Electronic Supplementary Information (ESI)

Capacity-prediction models for organic anode active materials of lithium-ion battery: Advances in the predictors using small data

Haruka Tobita,[a] Yuki Namiuchi,[b] Takumi Komura,[a] Hiroaki Imai,[a] Koki Obinata,[c] Masato Okada,[c] Yasuhiko Igarashi,*[,b] Yuya Oaki*[,a]

[a]Department of Applied Chemistry, Faculty of Science and Technology, Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama 223-8522, Japan

[b] Faculty of Engineering, Information and Systems, University of Tsukuba, 1-1-1 Tennodai, Tsukuba 305-8573, Japan

[c] Graduate School of Frontier Sciences, The University of Tokyo, 5-1-5 Kashiwanoha, Kashiwa 277-8561, Japan

E-mail: oakiyuya@applc.keio.ac.jp

Contents

Methods

Preparation of datasets. The training and test datasets were prepared using the data in our previous works and were summarized in Tables S1–S6, S8, and S9.[48,49] The training data for compounds **19**, **25**, **36**, **37**, **41**, **46**, **49**, **50**, **53**, and **54** were collected in the present work as follows.

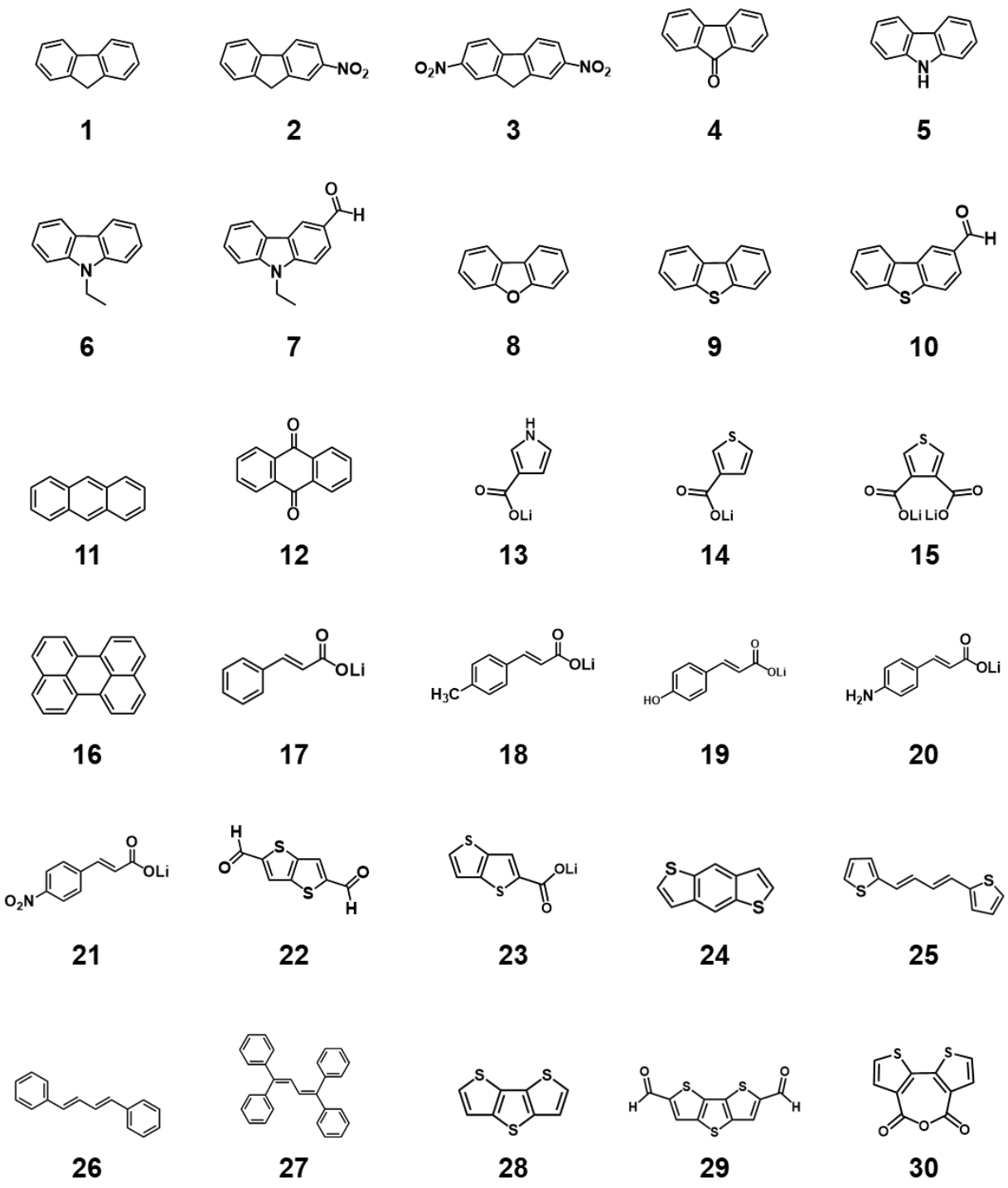
The specific capacity of compounds **19**, **25**, **36**, **37**, **41**, **46**, **49**, **50**, **53**, and **54** was measured using three-electrode setup in a twin-beaker cell. The mixture of these active material (30 wt%), acetylene black conductive carbon (60 wt %), and poly(vinylidene fluoride) (PVDF) binder (10 wt%) was prepared with a trace amount of *N*-methyl-2-pyrrolidone (NMP, Junsei Chemical, 99%) and then pasted on a copper mesh as a current collector. After evaporation of NMP, ca. 2 mg of the paste was loaded as the working electrode. The reference and counter electrodes were both metallic lithium. The electrolyte solution was ethylene carbonate (EC) and diethyl carbonate (DEC) (1/1 by volume) containing 1 mol dm^{-3} lithium perchlorate ($LiClO_4$). Galvanostatic charge-discharge measurement was performed using a chronopotentiometry (Hokuto Denko, HJ1001SD8) with a cut-off voltage 0.01–3.0 V vs. Li/Li^+ at 100 mA g^{-1}. The capacity was corrected to subtract the capacity of the conductive carbon (See Fig. S1). The explanatory variables for these compounds were calculated by density functional theory (DFT) using Gaussian 16 under B3LYP functional and 6–311G basis set (x_n, with the note *a* in Table 1) and Hansen Solubility Parameters in Practice (HSPiP, version 5.0.03) (x_n, with the note *b* in Table 1).

Machine learning. The descriptors of the model G3 were extracted by ES-LiR.[10–14,53] ES-LiR calculates the prediction accuracy of all the possible models based on the error by cross-validation for all 2^N combinations of variables, such as $\{x_1 \text{ only}\}$, $\{x_1, x_2\}$, $\{x_1, x_3\}$, $\{x_1, x_4\}$, …, $\{x_1, x_2, x_3\}$, $\{x_1, x_2, x_4\}$, … $\{x_1, x_2, \ldots, x_n\}$.[53] This exhaustive construction of the models was implemented in Python (ver. 3.7.6) and then the results are summarized in the weight diagram. In ES-BMA, we considered the uncertainty for all 2^N combinations of variables and introduced a method of quantitatively evaluating the confidence level of feature selection using a weighted average of the model posterior probabilities, which is called BMA.[S1] This method can evaluate the confidence level of feature selection and quantify the importance evaluation of features, which has been a qualitative one when using the weight diagram of the exhaustive search method. The summation over all combinations of indicator vectors can be calculated using the result of the exhaustive search, which is called as ES-BMA.[55] When this scheme is applied to the extraction of descriptors, we estimate the probability that i th variable ($i = 1, 2, \ldots, n$) is included in the descriptors by marginalizing over all combinations of variables including the i th variable.[55] The code used in the present work is available at https://github.com/okada-lab/exhbma. LASSO was performed by leave-one-out cross validation. When the reduced datasets were prepared, the data was randomly subtracted.

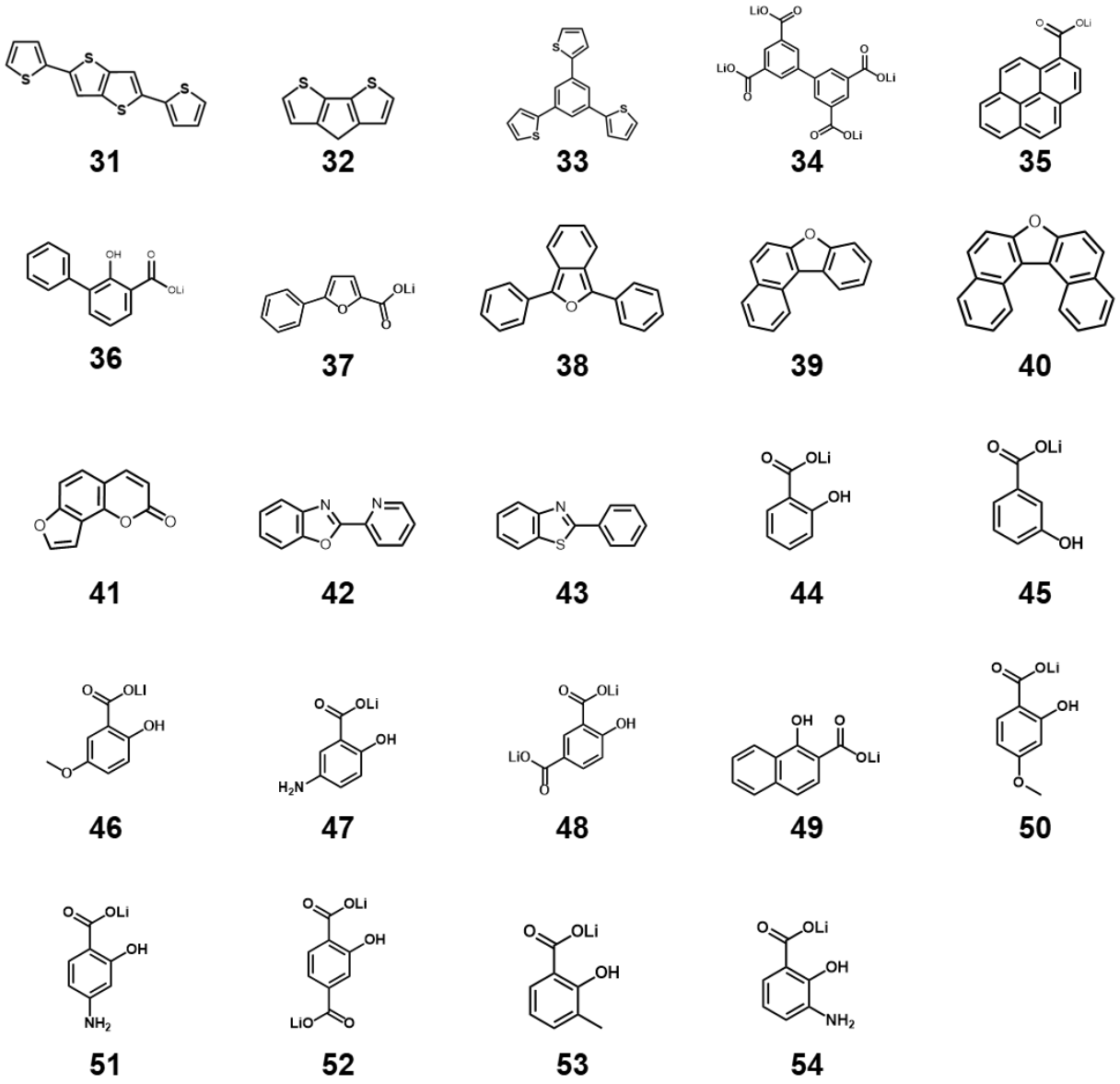
Additional Reference

S1. A. E. Raftery, D. Madigan and J. A. Hoeting, *J. Am. Stat. Assoc.*, 1997, **92**, 179.

Molecular structures of 1–54 for training data



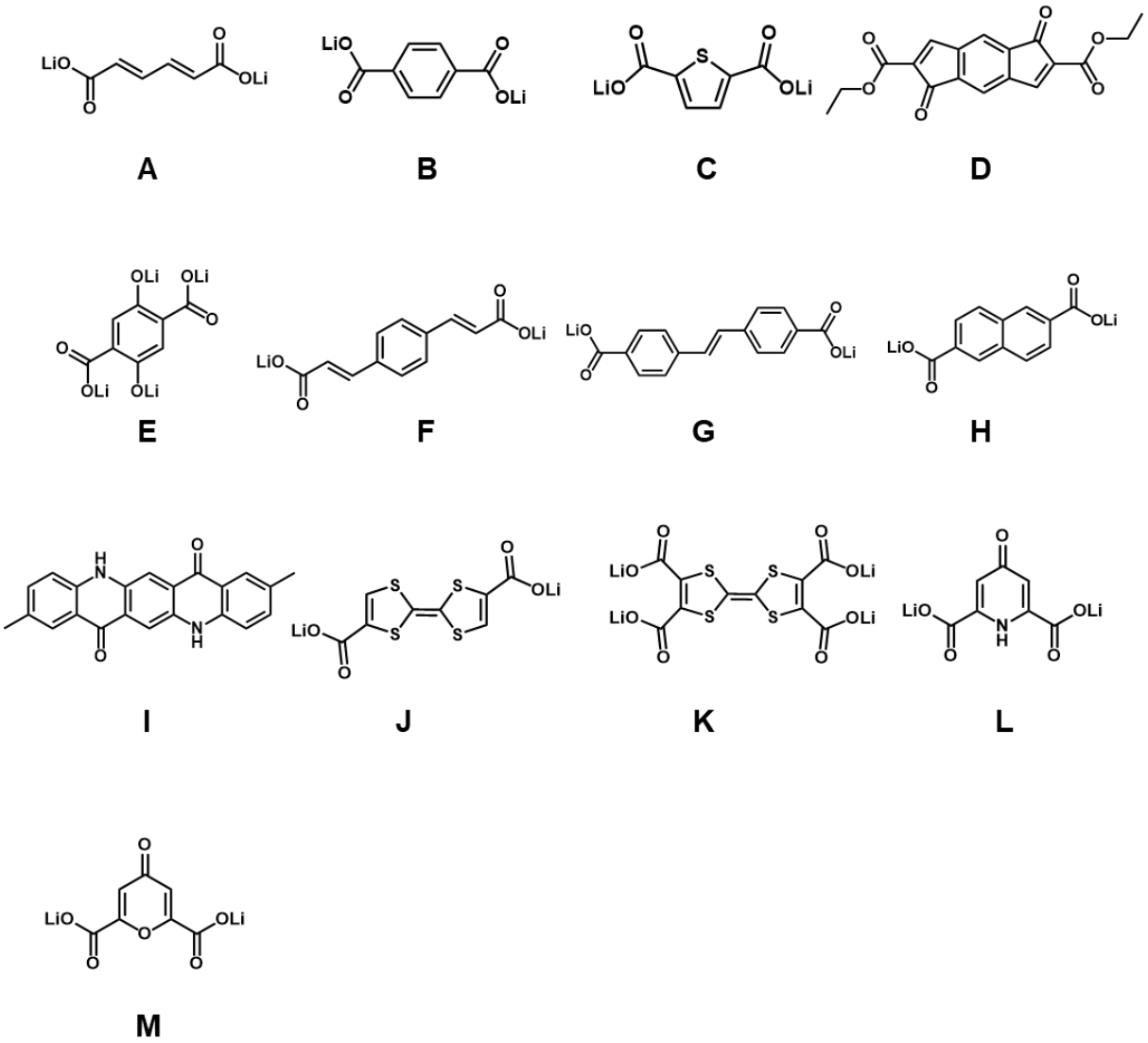
Scheme S1. Molecular structures of compounds **1**–**54** for the training.



Scheme S1 (continued). Molecular structures of compounds **1**–**54** for the training.

The reported specific capacity was summarized in Table 1.

Molecular structures of A–M for test data



Scheme S2. Molecular structures of compounds **A**–**M** for the test.

The reported specific capacity and literature number were summarized in Table 3.[21–29]

Additional data based on the charge-discharge measurement

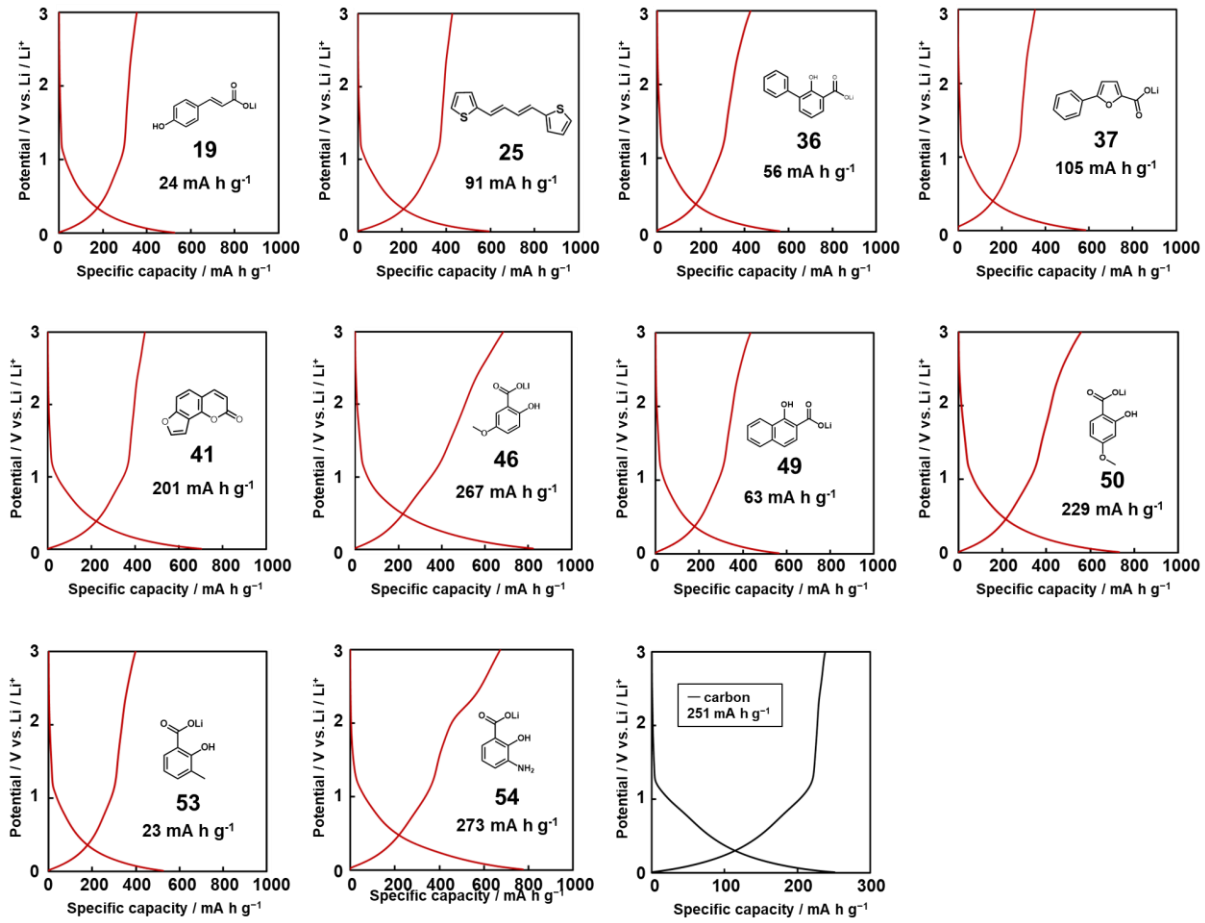


Fig. S1. Charge-discharge curves of the compounds **19**, **25**, **36**, **37**, **41**, **46**, **49**, **50**, **53**, and **54**.

The measured specific capacity was summarized in Table 1: 24 mA h g^{-1} (**19**), 91 mA h g^{-1} (**25**), 56 mA h g^{-1} (**36**), 105 mA h g^{-1} (**37**), 201 mA h g^{-1} (**41**), 267 mA h g^{-1} (**46**), 63 mA h g^{-1} (**49**), 229 mA h g^{-1} (**50**), 23 mA h g^{-1} (**53**), and 273 mA h g^{-1} (**54**).

The measured discharge capacity of the charge-discharge curve includes the capacity originating from conductive carbon. The corrected net capacity was estimated with the subtraction of the capacity originating from the conductive carbon using (eqn. S1),[49] the corrected actual specific capacity is C mA h g^{-1}, the measured specific capacity at 100 mA g^{-1} is C_{meas} mA h g^{-1}, the weight of the active material is W_a mg, the specific capacity of the conductive carbon at 50 mA g^{-1} is C_{AB} mA h g^{-1}, and weight of the conductive carbon is W_{CB} mg. As the weight ratio of the active material, AB, and binder was 3:6:1 for the high-throughput measurement, the specific capacity of the active material and conductive carbon was measured at 100 and 50 mA g^{-1}, respectively.

$$C = (C_{meas} \times W_a - C_{AB} \times W_{AB}) / W_a \quad \ldots \text{(eqn. S1)}$$

Training and test datasets G1

Table S1. Training dataset G1.

compounds	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	x14	x15	x16	x17	x19	x21	x22	x23	x24	x25	x26	y
1	-0.962	-0.469	0.0933	0.9331	1.3639	0.4931	1.055	1.8948	2.3255	0.5619	0.8398	1.4017	0.4308	1.2705	1.8325	166.22	161.241	0	12	10	2	1.4303	15.76	115	0
2	-3.015	-1.256	-1.063	-0.517	0.5363	1.7582	1.9514	2.4972	3.5509	0.1932	0.5459	0.7391	1.0536	1.5995	1.7927	211.22	126.8889	0	12	12	4	5.8514	11.36	155	0
3	-3.68	-3.09	-1.714	-1.371	-1.132	0.59	1.9658	2.3084	2.5473	1.3758	0.3426	1.7184	0.2389	0.5815	1.9574	256.22	104.6034	0	12	13	5	10.987	12	295	0
4	-2.627	-1.31	-0.365	0.4724	1.061	1.3171	2.2616	3.0994	3.688	0.9445	0.8379	1.7824	0.5886	1.4265	2.371	180.21	148.7236	0	12	10	3	4.3025	10.46	84	0
5	-0.903	-0.012	0.6071	0.9859	1.04	0.8904	1.5097	1.8885	1.9427	0.6193	0.3788	0.9981	0.0542	0.4329	1.0523	167.21	160.2864	0	12	10	2	0.9149	10.78	245	0
6	-0.855	0.04	0.6085	0.9296	1.0466	0.8947	1.4632	1.7843	1.9013	0.5685	0.3211	0.8896	0.117	0.4381	1.0066	195.27	137.2535	0	12	12	1	0.8547	9.83	68	19
7	-1.668	-1.189	-0.058	0.6378	0.6806	0.4784	1.6093	2.3054	2.3481	1.1309	0.6961	1.827	0.0427	0.7388	1.8697	223.28	120.0353	0	12	13	3	2.9149	16.08	90	732
8	-1.226	-0.271	-0.005	0.6694	1.2096	0.9549	1.221	1.8953	2.4355	0.2661	0.6743	0.9404	0.5402	1.2145	1.4806	168.2	159.3429	0	12	9	3	1.5018	12.35	83	126
9	-1.259	-0.408	-0.328	0.0773	0.6558	0.8509	0.9309	1.3364	1.9149	0.08	0.4055	0.4855	0.5785	0.984	1.064	184.26	145.4547	0	12	11	3	1.9955	13.45	99	0
10	-2.14	-1.582	-0.615	-0.214	0.0634	0.5584	1.5252	1.9266	2.2039	0.9668	0.4014	1.3682	0.2773	0.6787	1.6455	212.27	126.2613	0	12	12	4	4.5518	16.83	107	478
11	-1.874	-0.526	-0.048	0.8705	1.3244	1.3484	1.8265	2.7449	3.1988	0.4781	0.9184	1.3965	0.4539	1.3723	1.8504	178.23	150.3758	0	14	12	3	2.4483	14.76	218	0
12	-3.255	-1.955	-0.941	-0.642	1.0496	1.2994	2.3138	2.6124	4.3041	1.0145	0.2985	1.313	1.6918	1.9903	3.0047	208.22	128.7171	0	12	11	4	6.7927	9.87	284	0
13	-0.864	-0.314	0.7056	0.8724	0.9277	0.55	1.5696	1.7364	1.7916	1.0196	0.1668	1.1864	0.0552	0.222	1.2417	117.03	229.0138	1	5	10	2	1.178	10.27	148	84
14	-1.193	-1.172	0.0032	0.5404	0.5959	0.0212	1.1964	1.7337	1.7892	1.1752	0.5373	1.7125	0.0555	0.5928	1.768	134.08	199.8917	1	5	12	2	2.3653	9.25	138	135
15	-1.886	-1.841	-0.615	-0.566	-0.165	0.0441	1.2708	1.3192	1.7203	1.2267	0.0484	1.2752	0.4011	0.4495	1.6763	184.02	145.6444	2	6	17	5	5.0731	9.35	232	178
16	-2.151	-0.7	-0.395	-0.164	1.122	1.4507	1.7563	1.9867	3.2731	0.3056	0.2305	0.5361	1.2863	1.5168	1.8224	252.32	106.2202	0	20	14	4	3.4108	16.41	279	0

Table S2. Test dataset G1.

compounds	x23	x25	x26	y
A	4	10.15234	290	549
B	3	7.875278	300	851
C	4	9.35254	359	1143
E	3	9.763708	339	254
F	5	10.83421	300	178
G	6	9.089554	400	222
H	5	6.760917	192	176
L	5	11.53213	267	242
M	4	11.09369	257	230

Training and test datasets G2

Table S3. Training dataset G2.

compounds	x1	x2	x3	x4	x5	x16	x18	x19	x21	x22	x23	x24	x25	x26	x27	x28	x29	x31	x32	x33	x34	x35	x36	y
7	-1.66755	-1.18916	-0.05823	0.637849	0.680572	223.28	120.0353	0	12	13	3	2.914949	16.08	90	0	6.473428	-0.843	20.5	0.1	6	1	0	0.117647	0
8	-1.2259	-0.27103	-0.0049	0.669415	1.209573	168.2	159.3429	0	12	9	3	1.50183	12.35	83	0	1.180898	-0.522	20.69	3.4	4.2	0	1	0.076923	221
10	-2.1405	-1.58211	-0.61526	-0.21389	0.063404	212.27	126.2613	0	12	12	4	4.551751	16.83	107	1	2.729268	-0.358	20.1	3.9	4	0	0	0.133333	28
14	-1.19325	-1.17202	0.003157	0.54043	0.595943	134.08	199.8917	1	5	12	2	2.365267	9.25	138	1	3.728548	-0.587	19.9	9.8	9.1	0	0	0.375	64
15	-1.88552	-1.84144	-0.61472	-0.56628	-0.16518	184.02	145.6444	2	6	17	5	5.073133	9.35	232	1	6.249363	-0.581	20.4	13.7	9.6	0	0	0.454545	1147
17	-1.71354	-0.99351	-0.29035	0.280284	0.762752	154.09	173.934	1	9	14	3	2.997402	13.77	135	0	2.685945	-0.58	19	1.2	3	0	0	0.181818	355
20	-1.2659	-0.83486	-0.12572	0.532267	0.845749	169.11	158.4855	1	9	13	3	2.226486	12.01	170	0	6.036526	-0.894	20	6.9	11.1	0	0	0.25	105
21	-3.25592	-1.85477	-1.2444	-1.04739	-0.00599	199.09	134.6199	1	9	15	5	7.408467	8.65	285	0	3.720908	-0.575	19.8	9.2	7.4	0	0	0.416667	0
22	-3.44694	-1.98593	-0.7516	-0.41253	-0.15892	196.24	136.575	0	8	13	5	6.755923	13.18	274	2	3.949577	-0.437	20.5	6.2	11.4	0	0	0.333333	142
24	-1.4553	-0.43131	-0.05687	0.153204	0.393486	190.28	140.8529	0	10	12	3	1.943	15.09	198	2	0	-0.462	21.5	3.2	7.9	0	0	0.166667	289
26	-1.87137	-0.37906	-0.24382	-0.17035	1.124672	206.29	129.9214	0	16	15	4	2.664599	13.17	152	0	0	-0.211	18.9	1.7	4.2	0	0	0	372
27	-1.77694	-0.44247	-0.39784	-0.39512	-0.35893	358.48	74.76423	0	28	22	8	3.628	13.84	201	0	0.038004	-0.168	19.1	1.2	3	0	0	0	0
28	-1.60166	-0.31402	-0.24409	0.023674	0.440553	196.3	136.5333	0	8	13	3	2.159769	15.11	67	3	0.848	-0.455	21.61	4.23	9.79	0	0	0.272727	490
29	-3.43163	-2.21855	-0.98995	-0.82587	-0.67294	252.32	106.2202	0	8	16	6	8.394989	18.35	270	3	2.582761	-0.419	21.32	0.1	11.6	0	0	0.333333	0
30	-3.29013	-1.82561	-1.30016	-0.71702	-0.62151	236.26	113.4406	2	10	15	5	7.754433	12.05	263	2	6.6624	-0.511	21.54	11.5	10.66	0	0	0.333333	178
31	-2.40717	-1.23785	-0.28926	-0.21633	0.020409	296.18	90.49052	0	14	19	4	4.150609	15.94	254	4	0.0005	-0.432	21.48	3.4	10.93	0	0	0.222222	0
32	-1.42996	0.110206	0.27157	0.924643	1.130634	178.27	150.3421	0	8	12	1	1.429959	14.11	73	2	2.5281	-0.482	20.88	3.6	8.23	0	0	0.181818	729
33	-1.71241	-1.69064	-1.01362	0.29225	0.411164	324.47	82.6008	0	18	21	3	4.416682	14.85	159	3	0.9824	-0.432	20.96	3.37	9.59	0	0	0.142857	45
34	-1.48465	-1.48357	-1.30125	-1.29744	-0.78124	353.98	75.71468	4	16	32	8	7.503816	9.41	400	0	0.061041	-0.713	21.1	12.4	6.09	0	0	0.333333	512
35	-1.97636	-1.06533	-0.79811	-0.21524	-0.1015	252.2	106.2707	1	17	17	5	4.156541	13.01	274	0	2.987769	-0.608	21.34	5.26	5.31	0	0	0.105263	109
38	-2.04004	-0.51811	-0.46232	-0.29633	-0.28545	270.33	99.14358	0	20	16	5	3.602245	12.86	134	0	0.675173	-0.526	18.8	2.08	2.54	0	1	0.047619	0
39	-1.59404	-0.54559	-0.3249	-0.06395	1.001652	218.26	122.7961	0	16	13	4	2.528483	14.29	47	0	1.104507	-0.515	20.56	2.19	4.83	0	1	0.058824	277
40	-1.70343	-1.06097	-0.49715	-0.36708	0.104492	268.32	99.88627	0	20	15	4	3.62864	15.79	161	0	0.8982	-0.506	20.83	0.78	5.42	0	1	0.047619	0
42	-2.04875	-1.01743	-0.36273	-0.07211	1.154308	196.21	136.5959	0	12	11	4	3.501019	9.98	109	0	1.256983	-0.473	20.23	7.27	4.67	2	1	0.2	277
43	-1.93772	-0.52409	-0.41307	-0.06585	-0.05116	211.28	126.8529	0	13	12	5	2.991893	11.65	115	1	0.360998	-0.371	19.73	4.24	4.65	1	0	0.133333	141

Table S4. Test dataset G2.

compounds	x4	x16	x22	x23	x25	x35	y
A	-0.02014	153.97	16	4	10.15234	0	549
B	0.061498	178	16	3	7.875278	0	851
C	-0.11238	184.02	17	4	9.35254	0	1143
D	-0.77552	326.3	20	5	7.017834	0	125
E	0.096056	221.86	24	3	9.763708	0	254
F	-0.89961	230.07	21	5	10.83421	0	178
G	-0.89117	282.15	22	6	9.089554	0	222
H	-0.29007	228.06	19	5	6.760917	0	176
I	-0.3728	340.38	18	5	9.571834	0	306
J	-1.17281	304.22	25	6	14.86136	0	253
K	-1.71731	404.1	33	11	17.67852	0	344
L	-1.16438	194.98	17	5	11.53213	0	242
M	-1.23104	195.97	17	4	11.09369	1	230

Training and test datasets G3

Table S5. Training dataset G3.

compounds	x1	x2	x3	x4	x5	x16	x20	x21	x23	x25	x28	x29	x30	x31	x32	x33	x36	y
8	-1.22587	-0.27103	-0.0049	0.6694	1.209547	168.2	0	12	3	12.35	1.180898	-0.522	0.178	20.1	3.9	4	0.076923	221
10	-2.14045	-1.58207	-0.61525	-0.21388	0.063403	212.27	1	12	4	16.83	2.729268	-0.358	0.416	21.2	0.1	6.2	0.133333	28
14	-1.19322	-1.09825	0.030749	0.540418	0.595658	134.08	1	5	2	9.25	3.728548	-0.587	0.679	19.9	9.8	9.1	0.375	64
15	-1.88548	-1.8414	-0.61471	-0.56627	-0.16517	184.02	2	6	5	9.35	12.22115	-0.736	0.716	20.4	13.7	9.6	0.454545	1147
17	-1.7135	-0.99349	-0.29035	0.280277	0.762736	154.09	1	9	3	10.73965	2.225827	-0.599	0.658	19	4.7	6.5	0.181818	355
18	-1.62833	-0.96083	-0.24218	0.326809	0.794573	168.12	1	9	3	10.6033	1.689437	-0.614	0.656	19.1	4.9	6.2	0.166667	175
19	-1.58207	-0.96247	-0.43674	0.35783	0.589943	170.09	1	9	3	12.6937	3.040187	-0.615	0.655	19.9	7	12.5	0.25	24
20	-1.26642	-0.83457	-0.12599	0.532255	0.845458	169.11	1	9	5	12.01087	1.519543	-0.898	0.648	20	6.9	11.1	0.25	105
22	-3.44687	-1.98589	-0.75158	-0.41252	-0.15891	196.24	2	8	5	13.17801	3.949654	-0.437	0.472	20.5	6.2	11.4	0.333333	142
23	-1.83432	-1.11403	-0.07701	0.142588	0.450621	190.16	1	7	4	11.7652	2.726921	-0.587	0.687	20.8	8.4	10.2	0.363636	405
24	-1.45527	-0.4313	-0.05687	0.1532	0.393477	190.28	0	10	3	15.08675	0.00001	-0.462	0.336	21.5	3.2	7.9	0.166667	227
25	-2.17691	-0.73444	0.349394	0.490077	0.574977	218.33	0	12	2	13.52738	0.00002	-0.448	0.346	20.2	3.3	8.2	0.142857	91
26	-1.87133	-0.37905	-0.24381	-0.17034	1.124647	206.29	0	16	4	13.16776	0.000004	-0.211	0.162	18.9	1.7	4.2	0	310
28	-1.60166	-0.31402	-0.24409	0.023674	0.440553	196.3	0	8	3	15.09991	0.848018	-0.455	0.423	21.61	4.23	9.79	0.272727	490
30	-3.29013	-1.82561	-1.30016	-0.71702	-0.62151	236.26	2	10	5	12.09545	6.662516	-0.511	0.586	21.54	11.5	10.66	0.333333	178
31	-2.30263	-1.23785	-0.28926	-0.21633	0.020409	304.46	0	14	4	15.94	0.000524	-0.432	0.437	21.48	3.4	10.93	0.222222	30
32	-1.42996	0.110206	0.27157	0.874302	0.924643	178.27	0	8	1	14.09257	2.528194	-0.482	0.385	20.88	3.6	8.23	0.181818	798
34	-1.48465	-1.48357	-1.30125	-1.29744	-0.78124	353.98	4	16	8	9.41	0.061017	-0.713	0.704	21.1	12.4	6.09	0.333333	513
35	-1.97636	-1.06533	-0.79811	-0.21524	-0.1015	252.2	1	17	5	13.01	2.987839	-0.608	0.687	21.34	5.26	5.31	0.105263	109
36	-1.34506	-1.18043	-1.05689	-0.81906	-0.15973	220.15	1	13	5	11.04808	10.65623	-0.794	0.668	20.2	7.1	8.9	0.1875	56
37	-1.60438	-1.04002	-0.35865	-0.16953	0.524908	194.11	1	11	4	8.210359	2.194501	-0.586	0.682	18.5	6.8	4.9	0.214286	105
39	-1.59404	-0.54559	-0.3249	-0.06395	1.001652	218.26	0	16	4	14.29	1.104528	-0.515	0.186	20.56	2.19	4.83	0.058824	277
41	-2.2161	-1.28057	-0.32681	0.742599	0.929269	186.17	1	11	3	8.702873	4.317266	-0.504	0.483	19.8	8.5	6.1	0.214286	201
42	-2.04875	-1.01743	-0.36273	-0.07211	1.154308	196.21	0	12	4	9.98	1.257007	-0.473	0.312	20.23	7.27	4.67	0.2	277
43	-1.93772	-0.52409	-0.41307	-0.06585	-0.05116	211.28	0	13	5	11.65	0.361039	-0.371	0.395	19.73	4.24	4.65	0.133333	141
44	-1.36683	-1.00954	-0.42967	-0.27728	0.365993	144.05	1	7	4	10.5423	9.8109	-0.793	0.674	19.7	10.3	11.9	0.3	134
45	-1.26125	-1.11213	0.094151	0.389395	0.60627	144.05	1	7	2	11.58318	4.530913	-0.623	0.677	20	10	12.7	0.3	15
46	-1.3769	-1.22152	0.264767	0.383409	0.555657	174.08	1	7	2	11.11126	3.105764	-0.679	0.699	19.9	10.4	12.4	0.333333	267
47	-1.19077	-1.16111	0.342864	0.591848	0.716748	159.07	1	7	2	15.38636	4.241533	-0.914	0.693	20.9	11.3	16.7	0.363636	73
48	-1.29934	-1.20465	-0.94805	-0.83158	0.302863	194	2	8	4	12.04118	4.762919	-0.686	0.701	20.8	13.3	12.8	0.384615	318
49	-1.45227	-0.91512	-0.80192	0.403817	0.54314	194.11	1	11	3	11.15796	1.292809	-0.606	0.668	19.3	6.7	10.7	0.214286	63
50	-1.18206	-1.13227	0.156466	0.326809	0.567902	174.08	1	7	2	11.11126	3.50575	-0.689	0.695	19.9	10.4	12.4	0.333333	229
51	-1.05635	-0.84083	0.413069	0.461233	0.685999	159.07	1	7	2	15.38636	1.994699	-0.897	0.687	20.9	11.3	16.7	0.363636	133
52	-1.74207	-1.2071	-1.00791	-0.02204	0.324632	194	2	8	4	12.04118	1.750611	-0.679	0.699	20.8	13.3	12.8	0.384615	279
53	-1.29173	-1.23268	0.259325	0.281638	0.524908	158.08	1	7	2	10.61603	4.378474	-0.683	0.698	19.7	9.1	11.3	0.272727	23
54	-1.17744	-1.10832	0.334428	0.564364	0.747497	159.07	1	7	2	15.38636	3.495252	-0.911	0.695	20.9	11.3	16.7	0.363636	273

Table S6. Test dataset G3.

compounds	x2	x20	x25	x28	x33	x36	y
A	-1.01254	2	10.15234	0.002766	10.8	0.4	549
B	-1.02369	2	7.875278	0.001416	7	0.333333	851
C	-1.04873	2	9.35254	0.524158	9.6	0.454545	1143
D	-2.93611	4	7.017834	0.000116	7.1	0.25	125
E	-1.48411	2	9.763708	0.046831	6.9	0.428571	254
F	-1.00111	2	10.83421	0.001969	8	0.333333	178
G	-1.03485	2	9.089554	0.000472	4.9	0.2	222
H	-1.03458	2	6.760917	0.000102	6.2	0.25	176
I	-1.36084	2	9.571834	0.069349	7.8	0.153846	306
J	-1.40384	2	14.86136	0.004566	13.7	0.5	253
K	-2.11677	4	17.67852	1.351053	17.7	0.545455	344
L	-1.46642	3	11.53213	6.848167	12.6	0.461538	242
M	-1.74126	3	11.09369	7.937359	12.2	0.461538	230

RMSE values of the ten-fold cross validation

Table S7. RMSE values of the ten-fold cross validation for the models G1–G3 using the merged dataset.

Model G1												
Pattern	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)	Average	S. D.
Training	292.40	289.96	275.02	294.33	261.87	233.27	279.71	292.84	290.83	285.72	279.59	18.19
Test	212.10	222.74	366.08	186.04	446.21	643.46	376.66	121.31	174.10	285.50	303.42	149.71
Model G2												
Pattern	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)	Average	S. D.
Training	230.54	245.18	248.91	243.80	254.48	248.47	221.71	217.03	250.15	235.91	239.62	12.13
Test	344.23	239.21	203.84	267.10	90.48	210.87	388.88	417.86	116.88	333.43	261.28	104.53
Model G3												
Pattern	(i)	(ii)	(iii)	(iv)	(v)	(vi)	(vii)	(viii)	(ix)	(x)	Average	S. D.
Training	205.51	193.01	203.78	198.41	188.15	205.59	181.95	203.27	199.81	163.98	194.35	12.58
Test	94.34	233.84	129.29	200.09	281.37	95.12	323.50	133.12	214.06	480.16	218.49	113.98

The original training and test datasets were mixed and then divided into ten segments. The one segment and remaining nine segments were assigned to the test and training data, respectively. The RMSE values were calculated for the training and test data in each model (Table S7). In addition, the relationship between the estimated and measured values was summarized (Figs. S2–S4). This validation was performed by changing the assignments of the test data for the total ten patterns. Then, the average and standard deviation (S. D.) were calculated.

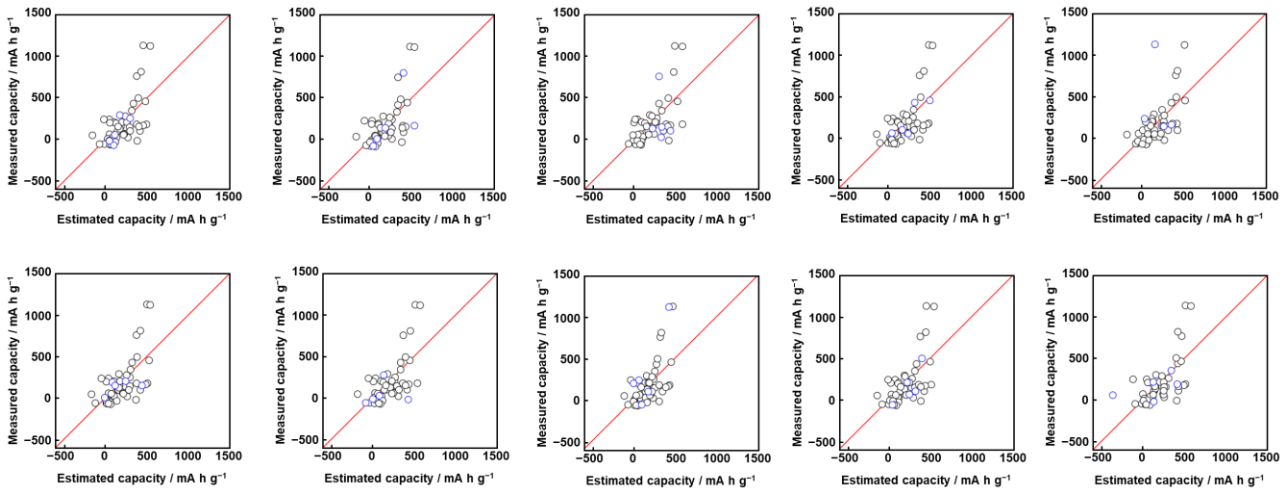


Fig. S2. Relationship between the estimated and measured capacity of the ten models prepared for the ten-fold cross validation using the dataset merging the training and test datasets G3.

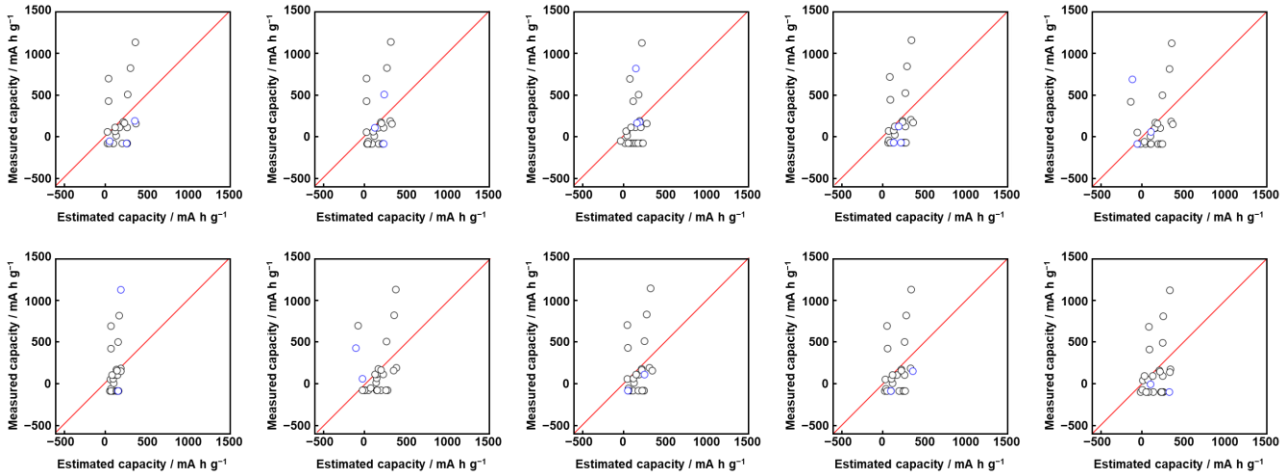


Fig. S3. Relationship between the estimated and measured capacity of the ten models prepared for the ten-fold cross validation using the dataset merging the training and test datasets G1.

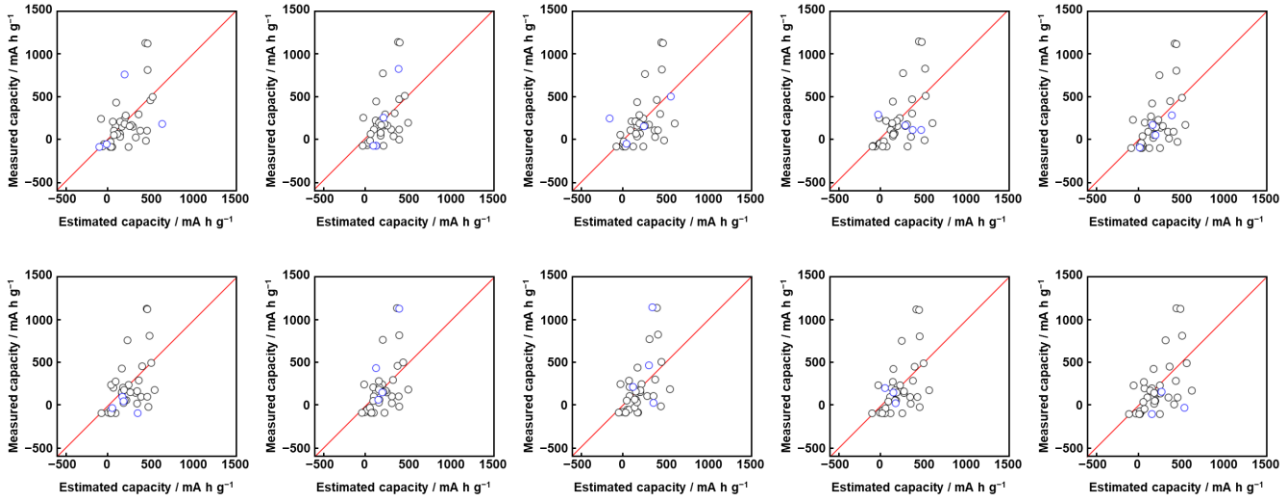


Fig. S4. Relationship between the estimated and measured capacity of the ten models prepared for the ten-fold cross validation using the dataset merging the training and test datasets G2.

Reduced training datasets G1′ and G2′

Table S8. Training dataset G1′.

compounds	x1	x2	x3	x4	x5	x16	x20	x21	x23	x25	x28	x29	x30	x31	x32	x33	x36	y
1	-0.96165	-0.46858	0.093335	0.933079	1.363835	166.22	0	12	2	13.24009	0.543053	-0.484	0.229	20	2.8	2.8	0	0
2	-3.06373	-1.45472	-0.83893	-0.33062	0.540146	211.22	0	12	4	12.18319	4.853086	-0.516	0.296	20.9	6.7	4.5	0.1875	0
3	-3.67953	-3.08958	-1.71377	-1.37118	-1.13227	256.22	0	12	5	11.99583	1.311101	-0.562	0.289	21.5	7.3	5.3	0.315789	0
4	-2.62699	-1.30996	-0.36545	0.47239	1.060972	180.21	1	12	3	10.48523	3.729398	-0.322	0.266	20.7	7.6	4	0.071429	0
5	-0.9026	-0.01225	0.607086	0.985869	1.04002	167.21	0	12	1	10.77544	1.636026	-0.903	0.333	20.6	7.1	6	0.076923	0
6	-0.85471	0.040001	0.608447	0.929541	1.04655	195.27	0	12	1	9.831582	2.175501	-0.844	0.338	20.1	7.2	4.4	0.066667	0
7	-1.66751	-1.18914	-0.05823	0.637835	0.680557	223.28	1	12	3	16.08	6.473428	-0.843	0.329	20.5	0.1	6	0.117647	0
8	-1.22587	-0.27103	-0.0049	0.6694	1.209547	168.2	0	12	3	12.35	1.180898	-0.522	0.178	20.69	3.4	4.2	0.076923	221
9	-1.25907	-0.40817	-0.32817	0.07728	0.655795	196.24	0	12	3	13.18	3.949577	-0.437	0.472	20.5	6.2	11.4	0.333333	0
10	-2.14045	-1.58207	-0.61525	-0.21388	0.063403	212.27	1	12	4	16.83	2.729268	-0.358	0.416	20.1	3.9	4	0.133333	28
11	-1.87432	-0.526	-0.04789	0.870493	1.324379	178.23	0	14	3	11.49043	0.000062	-0.16	0.157	20.4	1.4	4.8	0	0
12	-3.25421	-1.95459	-0.9407	-0.64219	1.049544	208.22	2	14	4	9.012769	0.000071	-0.343	0.201	20.9	11.7	4.8	0.125	0
13	-0.86423	-0.31348	0.705319	0.872397	0.927637	117.03	1	5	2	10.27132	1.432191	-0.7	0.663	20.3	13.5	11.3	0.375	0
14	-1.19322	-1.09825	0.030749	0.540418	0.595658	134.08	1	5	2	9.25	3.728548	-0.587	0.679	19.9	9.8	9.1	0.375	64
15	-1.88548	-1.8414	-0.61471	-0.56627	-0.16517	184.02	2	6	5	9.35	6.249363	-0.581	0.668	20.4	13.7	9.6	0.454545	1147
16	-2.15106	-0.70042	-0.39484	-0.21497	-0.16436	252.32	0	20	5	16.40579	0	-0.151	0.162	20.9	0.1	4.8	0	0

The training dataset was reduced to the 16 compounds used for the construction of the model G1.

Table S9. Training dataset G2′.

compounds	x1	x2	x3	x4	x5	x16	x20	x21	x23	x25	x28	x29	x30	x31	x32	x33	x36	y
7	-1.66751	-1.18914	-0.05823	0.637835	0.680557	223.28	1	12	3	16.08	6.473428	-0.843	0.329	20.5	0.1	6	0.117647	0
8	-1.22587	-0.27103	-0.0049	0.6694	1.209547	168.2	0	12	3	12.35	1.180898	-0.522	0.178	20.1	3.9	4	0.076923	221
10	-2.14045	-1.58207	-0.61525	-0.21388	0.063403	212.27	1	12	4	16.83	2.729268	-0.358	0.416	21.2	0.1	6.2	0.133333	28
14	-1.19322	-1.09825	0.030749	0.540418	0.595658	134.08	1	5	2	9.25	3.728548	-0.587	0.679	19.9	9.8	9.1	0.375	64
15	-1.88548	-1.8414	-0.61471	-0.56627	-0.16517	184.02	2	6	5	9.35	12.22115	-0.736	0.716	20.4	13.7	9.6	0.454545	1147
17	-1.7135	-0.99349	-0.29035	0.280277	0.762736	154.09	1	9	3	10.73965	2.225827	-0.599	0.658	19	4.7	6.5	0.181818	355
20	-1.26642	-0.83457	-0.12599	0.532255	0.845458	169.11	1	9	5	12.01087	1.519543	-0.898	0.648	20	6.9	11.1	0.25	105
21	-3.25612	-1.855	-1.24438	-1.04737	-0.00599	199.09	1	9	5	8.654479	8.512576	-0.59	0.672	19.8	9.2	7.4	0.357143	0
22	-3.44687	-1.98589	-0.75158	-0.41252	-0.15891	196.24	2	8	5	13.17801	3.949654	-0.437	0.472	20.5	6.2	11.4	0.333333	142
24	-1.45527	-0.4313	-0.05687	0.1532	0.393477	190.28	0	10	3	15.08675	0.00001	-0.462	0.336	21.5	3.2	7.9	0.166667	227
26	-1.87133	-0.37905	-0.24381	-0.17034	1.124647	206.29	0	16	4	13.16776	0.000004	-0.211	0.162	18.9	1.7	4.2	0	310
27	-1.7769	-0.44246	-0.39783	-0.39511	-0.35892	358.48	0	28	8	13.84	0.038294	-0.168	0.186	19.1	1.2	3	0	0
28	-1.60166	-0.31402	-0.24409	0.023674	0.440553	196.3	0	8	3	15.09991	0.848018	-0.455	0.423	21.61	4.23	9.79	0.272727	490
30	-3.43163	-2.21855	-0.98995	-0.82587	-0.67294	252.32	2	8	6	18.37149	2.582761	-0.419	0.487	21.32	0.1	11.6	0.333333	0
31	-3.29013	-1.82561	-1.30016	-0.71702	-0.62151	236.26	2	10	5	12.09545	6.662516	-0.511	0.586	21.54	11.5	10.66	0.333333	178
32	-2.30263	-1.23785	-0.28926	-0.21633	0.020409	304.46	0	14	4	15.94	0.000524	-0.432	0.437	21.48	3.4	10.93	0.222222	30
33	-1.42996	0.110206	0.27157	0.874302	0.924643	178.27	0	8	1	14.09257	2.528194	-0.482	0.385	20.88	3.6	8.23	0.181818	798
36	-1.71241	-1.69064	-1.01362	0.29225	0.411164	324.47	0	18	3	14.85	0.9824	-0.582	0.384	20.96	3.37	9.59	0.142857	55
37	-1.48465	-1.48357	-1.30125	-1.29744	-0.78124	353.98	4	16	8	9.41	0.061017	-0.713	0.704	21.1	12.4	6.09	0.333333	513
39	-1.97636	-1.06533	-0.79811	-0.21524	-0.1015	252.2	1	17	5	13.01	2.987839	-0.608	0.687	21.34	5.26	5.31	0.105263	109
43	-2.04004	-0.51811	-0.46232	-0.29633	-0.28545	270.33	0	20	5	12.86	0.675154	-0.526	0.187	18.8	2.08	2.54	0.047619	0
44	-1.59404	-0.54559	-0.3249	-0.06395	1.001652	218.26	0	16	4	14.29	1.104528	-0.515	0.186	20.56	2.19	4.83	0.058824	277
45	-1.70343	-1.06097	-0.49715	-0.36708	0.104492	268.32	0	20	4	15.79	0.898265	-0.506	0.197	20.83	0.78	5.42	0.047619	0
48	-2.04875	-1.01743	-0.36273	-0.07211	1.154308	196.21	0	12	4	9.98	1.257007	-0.473	0.312	20.23	7.27	4.67	0.2	277
49	-1.93772	-0.52409	-0.41307	-0.06585	-0.05116	211.28	0	13	5	11.65	0.361039	-0.371	0.395	19.73	4.24	4.65	0.133333	141

The training dataset was reduced to the 25 compounds used for the construction of the model G2.

Weight diagrams prepared using the reduced datasets

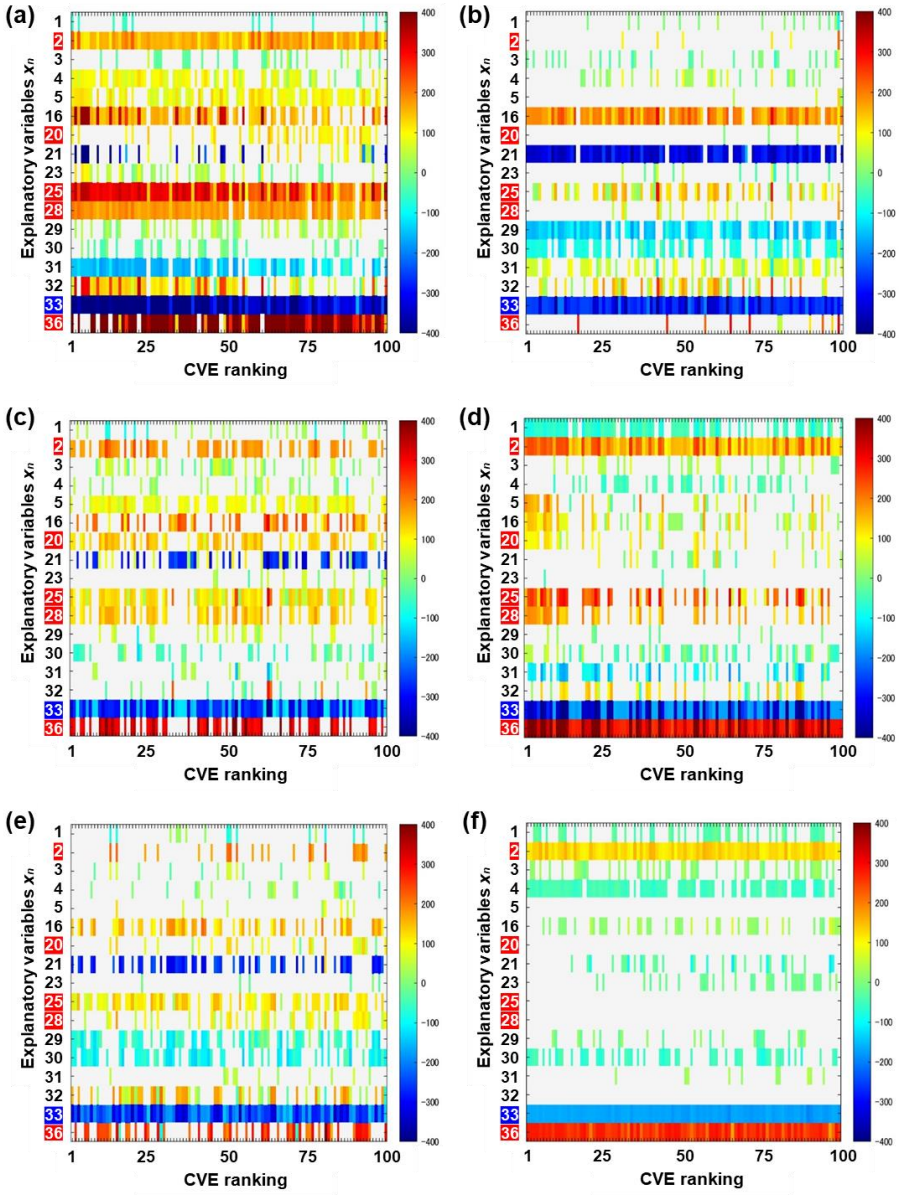


Fig. S5. Weight diagrams of the reduced datasets containing 35 *y* prepared in random six patterns.

The extractable x_n and its number N_x: (a) x_n (n = 2, 25, 28, 33, 36), N_x = 5. (b) x_n (n = 25, 33), N_x = 2. (c) x_n (n = 2, 20, 25, 28, 33, 36), N_x = 6. (d) x_n (n = 2, 25, 28, 33, 36), N_x = 5. (e) x_n (n = 2, 25, 33, 36), N_x = 4. (f) x_n (n = 2, 33, 36), N_x = 3. The results are summarized in Table 4.

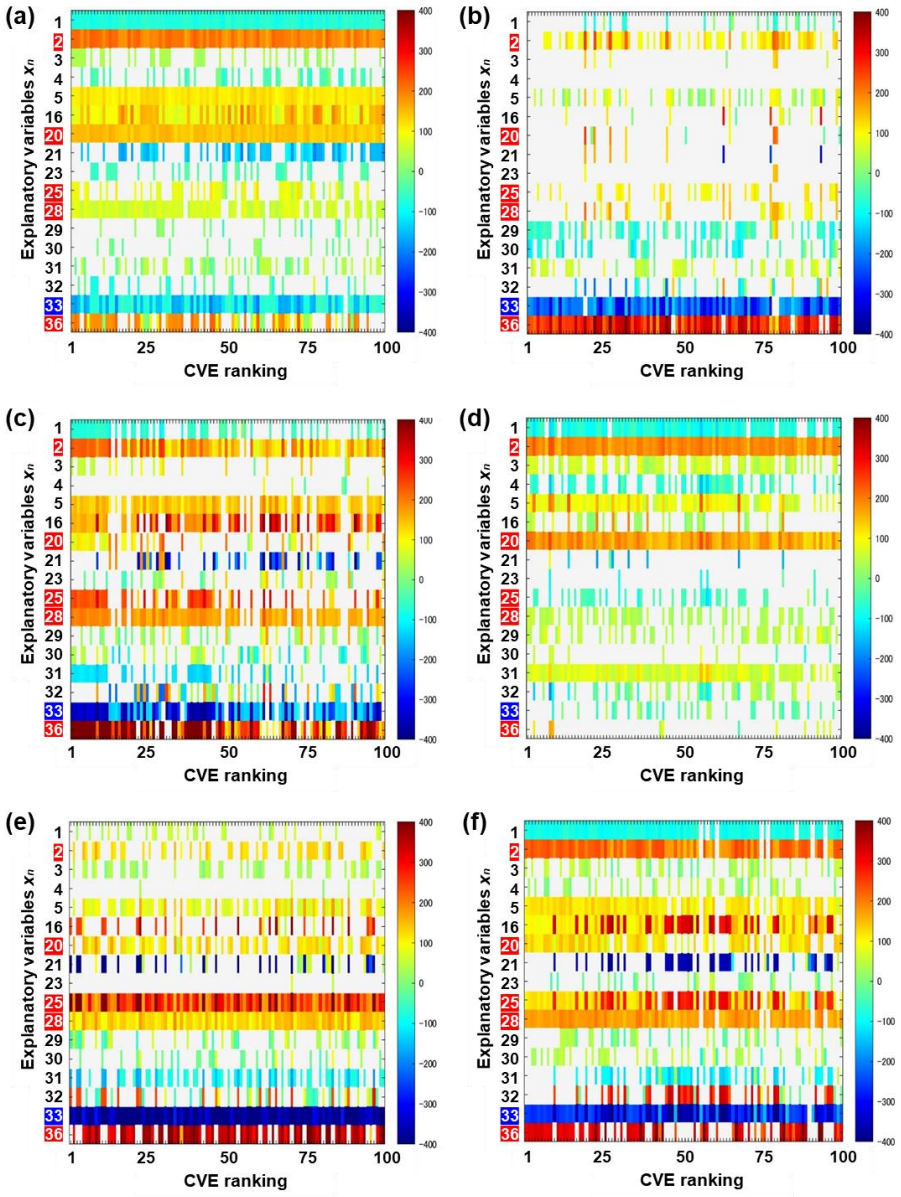


Fig. S6. Weight diagrams of the reduced datasets containing 34 *y* prepared in random six patterns.

The extractable x_n and its number N_x: (a) x_n (n = 2, 20, 28, 33, 36), N_x = 5. (b) x_n (n = 2, 33, 36), N_x = 3. (c) x_n (n = 2, 25, 28, 33, 36), N_x = 5. (d) x_n (n = 2, 20, 25, 28), N_x = 4. (e) x_n (n = 2, 20, 25, 28, 33, 36), N_x = 6. (f) x_n (n = 2, 20, 25, 28, 33, 36), N_x = 6. The results are summarized in Table 4.

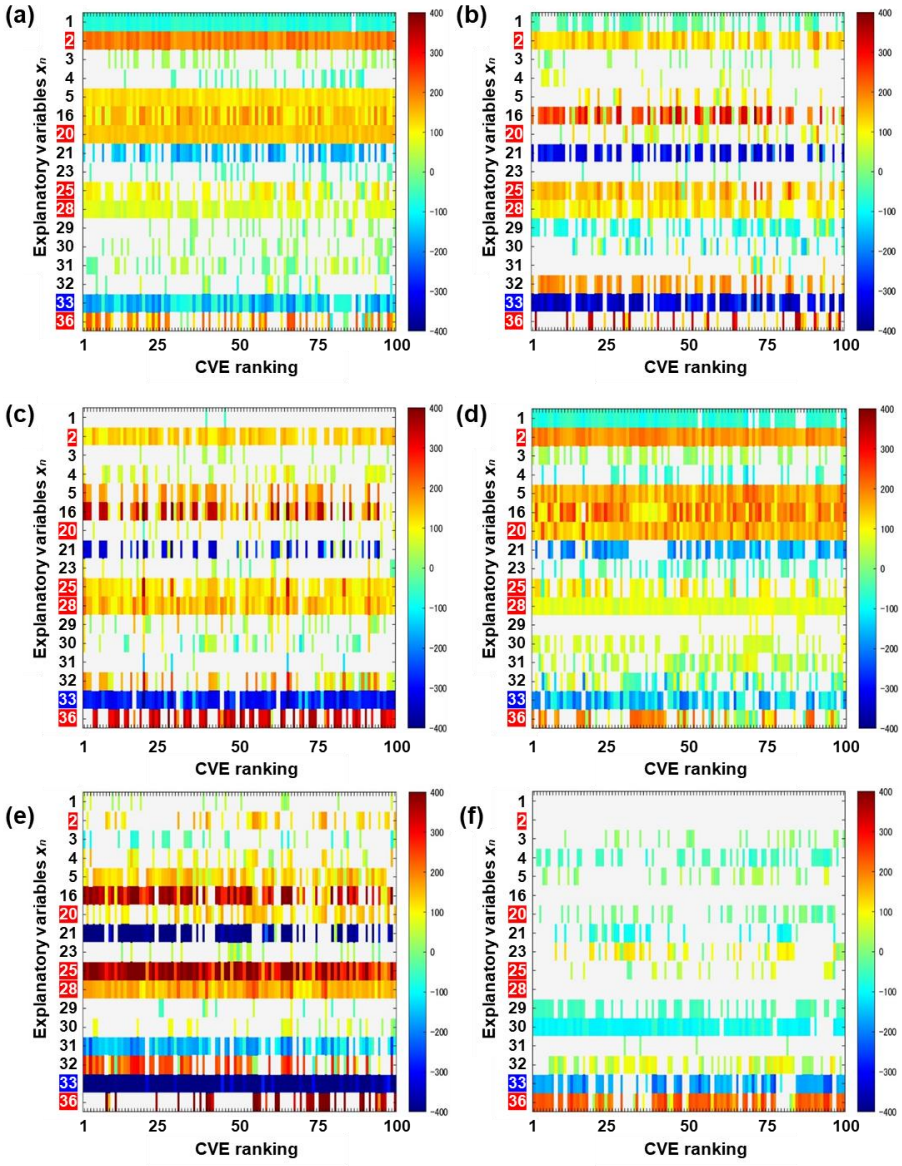


Fig. S7. Weight diagrams of the reduced datasets containing 33 y prepared in random six patterns.

The extractable x_n and its number N_x: (a) x_n (n = 2, 20, 28, 33, 36), N_x = 5. (b) x_n (n = 2, 25, 28, 33), N_x = 4. (c) x_n (n = 2, 25, 28, 33, 36), N_x = 5. (d) x_n (n = 2, 20, 28, 33, 36), N_x = 5. (e) x_n (n = 20, 25, 28, 33), N_x = 4. (f) x_n (n = 33, 36), N_x = 2. The results are summarized in Table 4.

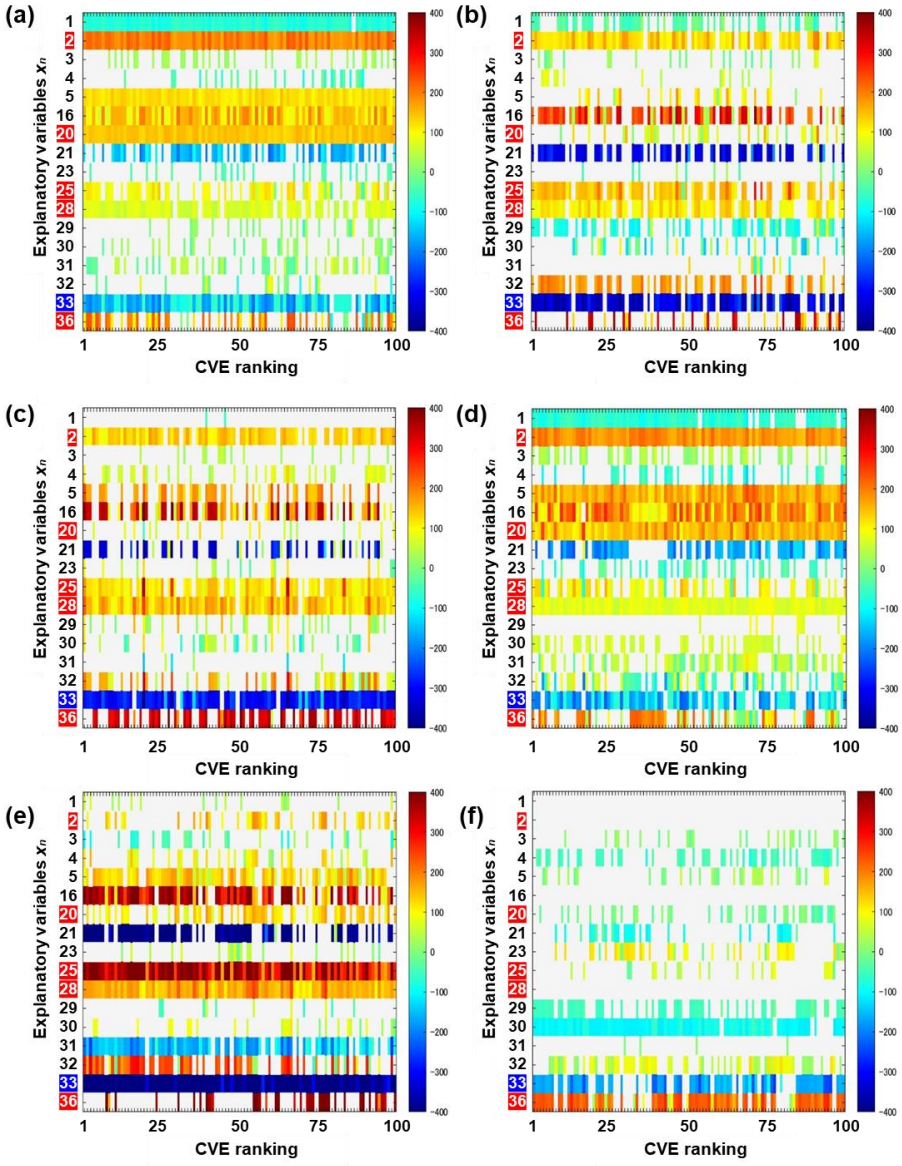


Fig. S8. Weight diagrams of the reduced datasets containing 30 *y* prepared in random six patterns.

The extractable x_n and its number N_x: (a) x_n (n = 2, 20, 28, 33), N_x = 4. (b) x_n (n = 25, 33), N_x = 2. (c) x_n (n = 2, 25, 28, 33, 36), N_x = 5. (d) x_n (n = 2, 20, 25, 28, 33), N_x = 5. (e) x_n (n = 20, 25, 28, 33), N_x = 4. (f) x_n (n = 33), N_x = 1. The results are summarized in Table 4.

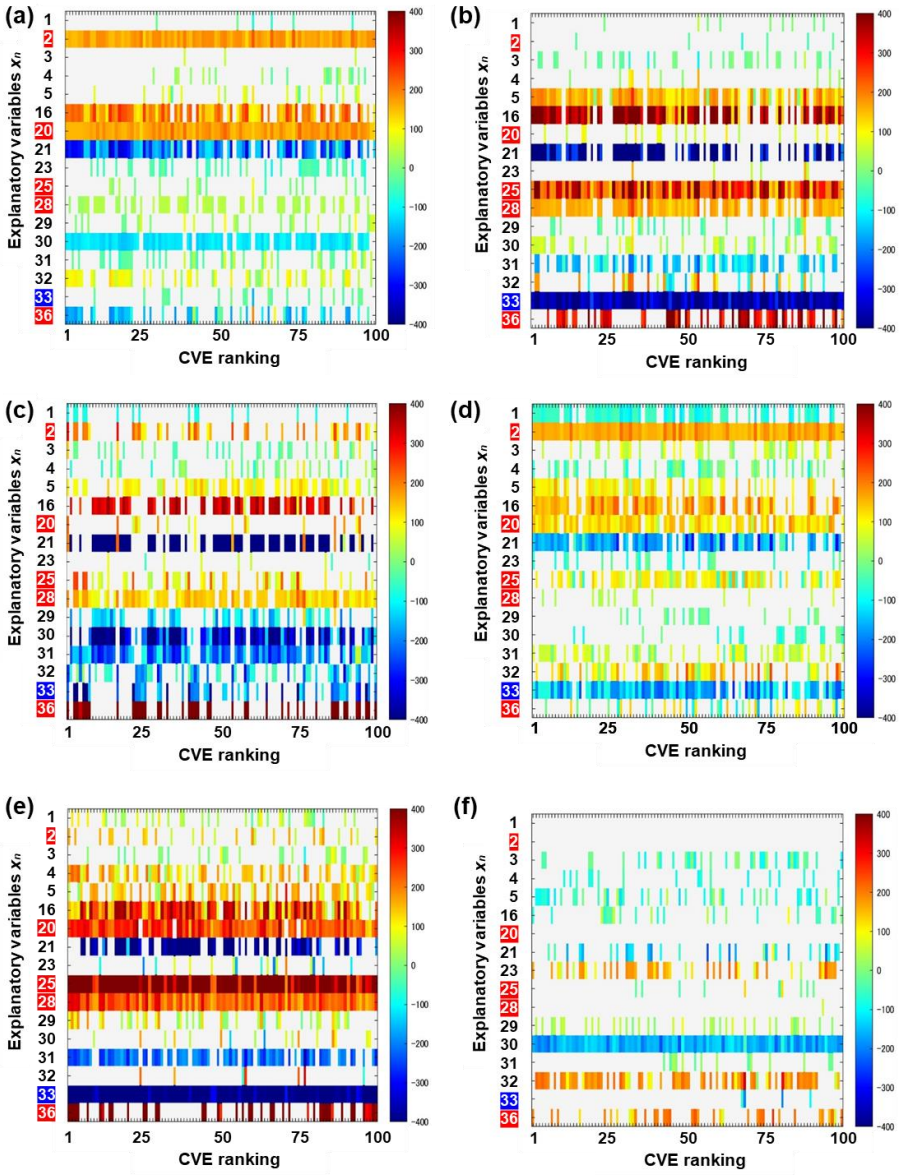


Fig. S9. Weight diagrams of the reduced datasets containing 27 *y* prepared in random six patterns.

The extractable x_n and its number N_x: (a) x_n (n = 2, 20, 28), N_x = 3. (b) x_n (n = 25, 28, 33, 36), N_x = 4. (c) x_n (n = 2, 28, 33, 36), N_x = 4. (d) x_n (n = 2, 20, 33), N_x = 3. (e) x_n (n = 20, 25, 28, 33), N_x = 4. (f) x_n (not available), N_x = 0. The results are summarized in Table 4.