


Parameters:

- $M = 100$
- $\theta_1 = 1.3$
- $\theta_M = 5.1$
- $\sigma = 0.05$
- $\alpha = 0.117$

```
R2: 0.966
MAE: 18.075
MSE: 1817.125
#coef < 1e-10: 1396
#coef < 1e-50: 1396
#coef == 0: 1022
```

Regularize more to improve interpretability



Parameters

- $M = 100$
- $\theta_1 = 1.3$
- $\theta_M = 5.1$
- $\sigma = 0.1$
- $\alpha = 1000$

```
R2: 0.976
MAE: 20.198
MSE: 1298.012
#coef < 1e-10: 1384
#coef < 1e-50: 1383
#coef == 0: 128
```

Notebook modifications

- Improved visualizations
- No bias/intercept in ridge regression