Supporting Information for

Representing Multi-Word Chemical Terms through Phrase-Level Preprocessing and Word Embedding

Liyuan Huang, Chen Ling*

* Corresponding author: chen.ling@toyota.com

Toyota Research Institute of North America, 1555 Woodridge Avenue, Ann Arbor, Michigan, United States, 48105

D1

Dataset contains the names of 50 lithium organic and inorganic compound extracted from PubChem. There are total 31 of them actually appear in the 119,166 patents, so we use them to do the evaluation. These 31 compound names are listed below.

1	lithium_hydroxide
2	lithium_tetrahydridoaluminate
3	lithium_ion
4	lithium_iodide
5	lithium_bromide
6	lithium_aluminate
7	lithium_metaborate
8	lithium_hydroxide_monohydrate
9	lithium_fluoride
10	lithium_chloride
11	lithium_stearate
12	lithium_acetoacetate
13	lithium_diisopropylamide
14	lithium_bis(trimethylsilyl)amide
15	lithium_3,5-diiodosalicylate
16	lithium_acetate
17	lithium_diphenylphosphide_solution
18	lithium_trifluoromethanesulfonate
19	bis(trifluoromethane)sulfonimide_lithium_salt
20	lithium_tri(tert-butoxy)aluminum_hydride
21	lithium_aluminium_hydride
22	lithium_tert-butoxide
23	lithium_perchlorate
24	lithium_hypochlorite
25	lithium_acetate_dihydrate
26	lithium_dihydrogen_phosphate
27	lithium_isopropoxide
28	lithium_lactate
29	lithium_bicarbonate
30	lithium_nonafluoro-1-butanesulfonate
31	lithium_hexafluorophosphate

D2

Dataset contains the two-word (2-grams) names of inorganic compound extracted from PubChem. There are total 932 of them actually appear in the 119,166 patents, so we use them to do the evaluation. These 932 compound names are listed below.

1	arsenic_acid
2	hydrobromic_acid
3	carbon_dioxide
4	carbon_monoxide
5	chloride_ion
6	hydrochloric_acid
7	hydrogen_sulfide
8	trithionic_acid
9	trimetaphosphoric_acid
10	hydrogen_selenide
11	arsenous_acid
12	bicarbonate_ion
13	hydrogen_cyanide
14	hydrogen_peroxide
15	nitric_acid
16	nitrous_oxide
17	oxygen_molecule
18	triphosphoric_acid
19	phosphoric_acid
20	diphosphoric_acid
21	hydrogen_ion
22	dimethyl_sulfide
23	selenic_acid
24	selenious_acid
25	sulfuric_acid
26	sulfur_dioxide
27	tungstic_acid
28	2-aminoethyl_diphenylborinate
29	3-nitrophenylboronic_acid
30	aluminum_fluoride
31	benzo[b]thiophene-2-boronic_acid
32	phenylarsine_oxide
33	potassium_chloride
34	potassium_iodide
35	sodium_chloride
36	sodium_fluoride
37	sodium_iodide

38	benzyltrimethylammonium_chloride
39	demecarium_bromide
40	sulfamic_acid
41	penthienate_bromide
42	calcium_acetate
43	isopropamide_iodide
44	carbon_disulfide
45	boron_trifluoride
46	acetyl_chloride
47	tetramethylammonium_chloride
48	tetramethylammonium_iodide
49	tert-butyl_hydroperoxide
50	triphenylsilyl_chloride
51	tetraethyl_orthosilicate
52	valethamate_bromide
53	benzenearsonic_acid
54	disulfide_bis(4-nitrophenyl)
55	metalkonium_chloride
56	triphenyl_phosphite
57	tributyl_phosphite
58	boric_acid
59	dilauroyl_peroxide
60	dimethyldioctadecylammonium_chloride
61	ethyl_nitrite
62	di-tert-butyl_peroxide
63	dipropyl_sulfide
64	dodecyltrimethylammonium_chloride
65	trimethyl_borate
66	trimethyl_phosphite
67	diphenyl_sulfide
68	sodium_cyanide
69	disodium_methanearsonate
70	methylarsonic_acid
71	hydroxymethanesulfinic_acid
72	dibenzyl_disulfide
73	potassium_cyanide
74	stannous_octoate

75	tris(2-ethylhexyl)_phosphite
76	azamethonium_bromide
77	diethyl_sulfide
78	silicon_carbide
79	carbonyl_sulfide
80	calcium_carbonate
81	sodium_carbonate
82	cyanogen_bromide
83	cyanogen_chloride
84	diammonium_carbonate
85	barium_carbonate
86	cadmium_carbonate
87	dibutoline_sulfate
88	sodium_sesquicarbonate
89	cesium_carbonate
90	domiphen_bromide
91	furtrethonium_iodide
92	isobutyl_nitrite
93	isobutyl_nitrate
94	cadmium_acetate
95	strontium_acetate
96	butyl_nitrite
97	dibutyl_sulfide
98	magnesium_carbonate
99	thiphenamil_hydrochloride
100	thonzonium_bromide
101	lithium_carbonate
102	silver_acetate
103	ferrous_carbonate
104	potassium_carbonate
105	trolnitrate_phosphate
106	mercuric_cyanide
107	diallyl_sulfide
108	methyl_nitrate
109	manganese_carbonate
110	lead_carbonate
111	tert-butyl_perbenzoate
112	benzenesulfinic_acid
113	methyl_nitrite
114	dimethyl_disulfide
115	ethyl_nitrate
116	n-propyl_nitrate
117	divinyl_sulfide

118	mercury_fulminate
119	dipropyl_disulfide
120	dibutyl_disulfide
121	brilliant_green
122	tetramethylammonium_tetrafluoroborate
123	tetramethyl_silicate
124	calcium_citrate
125	calcium_lactate
126	trimethylolpropane_phosphite
127	diphenyl_disulfide
128	lanthanum_acetate
129	sec-butyl_nitrite
130	heteronium_bromide
131	ammonium_bicarbonate
132	propylene_sulfide
133	dodecyltrimethylammonium_bromide
134	cupric_carbonate
135	sodium_aluminate
136	gallium_arsenide
137	beryllium_oxide
138	calcium_oxide
139	calcium_peroxide
140	cadmium_oxide
141	cadmium_sulfide
142	cadmium_selenide
143	iron_disulfide
144	magnesium_oxide
145	lead_dioxide
146	germanium_dioxide
147	potassium_hydroxide
148	sodium_hydroxide
149	manganese_dioxide
150	molybdenum_trioxide
151	sodium_peroxide
152	nickel(ii)_oxide
153	zinc_oxide
154	thorium_dioxide
155	tungsten_trioxide
156	antimony_pentoxide
157	vanadium_pentoxide
158	phosphorus_pentasulfide
159	phosphorus_sesquisulfide
160	zinc_sulfide

161	molybdenum_disulfide
162	lead_monoxide
163	ferrous_sulfide
164	cupric_oxide
165	cuprous_oxide
166	cobalt_sulfide
167	hydrofluoric_acid
168	ammonium_hydroxide
169	ferrous_oxide
170	3-hydroxyphenyltrimethylammonium_bromide
171	strontium_carbonate
172	isopropyl_nitrate
173	2-hydroxyethyl_disulfide
174	phenyltrimethylammonium_hydroxide
175	tetrabutylammonium_nitrate
176	diallyl_trisulfide
177	decyltrimethylammonium_bromide
178	diallyl_disulfide
179	didecyldimethylammonium_bromide
180	tetrapentylammonium_iodide
181	3-(trimethoxysilyl)propyl_methacrylate
182	tetramethylammonium_perchlorate
183	sulfur_hexafluoride
184	diphenylborinic_acid
185	nitrosyl_chloride
186	sulfuryl_fluoride
187	diisopropyl_sulfate
188	diammonium_citrate
189	glycidyltrimethylammonium_chloride
190	methylbenactyzium_bromide
191	cupric_nitrate
192	didodecyldimethylammonium_bromide
193	nickel_carbonate
194	didodecyldimethylammonium_chloride
195	ammonium_citrate
196	zinc_carbonate
197	dimethyl_trisulfide
198	denatonium_benzoate
199	chloric_acid
200	thiophosphoryl_chloride
201	butylboronic_acid
202	benzyltrimethylammonium_iodide

203	diphenylarsinic_acid
204	methyltrioctylammonium_chloride
205	benzyltrimethylammonium_bromide
206	calcium_borogluconate
207	ferrous_lactate
208	ammonium_nitrate
209	ferric_methanearsonate
210	sodium_silicate
211	potassium_pyrophosphate
212	zinc_dust
213	selenium_dioxide
214	cupric_chloride
215	magnesium_sulfate
216	mercuric_chloride
217	selenium_disulfide
218	titanium_tetrachloride
219	disodium_phosphate
220	trisodium_phosphate
221	perchloric_acid
222	sodium_monofluorophosphate
223	sodium_nitrate
224	tin_tetrachloride
225	palladium(ii)_chloride
226	cesium_chloride
227	antimony_pentachloride
228	hypochlorous_acid
229	iodic_acid
230	copper(i)_iodide
231	zirconium_oxychloride
232	nickel_chloride
233	thionyl_chloride
234	phosphorus_trichloride
235	ferrous_sulfate
236	monoammonium_phosphate
237	tetrasodium_pyrophosphate
238	potassium_persulfate
239	barium_sulfate
240	zinc_sulfate
241	chromic_acid
242	potassium_nitrate
243	sodium_sulfate
244	sodium_sulfite
245	trimagnesium_phosphate

246	bromic_acid
247	dipotassium_phosphate
248	chlorous_acid
249	calcium_biphosphate
250	sodium_tripolyphosphate
251	cerium_fluoride
252	ferrous_chloride
253	lead_chloride
254	lead_chromate
255	cupric_sulfate
256	silver_nitrate
257	sodium_thiosulfate
258	tin_dichloride
259	manganese_dichloride
260	ammonium_sulfamate
261	mercuric_iodide
262	sodium_chromate
263	sodium_hydrosulfite
264	calcium_sulfate
265	calcium_arsenate
266	potassium_dichromate
267	calcium_hypochlorite
268	potassium_sulfate
269	zinc_nitrate
270	zinc_phosphate
271	nitrous_acid
272	hydrazoic_acid
273	diammonium_phosphate
274	mercuric_sulfate
275	mercurous_sulfate
276	oxygen_difluoride
277	stannous_fluoride
278	ferric_fluoride
279	nitrogen_trifluoride
280	sulfur_tetrafluoride
281	antimony_pentafluoride
282	uranium_hexafluoride
283	silver_chloride
284	silver_perchlorate
285	silver_iodide
286	arsenic_triiodide
287	sodium_trimetaphosphate
288	manganese_sulfate

289	nickel_sulfate
290	bismuth_trichloride
291	cesium_bromide
292	bromine_trifluoride
293	potassium_chromate
294	strontium_chromate
295	ammonium_dichromate
296	cesium_iodide
297	deuterium_oxide
298	fluorosulfonic_acid
299	cobalt_bromide
300	copper(ii)_bromide
301	calcium_pyrophosphate
302	cadmium_fluoride
303	cadmium_iodide
304	cerium(iii)_chloride
305	chlorine_trifluoride
306	chlorosulfonic_acid
307	ammonium_perchlorate
308	iodine_monochloride
309	sulfuryl_chloride
310	ammonium_bisulfate
311	sulfur_trioxide
312	barium_nitrate
313	sulfur_monochloride
314	phosphorus_oxychloride
315	antimony_trichloride
316	silicon_tetrachloride
317	zirconium_tetrachloride
318	ferric_sulfate
319	magnesium_perchlorate
320	hydriodic_acid
321	strontium_nitrate
322	aluminum_sulfate
323	ferric_phosphate
324	mercuric_nitrate
325	chlorine_dioxide
326	ferric_pyrophosphate
327	lead_nitrate
328	palladium_nitrate
329	uranyl_nitrate
330	sodium_selenite
331	sodium_tellurite

332	tellurous_acid
333	cadmium_dichloride
334	cerium_nitrate
335	trimethyloctylammonium_chloride
336	decyltrimethylammonium_chloride
337	mercurous_chloride
338	cadmium_sulfate
339	calcium_nitrate
340	calcium_thiosulfate
341	cobalt_sulfate
342	calcium_chlorate
343	dysprosium_nitrate
344	ammonium_sulfite
345	boron_tribromide
346	boron_trichloride
347	barium_chromate
348	cesium_sulfate
349	cadmium_nitrate
350	magnesium_chlorate
351	magnesium_nitrate
352	orthoperiodic_acid
353	nitrogen_tetroxide
354	sodium_dichromate
355	ammonium_iodide
356	ammonium_paramolybdate
357	ammonium_bromide
358	ammonium_chloride
359	tricalcium_silicate
360	isobutyl_xanthate
361	chromium_hexacarbonyl
362	nickel_nitrate
363	cesium_fluoride
364	sodium_selenate
365	itramin_tosylate
366	aminoethyl_nitrate
367	gallium_trichloride
368	gold_trichloride
369	platinum_tetrachloride
370	titanium_dioxide
371	sodium_tungstate
372	aluminium_nitrate
373	beryllium_sulfate
374	zinc_chromate

375	dichromic_acid
376	beryllium_nitrate
377	potassium_ferricyanide
378	tetrathionic_acid
379	calcium_chromate
380	indium_nitrate
381	thorium_nitrate
382	calcium_metasilicate
383	trimethidinium_methosulfate
384	sodium_tetrachloroaurate
385	trimethylphenylammonium_bromide
386	ferric_nitrilotriacetate
387	sodium_hydrosulfide
388	potassium_metabisulfite
389	fluoroboric_acid
390	sodium_hexafluorosilicate
391	ammonium_hexafluorosilicate
392	hexachloroiridic_acid
393	tert-butyldimethylsilyl_chloride
394	tin_dioxide
395	8-quinolinesulfonyl_chloride
396	molybdenum_dioxide
397	osmium_tetroxide
398	aluminum_phosphide
399	mercuric_oxide
400	indium_phosphide
401	cetalkonium_chloride
402	triethyl_phosphite
403	dilauryl_thiodipropionate
404	cetyldimethylethylammonium_bromide
405	carbamide_peroxide
406	urea_nitrate
407	tridihexethyl_iodide
408	dimethyldiallylammonium_chloride
409	sodium_azide
410	vanadyl_sulfate
411	diammonium_22'-azino-bis(3-ethylbenzothiazoline-6-sulfonate)
412	tetramethylammonium_hydroxide
413	amyl_nitrate
414	ferric_citrate
415	sodium_molybdate
416	nitrogen_trichloride

417	dipotassium_tetrachloroplatinate
418	tellurium_tetrachloride
419	gadolinium_chloride
420	molybdenum_pentachloride
421	potassium_thiosulfate
422	nitrogen_trioxide
423	nickel_hydroxide
424	lead_selenide
425	arsenic_sulfide
426	nitrogen_triiodide
427	barium_perchlorate
428	gallium_nitrate
429	sodium_orthovanadate
430	pentapotassium_triphosphate
431	bromine_chloride
432	potassium_tetrathionate
433	barium_bromate
434	magnesium_peroxide
435	zirconium_silicate
436	ruthenium_chloride
437	hexachloroplatinic_acid
438	magnesium_fluosilicate
439	barium_oxide
440	zirconium_oxide
441	mercuric_sulfide
442	phenyl_trimethicone
443	octyldimethylamine_oxide
444	ethyl_hydroperoxide
445	vinylsulfonic_acid
446	silver_fulminate
447	tellurium_dioxide
448	stannous_sulfate
449	titanium_trichloride
450	ammonium_persulfate
451	cuprous_chloride
452	sodium_persulfate
453	tripotassium_phosphate
454	phosphorus_trifluoride
455	potassium_selenate
456	potassium_pyrosulfate
457	disulfuric_acid
458	rubidium_chloride
459	dicetyl_peroxydicarbonate

460	2-ethylhexyl_nitrate
461	aluminum_phosphate
462	denatonium_saccharide
463	phosphoramidic_acid
464	propaneperoxoic_acid
465	periodic_acid
466	potassium_tellurite
467	phenethylboronic_acid
468	oxalyl_chloride
469	ethyltrimethylammonium_iodide
470	benzyltriethylammonium_chloride
471	tetramethylammonium_bromide
472	silver_bromide
473	selenium_tetrachloride
474	boron_nitride
475	nitrogen_pentoxide
476	dimethyl_cyanocarbonimidodithioate
477	lithium_sulfate
478	24-diaminobenzenesulfonic_acid
479	phenylboronic_acid
480	benzyltrimethylammonium_hydroxide
481	5-amino-2-methylbenzenesulfonic_acid
482	arsonoacetic_acid
483	phenyltrimethylammonium_chloride
484	tetrabutylammonium_iodide
485	perfluorohexanesulfonic_acid
486	tetramethylammonium_fluoride
487	perfluorobutanesulfonic_acid
488	tetrabutylammonium_tetrafluoroborate
489	thionyl_bromide
490	tetraphenylarsonium_bromide
491	tetraphenylarsonium_chloride
492	dimethyl_telluride
493	dimethyl_sulfite
494	diethyl_telluride
495	trimethylphosphine_oxide
496	tetrapentylammonium_bromide
497	tetrabutylammonium_chloride
498	stearyltrimethylammonium_bromide
499	fubrogonium_iodide
500	penoctonium_bromide
501	diponium_bromide
502	mebezonium_iodide

503	oxitefonium_bromide
504	otilonium_bromide
505	2-nitrophenyl_azide
506	ferric_hydroxide
507	sodium_oxide
508	magnesium_hydroxide
509	uranium_trioxide
510	tetrabutylammonium_bromide
511	tetrabutylammonium_perchlorate
512	tetramethylammonium_nitrate
513	tetrahexylammonium_iodide
514	diallyl_tetrasulfide
515	methyltributylammonium_iodide
516	tert-amyl_hydroperoxide
517	dimethyldioctadecylammonium_bromide
518	phenyltrimethylammonium_tribromide
519	tetrahexylammonium_bromide
520	tetrahexylammonium_perchlorate
521	tetrapentylammonium_chloride
522	4-bromophenylboronic_acid
523	p-tolylboronic_acid
524	tetrasulfide_dimethyl
525	tetrahexylammonium_chloride
526	tributylammonium_chloride
527	tridodecylmethylammonium_chloride
528	trimethylsilyl_cyanide
529	molybdic_acid
530	arsenic_pentafluoride
531	magnesium_bromide
532	platinum_diiodide
533	ruthenium_trichloride
534	triethylmethylammonium_chloride
535	isopropylsulfonyl_chloride
536	strontium_titanate
537	gallium_phosphide
538	4-(methylthio)benzoic_acid
539	hypobromous_acid
540	lanthanum_bromide
541	xenon_difluoride
542	aminomethanesulfonic_acid
543	zirconium_hydroxide
544	calcium_fluoride

545	bis(4-tert-butylcyclohexyl)_peroxydicarbonate
546	tetrahexylammonium_benzoate
547	scandium_hydroxide
548	methanesulfinic_acid
549	manganese_sulfide
550	titanium_hydroxide
551	silver(i)_oxide
552	3-(triethoxysilyl)propyl_methacrylate
553	di-tert-butyl_dicarbonate
554	cadmium_telluride
555	tris(24-di-tert-butylphenyl)_phosphite
556	tributylmethylammonium_chloride
557	1-phenylethyl_hydroperoxide
558	3-aminophenylboronic_acid
559	4-azidophenacyl_bromide
560	glycidyl_nitrate
561	titanium_nitride
562	4-fluoro-3-nitrophenyl_azide
563	phenyltrimethylammonium_iodide
564	4-(phenylsulfanyl)butanoic_acid
565	dibenzyldimethylammonium_chloride
566	hydrogen_sulfite
567	tritium_oxide
568	chlorate_ion
569	cuprous_ion
570	trimethylarsine_oxide
571	potassium_triiodide
572	methacrylamidopropyltrimethylammonium_chloride
573	bismuth_nitrate
574	phosphorous_acid
575	carbon_sulfide
576	sulfur_oxide
577	tetradecylthioacetic_acid
578	selenium_trioxide
579	ethidium_diazide
580	indium_nitride
581	ruthenium_tetraoxide
582	titanium_tetrafluoride
583	magnesium_pyrophosphate
584	niobium_oxide
585	methyl_4-(methylthio)benzoate
586	dimethylboron_bromide

587	nitrosyl_fluoride
588	lead_tetrachloride
589	vanadium_oxytrifluoride
590	hypoiodous_acid
591	peroxynitrous_acid
592	diphenyl_azidophosphate
593	diethylaminosulfur_trifluoride
594	octacalcium_phosphate
595	25-dioxopyrrolidin-1-yl_2-(acetylthio)acetate
596	stannic_fluoride
597	calcium_borate
598	lanthanum_hydroxide
599	2-amino-4-arsonobutanoic_acid
600	carbon_suboxide
601	cyanogen_fluoride
602	methylboronic_acid
603	xenon_hexafluoride
604	sulfur_difluoride
605	samarium_diiodide
606	nitric_oxide
607	sodium_metaborate
608	sodium_dithionate
609	diquafosol_tetrasodium
610	succinimidyl_carbonate
611	3-(dansylamino)phenylboronic_acid
612	gadolinium_oxalate
613	didodecyl_disulfide
614	4-biphenylboronic_acid
615	4-iodophenylboronic_acid
616	phosphorochloridic_acid
617	phosphorodithioic_acid
618	methylsulfenyl_bromide
619	35-bis(trifluoromethyl)phenylboronic_acid
620	bis(244-trimethylpentyl)phosphinic_acid
621	gallium_oxide
622	lanthanum_sulfate
623	dysprosium_oxide
624	neodymium_oxide
625	yttrium_oxide
626	iodine_pentoxide
627	calcium_aluminate
628	barium_hexaferrite

629	ammonium_cyanide
630	copper_silicate
631	cerium_phosphate
632	cerium_sulfate
633	cerium(iv)_sulfate
634	zinc_selenite
635	silver_tetrafluoroborate
636	sodium_percarbonate
637	uranyl_phosphate
638	zinc_dithiophosphate
639	benzyltributylammonium_chloride
640	ferrocene_carbamate
641	tin_sulfate
642	zirconium_phosphate
643	phosphoramidothioic_acid
644	cerium_trichloride
645	ammonium_borate
646	potassium_ferrite
647	indium_formate
648	copper_hydroxide
649	iron_acetate
650	tetrabutylammonium_hexafluorophosphate
651	ferric_thiocyanate
652	ferrous_thiocyanate
653	phosphorodiamidic_acid
654	tetrabutylammonium_cyanide
655	silver_fluoride
656	diphenyl(246-trimethylbenzoyl)phosphine_oxide
657	calcium_selenate
658	magnesium_selenate
659	phosphorothioic_acid
660	aluminum_hexafluorosilicate
661	lanthanum_carbonate
662	tetrabutylammonium_tetrahydroborate
663	molybdenum_blue
664	ruthenium_hexacyanide
665	magnesium_hydroxycarbonate
666	vanadium_disulfide
667	4-azidophenyl_isothiocyanate
668	nickel-palladium_alloy
669	rubidium_sulfate
670	ferrous_gluconate

671	4-methoxyphenylboronic_acid
672	14-benzenediboronic_acid
673	potassium_bromide
674	sodium_bromide
675	methyl_2-(methylthio)benzoate
676	2-formylphenylboronic_acid
677	platinum_dioxide
678	permanganic_acid
679	propane-13-disulfonic_acid
680	hydroxylamine_hydrochloride
681	sodium_arsenite
682	prostaglandin_h2
683	geranyl_diphosphate
684	ammonium_metavanadate
685	potassium_permanganate
686	sodium_bicarbonate
687	potassium_bicarbonate
688	potassium_periodate
689	potassium_perchlorate
690	sodium_chlorate
691	potassium_nitrite
692	sodium_bisulfate
693	potassium_acetate
694	ammonium_carbonate
695	chromium(iii)_oxide
696	ferric_acetate
697	dimanganese_decacarbonyl
698	boric_oxide
699	beryllium_hydroxide
700	dirhenium_decacarbonyl
701	hydroperoxy_radical
702	sodium_perchlorate
703	potassium_fluoride
704	cyclohexyl_nitrite
705	arteannuic_acid
706	3-thiopheneboronic_acid
707	4-formylphenylboronic_acid
708	ferric_gluconate
709	sodium_metabisulfite
710	tetrabutylammonium_hydroxide
711	sodium_benzenesulfinate
712	triethyloxonium_tetrafluoroborate
713	tetrabutylammonium_fluoride

714	prussian_blue
715	tungsten_carbide
716	benzyltributylammonium_bromide
717	2-methylphenylboronic_acid
718	sodium_methanesulfinate
719	chloromethyl_chlorosulfate
720	tetraoctylammonium_bromide
721	1-hexyl-3-methylimidazolium_hexafluorophosphate
722	1-hexyl-3-methylimidazolium_tetrafluoroborate
723	2-acetylphenylboronic_acid
724	23-difluorophenylboronic_acid
725	pyridine-3-boronic_acid
726	pyridine-4-boronic_acid
727	8-quinolineboronic_acid
728	4-vinylphenylboronic_acid
729	tris(2-carboxyethyl)phosphine_hydrochloride
730	4-aminophenylboronic_acid
731	bis(2-methoxyethyl)aminosulfur_trifluoride
732	benzenethiosulfonic_acid
733	trimethyloxonium_tetrafluoroborate
734	dimethyl(methylthio)sulfonium_tetrafluoroborate
735	4-azido-2356-tetrafluorobenzoic_acid
736	2-bromopyridine-5-boronic_acid
737	4-carboxy-3-fluorophenylboronic_acid
738	23456-pentafluorophenyl_1-ethylenesulfonate
739	4-methyl-1-naphthaleneboronic_acid
740	methyltrioctylammonium_iodide
741	3-methacrylamidophenylboronic_acid
742	nitrogen_dioxide
743	ammoniated_mercury
744	4-azidobenzoic_acid
745	sodium_tetrathiocarbonate
746	mercuric_perchlorate
747	chromitope_sodium
748	sodium_peroxocarbonate
749	tantalum;_titanium
750	platinum-iridium_alloy
751	cethexonium_bromide
752	silicon_nitride

753	phosphonic_acid
754	metaphosphoric_acid
755	phenylacetyl_disulfide
756	rubidium_iodide
757	tris(222-trifluoroethyl)_borate
758	chloramine_t
759	niobium_carbide
760	lithium_tetrafluoroborate
761	sodium_borohydride
762	triacetone_triperoxide
763	4-mercaptophenylboronic_acid
764	scandium_oxide
765	zinc_phosphide
766	sodium_niobate
767	quinoline-5-boronic_acid
768	tetrametaphosphate_ion
769	calcium_dichloride
770	tetrafluoroaluminate_ion
771	tin_powder
772	13-hydroperoxy-911-octadecadienoic_acid
773	magnesium_chloride
774	calcium_dibromide
775	methylnaltrexone_bromide
776	iron_fumarate
777	tetrabutylammonium_azide
778	manganese_carbonyl
779	ammonium_sulfate
780	sodium_zirconate
781	calcium_tungstate
782	phenylphosphinic_acid
783	bismuth_oxychloride
784	tetramethylammonium_borohydride
785	magnesium_hexaborate
786	aluminum_boride
787	potassium_chlorate
788	cobalt_tetraoxide
789	ethylammonium_nitrate
790	cuprous_iodide
791	ferrous_fumarate
792	stilonium_iodide
793	dioleyldimethylammonium_chloride
794	tylosin_phosphate
795	clofilium_phosphate
796	ammonium_pertechnetate
797	ammonium_hexafluorophosphate
798	dihexadecyldimethylammonium_bromide
799	zinc_hydroxide
800	lithium_hexafluoroarsenate
801	potassium_phosphite
802	ferrous_bisglycinate
803	bismuth_phosphate
804	drometrizole_trisiloxane
805	ferrous_phosphate
806	sodium_thiophosphate
807	tetrabutylammonium_borohydride
808	dodecyltrimethylammonium_iodide
809	hafnium_chloride
810	246-triphenylpyrylium_tetrafluoroborate
811	aluminum_oxide
812	hexyltrimethylammonium_bromide
813	samarium_cobalt
814	cobalt_hydroxide
815	ammonium_monofluorophosphate
816	yttrium_silicate
817	calcium_bicarbonate
818	1122-tetrafluoroethanesulfonic_acid
819	4-amino-3-fluorophenylboronic_acid
820	ammonium_dinitramide
821	borax_glass
822	cupric_bromide
823	potassium_azide
824	pyridine_borane
825	nickel(iii)_oxide
826	3-acrylamidophenylboronic_acid
827	aluminum_carbonate
828	ammonium_bisulfite
829	ferrous_succinate
830	sodium_tetrathionate
831	artemisinic_acid
832	disodium_tetrachloropalladate
833	1-carbethoxycyclopropyltriphenylphosphonium_tetrafluoroborate
834	dihydroartemisinic_acid
835	nitrosonium_tetrafluoroborate
836	fluosilicic_acid

837	2-picoline_borane
838	sodium_selenosulfate
839	1-butyl-4-methylpyridinium_tetrafluoroborate
840	ferrocyanic_acid
841	nitryl_chloride
842	aluminum_telluride
843	silyl_trifluoromethanesulfonate
844	tetrabutylphosphonium_tetrafluoroborate
845	lithium_borate
846	manganese_silicate
847	silver_carbonate
848	tetrahexylammonium_hydroxide
849	nickel_sulfite
850	ferrous_citrate
851	tetrapentylammonium_hydroxide
852	phytyl_diphosphate
853	ammonium_tetrathiomolybdate
854	zinc_silicate
855	ammonium_chloroplatinate
856	dodecamolybdophosphoric_acid
857	thallium_carbonate
858	calcium_titanate
859	barium_ferrite
860	naloxone_methiodide
861	bismuth_subsalicylate
862	dihydroxyaluminum_aminoacetate
863	antimony_pentasulfide
864	bismuth_subcarbonate
865	aluminum_carbide
866	potassium_tetrafluoroaluminate
867	tetracalcium_phosphate
868	antimony_hydroxide
869	sodium_aluminosilicate
870	sodium_cyanoborohydride
871	bismuth_aluminate
872	fluorosilicic_acid
873	thallium_trifluoroacetate
874	argon_fluoride
875	1-decyl-3-methylimidazolium_bromide
876	aluminum_chlorhydroxide
877	potassium_tetrahydroborate
878	sodium_borohydrate

879	monosodium_methanearsonate
880	sodium_cyanotriphenylborate
881	lithium_niobate
882	potassium_iodate
883	sodium_hypochlorite
884	sodium_bisulfite
885	sodium_periodate
886	sodium_nitrite
887	sodium_bromate
888	sodium_chlorite
889	potassium_chlorite
890	monosodium_phosphate
891	sodium_permanganate
892	potassium_bromate
893	sodium_hypobromite
894	sodium_hypophosphite
895	sodium_iodate
896	sodium_tetraethylborate
897	sodium_tantalate
898	potassium_sulfamate
899	sodium_hydroxymethanesulfinate
900	sodium_ethylxanthate
901	cesium_thiocyanate
902	sodium_difluorophosphate
903	sodium_perbromate
904	gallium_citrate
905	omeprazole/sodium_bicarbonate
906	iron_gluconate
907	lanthanum_permanganate
908	silicotungstic_acid
909	isoquinolin-8-ylboronic_acid
910	sodium_artesunate
911	(diethylamino)difluorosulfonium_tetrafluoroborate
912	1-dodecyl-3-methylimidazolium_tetrafluoroborate
913	ammonium_tetrathiotungstate
914	decyldiethylphosphine_oxide
915	difluoro(morpholino)sulfonium_tetrafluoroborate
916	zinc_aspartate
917	aluminum_hydroxyphosphate
918	aluminum_chlorhydrate
919	americium_dioxide

920	soda_lime
921	artelinic_acid
922	gadofosveset_trisodium
923	chromium(iv)_oxide
924	nepheline_syenite
925	magnesium_aluminosilicate
926	aqua_regia

927	sucroferric_oxyhydroxide
928	titanium_diboride
929	zirconium_diboride
930	tetrapropylammonium_perruthenate
931	hyponitrous_acid
932	lodenafil_carbonate

D3

Dataset contains the three-word (3-grams) names of inorganic compound extracted from PubChem. There are total 102 of them actually appear in the 119,166 patents, so we use them to do the evaluation. These 102 compound names are listed below.

1	boron_trifluoride_etherate
2	zinc_butyl_xanthate
3	potassium_silver_cyanide
4	2-chloroethyl_methyl_sulfide
5	ethyl_methyl_sulfide
6	ethyl_vinyl_sulfide
7	ferric_ammonium_citrate
8	beryllium_aluminum_silicate
9	chloromethyl_methyl_sulfide
10	3-chloro-2-hydroxypropyltrimethyl_ammonium_chloride
11	calcium_hydrogen_phosphate
12	copper_sulfate_pentahydrate
13	aluminum_chloride_hexahydrate
14	aluminum_nitrate_nonahydrate
15	cobalt_nitrate_hexahydrate
16	aluminum_potassium_sulfate
17	ferrous_ammonium_sulfate
18	calcium_sulfate_dihydrate
19	nickel_ammonium_sulfate
20	disodium_phosphate_dodecahydrate
21	trisodium_phosphate_dodecahydrate
22	sodium_thiosulfate_pentahydrate
23	chromium_potassium_sulfate
24	chlorinated_trisodium_phosphate
25	nickel_nitrate_hexahydrate
26	sodium_metasilicate_nonahydrate
27	sodium_sulfate_decahydrate
28	ferrous_sulfate_heptahydrate
29	ranitidine_bismuth_citrate
30	allyl_methyl_sulfide
31	ethyl_dihydrogen_phosphate
32	basic_blue_7
33	tert-butyl_phenyl_carbonate
34	tetramethylammonium_hydroxide_pentahydrate
35	methyl_p-nitrophenyl_carbonate
36	tetrabutylammonium_hydrogen_sulfate
37	potassium_aluminum_silicate
38	magnesium_hydrogen_phosphate
39	sodium_tungstate_dihydrate
40	sodium_carbonate_decahydrate
41	potassium_titanium_oxide
42	zirconyl_chloride_octahydrate
43	ammonium_dihydrogen_citrate
44	lanthanum_chloride_heptahydrate
45	calcium_magnesium_phosphate
46	aluminum_magnesium_hydroxide
47	manganese_sulfate_monohydrate
48	ammonium_ferric_sulfate
49	zinc_perchlorate_hexahydrate
50	phosphonitrilic_chloride_trimer
51	potassium_dihydrogen_phosphate
52	manganese_chloride_tetrahydrate
53	cystine_dimethyl_ester
54	potassium_ethyl_xanthate
55	aluminum_sodium_phosphate
56	calcium_sulfate_hemihydrate
57	aluminum_magnesium_silicate
58	sodium_pyrophosphate_decahydrate
59	calcium_citrate_tetrahydrate
60	diammonium_hydrogen_citrate
61	nickel_sulfate_hexahydrate
62	magnesium_dihydrogen_phosphate
63	strontium_chloride_hexahydrate
64	aluminum_hydroxide_hydrate
65	aluminum_magnesium_oxide
66	magnesium_iron_silicate
67	artemether_and_lumefantrine
68	aluminum_chloride_anhydrous
69	cadmium_acetate_dihydrate
70	ammonium_cerium(iv)_nitrate
71	dried_aluminium_hydroxide
72	nickel_chloride_dihydrate

73	lithium_aluminium_deuteride
74	triethylammonium_hydrogen_carbonate
75	ferrocene_monocarboxylic_acid
76	ammonium_citrate_dibasic
77	cetyl_ammonium_bromide
78	zinc_nitrate_hexahydrate
79	calcium_sodium_metaphosphate
80	barium_hydroxide_monohydrate
81	ferric_nitrate_nonahydrate
82	calcium_nitrate_tetrahydrate
83	copper_iron_oxide
84	barium_hydroxide_hydrate
85	barium_hydroxide_octahydrate
86	phenyltrimethylammonium_bromide_dibromide
87	zinc_phosphate_tetrahydrate

88	potassium_magnesium_citrate
89	manganese_sulfate_tetrahydrate
90	acid_blue_129
91	acidulated_phosphate_fluoride
92	rubidium_hydrogen_sulfate
93	dihydroxyaluminum_sodium_carbonate
94	potassium_hydroxide_monohydrate
95	ammonium_ferric_citrate
96	ferric_citrate_hydrate
97	barium_perchlorate_trihydrate
98	ferric_sulfate_heptahydrate
99	ferric_oxide_red
100	lithium_manganese_oxide
101	nickel_zinc_ferrite
102	basic_bismuth_nitrate

D4

Dataset contains the three-word (4-grams) names of inorganic compound extracted from PubChem. There are total 13 of them actually appear in the 119,166 patents, so we use them to do the evaluation. These 13 compound names are listed below.

1	didecyl_dimethyl_ammonium_chloride
2	aluminum_potassium_sulfate_dodecahydrate
3	calcium_hydrogen_phosphate_dihydrate
4	methylsulfamic_acid_3-(2-methoxyphenoxy)-2-(((methylamino)sulfonyl)oxy)propyl_ester
5	boron_trifluoride_diethyl_etherate
6	phosphonic_acid_dimethyl_ester
7	phosphonic_acid_bis(1-methylethyl)_ester
8	terephthalylidene_dicamphor_sulfonic_acid
9	menthyl_ethylene_glycol_carbonate
10	magnesium_ammonium_phosphate_hexahydrate
11	dodecyl_trimethyl_ammonium_hydroxide
12	sodium_dihydrogen_phosphate_dihydrate
13	lithium_magnesium_sodium_silicate

D5

Dataset contains the names of “lithium battery” related terms manually extracted from Wikipedia. There are total 42 of them actually appear in the 119,166 patents, so we use them to do the evaluation. These 42 compound names are listed below.

1	lithium_battery
2	lithium-ion_battery
3	li-ion_battery
4	rechargeable_battery
5	portable_electronics
6	electric_vehicles
7	lithium_ions
8	negative_electrode
9	positive_electrode
10	intercalated_lithium_compound
11	electrode_material
12	metallic_lithium
13	non-rechargeable_lithium_battery
14	energy_density
15	memory_effect
16	flammable_electrolyte
17	electric_tool
18	medical_equipment
19	life_extension
20	charging_speed
21	non-flammable_electrolyte
22	electrochemical_unit

23	rechargeable_cells
24	lithium-ion_cells
25	lithium_metal
26	electrochemical_intercalation
27	organic_electrolytes
28	lithium_polymer_battery
29	lithium-ion_polymer_battery
30	lithium_salt
31	lithium_ion_diffusion
32	lithium_atom
33	lithium_anode
34	self_discharge
35	battery_life
36	lithium_plating
37	lithium_content
38	bipolar_battery
39	solid-state_battery
40	lithium-ion_capacitor
41	lithium-air_battery
42	solid-state_lithium-ion_battery

D6

Dataset contains the names of “unit operations & separation process” terms manually extracted from Wikipedia. There are total 36 of them actually appear in the 119,166 patents, so we use them to do the evaluation. These 36 compound names are listed below.

1	fluid_flow
2	heat_transfer
3	heat_exchange
4	mass_transfer
5	gas_absorption
6	gas_liquefaction
7	chemical_reaction
8	reactive_distillation
9	catalytic_reactions
10	cyclonic_separation
11	thin-layer_chromatography
12	high-performance_liquid_chromatography
13	countercurrent_chromatography
14	droplet_countercurrent_chromatography
15	paper_chromatography
16	ion_chromatography
17	size-exclusion_chromatography
18	affinity_chromatography
19	centrifugal_partition_chromatography
20	gas_chromatography
21	inverse_gas_chromatography
22	capillary_electrophoresis
23	electrostatic_separation
24	liquid-liquid_extraction
25	solid_phase_extraction
26	supercritical_fluid_extraction
27	field_flow_fractionation
28	froth_flotation
29	dissolved_air_flotation
30	fractional_distillation
31	fractional_freezing
32	oil-water_separation
33	magnetic_separation
34	gravity_separation
35	vapor-liquid_separation
36	zone_refining

D1*, D2*, and D3*

D1*, D2*, and D3* are specially crafted datasets contain compound names, PubChem formulas, and commonly used formulas based on D1, D2, and D3, respectively. PubChem formulas were extracted using PubChem API. Commonly used formulas were manually selected and matched to the compound names. D1* has 18 records, D2* has 75 records, and D3* has 20 records.

D1*

	Compound_Name	PubChem_Formula	General_Formula
1	lithium_aluminate	allio2	lialo2
2	lithium_aluminium_hydride	alh4li	lialh4
3	lithium_bicarbonate	chlio3	lihco3
4	lithium_bromide	brli	libr
5	lithium_chloride	clli	licl
6	lithium_cyanide	clin	licn
7	lithium_dihydrogen_phosphate	h2lio4p	lih2po4
8	lithium_fluoride	fli	lif
9	lithium_hexafluorophosphate	f6lip	lipf6
10	lithium_hydroxide	hlio	lioh
11	lithium_hydroxide_monohydrate	h3lio2	lioh.h2o
12	lithium_hypochlorite	cllio	liclo
13	lithium_iodide	ili	lii
14	lithium_metaborate	blio2	libo2
15	lithium_nitride	h2li3n+2	li3n
16	lithium_perchlorate	cllio4	liclo4
17	lithium_tetrahydridoaluminate	alh4li	lialh4
18	lithium_trifluoromethanesulfonate	cf3lio3s	cf3so3li

D2*

	Compound_Name	PubChem_Formula	General_Formula
1	zinc_dust	zn	zn
2	zinc_sulfide	szn	zns
3	tin_powder	sn	sn
4	lead_selenide	pbse	pbse
5	phosphorus_pentasulfide	p4s10	p4s10
6	zinc_oxide	ozn	zno
7	zinc_phosphate	o8p2zn3	zn3(po4)2
8	vanadium_pentoxide	o5v2	v2o5
9	antimony_pentoxide	o5sb2	sb2o5
10	zinc_sulfate	o4szn	znso4

11	zirconium_silicate	o4sizr	zrsio4
12	zinc_silicate	o4sizn2	zn2sio4
13	ruthenium_tetraoxide	o4ru	ruo4
14	osmium_tetroxide	o4os	oso4
15	yttrium_oxide	o3y2	y2o3
16	tungsten_trioxide	o3w	wo3
17	strontium_titanate	o3srti	srtio3
18	scandium_oxide	o3sc2	sc2o3
19	sulfur_trioxide	o3s	so3
20	zirconium_oxide	o2zr	zro2
21	titanium_dioxide	o2ti	tio2
22	tellurium_dioxide	o2te	teo2
23	tin_dioxide	o2sn	sno2
24	selenium_dioxide	o2se	seo2
25	sulfur_dioxide	o2s	so2
26	platinum_dioxide	o2pt	pto2
27	lead_dioxide	o2pb	pbo2
28	titanium_nitride	nti	tin
29	nitrogen_dioxide	no2	no2
30	sodium_nitrate	nnao3	nano3
31	sodium_nitrite	nnao2	nano2
32	nickel_sulfate	nio4s	niso4
33	neodymium_oxide	nd2o3	nd2o3
34	niobium_oxide	nb2o5	nb2o5
35	sodium_niobate	nanbo3	nanbo3
36	tetrasodium_pyrophosphate	na4o7p2	na4p2o7
37	sodium_trimetaphosphate	na3o9p3	na3p3o9
38	sodium_orthovanadate	na3o4v	na3vo4
39	trisodium_phosphate	na3o4p	na3po4
40	sodium_persulfate	na2o8s2	na2s2o8
41	sodium_metabisulfite	na2o5s2	na2s2o5
42	sodium_tungstate	na2o4w	na2wo4
43	sodium_hydrosulfite	na2o4s2	na2s2o5
44	sodium_sulfate	na2o4s	na2so4
45	sodium_silicate	na2o3si	na2sio3
46	sodium_thiosulfate	na2o3s2	na2s2o4
47	sodium_sulfite	na2o3s	na2so3
48	sodium_peroxide	na2o2	na2o2
49	silicon_nitride	n4si3	si3n4
50	sodium_azide	n3na	nan3
51	zinc_nitrate	n2o6zn	zn(no3)2
52	strontium_nitrate	n2o6sr	sr(no3)2
53	palladium_nitrate	n2o6pd	pd(no3)2

54	lead_nitrate	n2o6pb	pb(no3)2
55	nitrous_oxide	n2o	n2o
56	nickel_nitrate	n2nio6	ni(no3)2
57	molybdenum_disulfide	mos2	mos2
58	molybdenum_trioxide	moo3	moo3
59	molybdenum_dioxide	moo2	moo2
60	manganese_sulfate	mno4s	mnso4
61	manganese_dioxide	mno2	mno2
62	sodium_permanganate	mnnao4	namno4
63	magnesium_sulfate	mgo4s	mgso4
64	magnesium_oxide	mgo	mgo
65	magnesium_nitrate	mgn2o6	mg(no3)2
66	trimagnesium_phosphate	mg3o8p2	mg3(po4)2
67	lithium_niobate	linbo3	linbo3
68	lithium_sulfate	li2o4s	li2so4
69	potassium_nitrate	kno3	kno3
70	potassium_permanganate	kmno4	kmno4
71	tripotassium_phosphate	k3o4p	k3po4
72	potassium_persulfate	k2o8s2	k2s2o8
73	potassium_sulfate	k2o4s	k2so4
74	platinum-iridium_alloy	irpt	ptir
75	rubidium_iodide	irb	rbi

D3*

	Compound_Name	PubChem_Formula	General_Formula
1	lithium_manganese_oxide	limn2o4	limn2o4
2	sodium_tungstate_dihydrate	h4na2o6w	na2wo4.2h2o
3	potassium_dihydrogen_phosphate	h2ko4p	kh2po4
4	sodium_sulfate_decahydrate	h20na2o14s	na2so4.10h2o
5	zinc_nitrate_hexahydrate	h12n2o12zn	zn(no3)2.6h2o
6	nickel_nitrate_hexahydrate	h12n2nio12	ni(no3)2.6h2o
7	sodium_thiosulfate_pentahydrate	h10na2o8s2	na2s2o3.5h2o
8	ferric_nitrate_nonahydrate	feh18n3o18	fe(no3)3.9h2o
9	ferrous_sulfate_heptahydrate	feh14o11s	feso4.7h2o
10	copper_sulfate_pentahydrate	cuh10o9s	cuso4.5h2o
11	cobalt_nitrate_hexahydrate	coh12n2o12	co(no3)2.6h2o
12	manganese_chloride_tetrahydrate	cl2h8mno4	mncl2.4h2o
13	sodium_carbonate_decahydrate	ch20na2o13	na2co3.10h2o
14	ammonium_cerium(iv)_nitrate	ceh4n6o15	(nh4)2ce(no3)6
15	calcium_hydrogen_phosphate	caho4p	cahpo4
16	calcium_nitrate_tetrahydrate	cah8n2o10	ca(no3)2.4h2o
17	barium_hydroxide_octahydrate	bah18o10	ba(oh)2.8h2o

18	lithium_aluminium_deuteride	alh4li	lialD4
19	aluminum_nitrate_nonahydrate	alh18n3o18	al(no3)3.9h2o
20	aluminum_chloride_hexahydrate	alcl3h12o6	alcl3.6h2o

Dataset for Clustering Evaluations

Dataset used for clustering evaluation, and it is composed of a total of 72 examples evenly sub-sampled from D1, D5 and D6 (24 examples each). There's one restriction applied here, which is the example chemical phrase names and their component words should be included in the vocabularies of two approaches $\mathbf{W}_c$ and $\mathbf{W}_p$.

	Chemical Phrases	Origin Dataset	Label (Self-assigned)
1	lithium_hydroxide	D1	0
2	lithium_tetrahydridoaluminate	D1	0
3	lithium_iodide	D1	0
4	lithium_bromide	D1	0
5	lithium_aluminate	D1	0
6	lithium_metaborate	D1	0
7	lithium_hydroxide_monohydrate	D1	0
8	lithium_fluoride	D1	0
9	lithium_chloride	D1	0
10	lithium_stearate	D1	0
11	lithium_diisopropylamide	D1	0
12	lithium_bis(trimethylsilyl)amide	D1	0
13	lithium_acetate	D1	0
14	lithium_trifluoromethanesulfonate	D1	0
15	lithium_aluminium_hydride	D1	0
16	lithium_tert-butoxide	D1	0
17	lithium_perchlorate	D1	0
18	lithium_hypochlorite	D1	0
19	lithium_acetate_dihydrate	D1	0
20	lithium_dihydrogen_phosphate	D1	0
21	lithium_isopropoxide	D1	0
22	lithium_lactate	D1	0
23	lithium_bicarbonate	D1	0
24	lithium_hexafluorophosphate	D1	0
25	lithium_battery	D5	1
26	lithium-ion_battery	D5	1
27	li-ion_battery	D5	1
28	rechargeable_battery	D5	1
29	bipolar_battery	D5	1
30	solid-state_battery	D5	1
31	lithium_ions	D5	1
32	negative_electrode	D5	1
33	positive_electrode	D5	1
34	metallic_lithium	D5	1

35	energy_density	D5	1
36	lithium_plating	D5	1
37	lithium-air_battery	D5	1
38	lithium-ion_capacitor	D5	1
39	charging_speed	D5	1
40	non-flammable_electrolyte	D5	1
41	rechargeable_cells	D5	1
42	lithium-ion_cells	D5	1
43	lithium_metal	D5	1
44	electrochemical_intercalation	D5	1
45	lithium_salt	D5	1
46	lithium_ion_diffusion	D5	1
47	lithium_anode	D5	1
48	self_discharge	D5	1
49	fluid_flow	D6	2
50	heat_transfer	D6	2
51	heat_exchange	D6	2
52	mass_transfer	D6	2
53	chemical_reaction	D6	2
54	reactive_distillation	D6	2
55	catalytic_reactions	D6	2
56	thin-layer_chromatography	D6	2
57	high-performance_liquid_chromatography	D6	2
58	countercurrent_chromatography	D6	2
59	ion_chromatography	D6	2
60	size-exclusion_chromatography	D6	2
61	affinity_chromatography	D6	2
62	gas_chromatography	D6	2
63	capillary_electrophoresis	D6	2
64	electrostatic_separation	D6	2
65	liquid-liquid_extraction	D6	2
66	solid_phase_extraction	D6	2
67	supercritical_fluid_extraction	D6	2
68	field_flow_fractionation	D6	2
69	froth_flotation	D6	2
70	fractional_distillation	D6	2
71	oil-water_separation	D6	2
72	magnetic_separation	D6	2

Dataset for Synonyms Finding

Dataset used for synonym finding application, and it is composed of a total of 30 examples randomly and evenly sampled from D1, D5 and D6 (10 examples each). There's one restriction applied here, which is the example chemical phrase names and their component words should be included in the vocabulary of $\mathbf{W}_p$. Top 10 nearest neighbor words of target words and similarities between them are shown below.

Target Words	Rank	Synonyms Candidates	Cosine Similarity
lithium_aluminium_hydride	1	lithium_aluminum_hydride	0.9288
lithium_aluminium_hydride	2	lithium_borohydride	0.9078
lithium_aluminium_hydride	3	borane-tetrahydrofuran_complex	0.8682
lithium_aluminium_hydride	4	lithium_aluminiumhydride	0.8480
lithium_aluminium_hydride	5	borane-thf_complex	0.8477
lithium_aluminium_hydride	6	sodium_borohydride	0.8423
lithium_aluminium_hydride	7	diisobutylaluminium_hydride	0.8391
lithium_aluminium_hydride	8	borane_tetrahydrofuran_complex	0.8386
lithium_aluminium_hydride	9	lialh	0.8323
lithium_aluminium_hydride	10	lialh4	0.8292
lithium_tetrahydridoaluminate	1	bensophenone	0.7517
lithium_tetrahydridoaluminate	2	lithium_aluminum_hydride	0.7441
lithium_tetrahydridoaluminate	3	lithium_aluminium_hydride	0.7035
lithium_tetrahydridoaluminate	4	lithium_borohydride	0.6979
lithium_tetrahydridoaluminate	5	borane-methyl_sulfide_complex	0.6525
lithium_tetrahydridoaluminate	6	hydride_g	0.6517
lithium_tetrahydridoaluminate	7	borane-dimethyl_sulfide_complex	0.6504
lithium_tetrahydridoaluminate	8	lialh	0.6489
lithium_tetrahydridoaluminate	9	sodium_borohydride	0.6471
lithium_tetrahydridoaluminate	10	borane-tetrahydrofuran_complex	0.6438
lithium_bicarbonate	1	lithium_carbonate_sodium_carbonate	0.8796
lithium_bicarbonate	2	sodium_bicarbonate_potassium_bicarbonate	0.8759
lithium_bicarbonate	3	potassium_hydroxide_cesium_hydroxide	0.8603
lithium_bicarbonate	4	sodium_hydrogencarbonate_potassium_hydrogencarbonate	0.8361
lithium_bicarbonate	5	lithium_hydroxide_sodium_hydroxide	0.8355
lithium_bicarbonate	6	sodium_hydroxide_potassium_hydroxide	0.8348
lithium_bicarbonate	7	sodium_hydrogen_carbonate_potassium	0.8341
lithium_bicarbonate	8	sodium_carbonate_potassium_carbonate	0.8328
lithium_bicarbonate	9	lithium_hydrogen_carbonate	0.8328
lithium_bicarbonate	10	alkali_metal_hydroxides	0.8318
lithium_acetate_dihydrate	1	magnesium_acetate_tetrahydrate	0.7936
lithium_acetate_dihydrate	2	magnesium_nitrate_hexahydrate	0.7337

lithium_acetate_dihydrate	3	lithium_nitrate	0.7333
lithium_acetate_dihydrate	4	lithium_sulfate	0.7279
lithium_acetate_dihydrate	5	zinc_acetate	0.7169
lithium_acetate_dihydrate	6	magnesium_formate	0.7116
lithium_acetate_dihydrate	7	calcium_acetate	0.7088
lithium_acetate_dihydrate	8	acetate_tetrahydrate	0.7078
lithium_acetate_dihydrate	9	dihydrate_potassium	0.7042
lithium_acetate_dihydrate	10	hexahydrate	0.7041
lithium_lactate	1	lithium_citrate	0.7560
lithium_lactate	2	lithium_benzoate	0.7239
lithium_lactate	3	lithium_oxalate	0.7196
lithium_lactate	4	lithium_sulfate	0.7086
lithium_lactate	5	lithium_nitrate_lithium	0.7054
lithium_lactate	6	oxalate_lithium	0.7048
lithium_lactate	7	lithium_carbonate_lithium_nitrate	0.6833
lithium_lactate	8	lithium_formate	0.6747
lithium_lactate	9	lithium_dihydrogen_phosphate	0.6740
lithium_lactate	10	formate_lithium	0.6710
lithium_isopropoxide	1	lithium_methoxide_lithium_ethoxide	0.8663
lithium_isopropoxide	2	lithium_propoxide	0.8530
lithium_isopropoxide	3	sodium_propoxide	0.8498
lithium_isopropoxide	4	potassium_propoxide	0.8495
lithium_isopropoxide	5	sodium_isopropoxide	0.8490
lithium_isopropoxide	6	potassium_n-propoxide	0.8418
lithium_isopropoxide	7	sodium_isopropoxide_potassium_isopropoxide	0.8337
lithium_isopropoxide	8	potassium_methoxide_potassium_ethoxide	0.8326
lithium_isopropoxide	9	lithium_phenoxide	0.8318
lithium_isopropoxide	10	sodium_n-propoxide	0.8317
lithium_metaborate	1	lithium_tetraborate	0.7388
lithium_metaborate	2	lithium_oxalate	0.6813
lithium_metaborate	3	lithium_nitrate	0.6803
lithium_metaborate	4	pentaborate	0.6714
lithium_metaborate	5	hexaborate	0.6661
lithium_metaborate	6	strontium_hydroxide	0.6640
lithium_metaborate	7	barium_acetate	0.6601
lithium_metaborate	8	cesium	0.6589
lithium_metaborate	9	potassium_metaborate	0.6471
lithium_metaborate	10	potassium	0.6426
lithium_hexafluorophosphate	1	supporting_salt	0.8676
lithium_hexafluorophosphate	2	lipf	0.8573
lithium_hexafluorophosphate	3	lipf6	0.8569
lithium_hexafluorophosphate	4	libf4	0.8398
lithium_hexafluorophosphate	5	electrolyte_salt	0.8302

lithium_hexafluorophosphate	6	ethylene_carbonate_diethyl_carbonate	0.8238
lithium_hexafluorophosphate	7	mixed_solvent_propylene_carbonate	0.8184
lithium_hexafluorophosphate	8	lithium_phosphate_hexafluoride	0.8159
lithium_hexafluorophosphate	9	mixed_solvent_ethylene_carbonate	0.8154
lithium_hexafluorophosphate	10	libf	0.8154
lithium_trifluoromethanesulfonate	1	lithium_hexafluoroarsenate	0.8451
lithium_trifluoromethanesulfonate	2	lithium_perchlorate	0.8192
lithium_trifluoromethanesulfonate	3	lithium_bis(trifluoromethanesulfonyl)imide	0.8057
lithium_trifluoromethanesulfonate	4	bistrifluoromethylsulfonylimide_lithium	0.7991
lithium_trifluoromethanesulfonate	5	lithium_hexafluoroantimonate	0.7980
lithium_trifluoromethanesulfonate	6	lithium_trifluoromethane_sulfonate	0.7909
lithium_trifluoromethanesulfonate	7	lithium_perchlorate_liclo4	0.7852
lithium_trifluoromethanesulfonate	8	lithium_borofluoride	0.7810
lithium_trifluoromethanesulfonate	9	litfsi_lithium	0.7740
lithium_trifluoromethanesulfonate	10	lithium_trifluoromethasulfonate	0.7713
lithium_bis(trimethylsilyl)amide	1	lithium_diisopropylamide	0.9353
lithium_bis(trimethylsilyl)amide	2	lithium_hexamethyldisilazide	0.9315
lithium_bis(trimethylsilyl)amide	3	sodium_bis(trimethylsilyl)amide	0.9155
lithium_bis(trimethylsilyl)amide	4	lihmds	0.9092
lithium_bis(trimethylsilyl)amide	5	lithium_diisopropyl_amide	0.9053
lithium_bis(trimethylsilyl)amide	6	potassium_bis(trimethylsilyl)amide	0.9017
lithium_bis(trimethylsilyl)amide	7	potassium_tert-butoxide	0.9004
lithium_bis(trimethylsilyl)amide	8	sodium_hexamethyldisilazide	0.9003
lithium_bis(trimethylsilyl)amide	9	lithium_diisopropylamine	0.8975
lithium_bis(trimethylsilyl)amide	10	potassium_hexamethyldisilazide	0.8876
lithium-ion_cells	1	lithium-ion_batteries	0.8704
lithium-ion_cells	2	rechargeable_lithium	0.8541
lithium-ion_cells	3	lithium_ion_batteries	0.8532
lithium-ion_cells	4	lithium-ion	0.8508
lithium-ion_cells	5	li-ion	0.8485
lithium-ion_cells	6	lithium_batteries	0.8444
lithium-ion_cells	7	lithium-ion_battery	0.8371
lithium-ion_cells	8	li-ion_battery	0.8323
lithium-ion_cells	9	rechargeable	0.8322
lithium-ion_cells	10	li-ion_batteries	0.8279
lithium-air_battery	1	air_battery	0.8020
lithium-air_battery	2	lithium_battery	0.7974
lithium-air_battery	3	lithium_ion_battery	0.7943
lithium-air_battery	4	metal-air_battery	0.7930
lithium-air_battery	5	electrochemical_device	0.7919
lithium-air_battery	6	lithium-ion_battery	0.7906
lithium-air_battery	7	lithium-sulfur_battery	0.7831
lithium-air_battery	8	lithium_metal_battery	0.7808

lithium-air_battery	9	lithium_ion	0.7791
lithium-air_battery	10	li-ion_battery	0.7604
life_extension	1	life_span	0.5476
life_extension	2	long_term	0.4918
life_extension	3	key_factors	0.4912
life_extension	4	life-span	0.4867
life_extension	5	pollution	0.4844
life_extension	6	become_important	0.4821
life_extension	7	environmental_friendliness	0.4807
life_extension	8	prerequisite	0.4806
life_extension	9	gaining	0.4762
life_extension	10	collagen_antibody-induced	0.4759
portable_electronics	1	electric_vehicles	0.8681
portable_electronics	2	consumer_electronics	0.8637
portable_electronics	3	portable_electronic_devices	0.8301
portable_electronics	4	power_tools	0.8281
portable_electronics	5	electric_cars	0.8258
portable_electronics	6	uninterruptible_power_supplies	0.8244
portable_electronics	7	electrical_vehicles	0.8221
portable_electronics	8	laptop_computers	0.8200
portable_electronics	9	cell_phones	0.8093
portable_electronics	10	portable_devices	0.8035
energy_density	1	energy_densities	0.8844
energy_density	2	higher_energy_density	0.8571
energy_density	3	power_density	0.8321
energy_density	4	volumetric_energy_density	0.8264
energy_density	5	high_energy_density	0.8225
energy_density	6	wh_kg	0.8096
energy_density	7	rate_capability	0.8053
energy_density	8	capacity_battery	0.8052
energy_density	9	cycle_life	0.8031
energy_density	10	long_cycle_life	0.8021
electrochemical_intercalation	1	lithium_insertion	0.7556
electrochemical_intercalation	2	lithium_intercalation	0.7399
electrochemical_intercalation	3	delithiation	0.7241
electrochemical_intercalation	4	lithium_insertion_extraction	0.7152
electrochemical_intercalation	5	intercalation_compound	0.7110
electrochemical_intercalation	6	intercalation	0.7095
electrochemical_intercalation	7	deintercalation	0.7094
electrochemical_intercalation	8	intercalation_lithium	0.7070
electrochemical_intercalation	9	electrochemical_cycling	0.7040
electrochemical_intercalation	10	li_ion	0.7011
lithium_ion_diffusion	1	diffusion_lithium_ions	0.7752

lithium_ion_diffusion	2	diffusion_paths	0.7614
lithium_ion_diffusion	3	electronic_conductivity	0.7604
lithium_ion_diffusion	4	li_ions	0.7366
lithium_ion_diffusion	5	intercalation_deintercalation	0.7360
lithium_ion_diffusion	6	li_intercalation	0.7330
lithium_ion_diffusion	7	intercalation_deintercalation_lithium_ions	0.7274
lithium_ion_diffusion	8	lithium_diffusion	0.7253
lithium_ion_diffusion	9	rate_performance	0.7206
lithium_ion_diffusion	10	cyclability	0.7196
lithium_plating	1	sei_layer	0.7336
lithium_plating	2	capacity_loss	0.7195
lithium_plating	3	dendrite_growth	0.7181
lithium_plating	4	self-discharge	0.7175
lithium_plating	5	battery_discharge	0.7171
lithium_plating	6	lithium_anode	0.7160
lithium_plating	7	overcharge	0.7105
lithium_plating	8	electrolyte_decomposition	0.7078
lithium_plating	9	charge_discharge_cycles	0.7047
lithium_plating	10	deep_discharge	0.7007
lithium_salt	1	electrolyte_salt	0.7944
lithium_salt	2	supporting_salt	0.7652
lithium_salt	3	lithium_salt_dissolved_in	0.7519
lithium_salt	4	lipf	0.7453
lithium_salt	5	lithium_hexafluorophosphate	0.7428
lithium_salt	6	lithium_salts	0.7373
lithium_salt	7	libf	0.7361
lithium_salt	8	electrolytic_salt	0.7358
lithium_salt	9	lithium_hexafluorophosphate_lipf6	0.7337
lithium_salt	10	electrolyte_lithium_salt	0.7272
battery_life	1	battery	0.7519
battery_life	2	recharge	0.7272
battery_life	3	life_battery	0.7236
battery_life	4	capacity_battery	0.7222
battery_life	5	battery_capacity	0.7177
battery_life	6	power_capability	0.7106
battery_life	7	battery_voltage	0.7098
battery_life	8	battery_charging	0.7050
battery_life	9	self-discharge	0.7029
battery_life	10	remaining_capacity	0.6976
fractional_distillation	1	distillation	0.8796
fractional_distillation	2	vacuum_distillation	0.8340
fractional_distillation	3	simple_distillation	0.8221
fractional_distillation	4	distilling	0.7825

fractional_distillation	5	flash_distillation	0.7751
fractional_distillation	6	extraction_distillation	0.7736
fractional_distillation	7	distillation_residue	0.7689
fractional_distillation	8	short_path_distillation	0.7679
fractional_distillation	9	solvent_extraction	0.7630
fractional_distillation	10	extractive_distillation	0.7605
heat_exchange	1	heat_exchanger	0.8547
heat_exchange	2	heat_exchangers	0.8306
heat_exchange	3	heat_transfer	0.8012
heat_exchange	4	indirect_heat_exchange	0.7941
heat_exchange	5	boiler	0.7901
heat_exchange	6	cooling_medium	0.7861
heat_exchange	7	steam_generator	0.7774
heat_exchange	8	heat_removal	0.7692
heat_exchange	9	compressor	0.7685
heat_exchange	10	cooler	0.7671
catalytic_reactions	1	catalysts	0.7730
catalytic_reactions	2	catalysis	0.7531
catalytic_reactions	3	heterogeneous_catalysts	0.7362
catalytic_reactions	4	hydrogenation_reactions	0.7334
catalytic_reactions	5	oxidative_dehydrogenation	0.7213
catalytic_reactions	6	ocm_reaction	0.7190
catalytic_reactions	7	catalytic_processes	0.7152
catalytic_reactions	8	oxidation_reactions	0.7073
catalytic_reactions	9	metal_catalysts	0.6981
catalytic_reactions	10	catalytic	0.6980
chemical_reaction	1	chemical_reactions	0.7371
chemical_reaction	2	chemical_transformation	0.6625
chemical_reaction	3	chemical_bonds	0.6596
chemical_reaction	4	reactive	0.6425
chemical_reaction	5	reaction_taking_place	0.6404
chemical_reaction	6	chemical_species	0.6376
chemical_reaction	7	reactive_site	0.6353
chemical_reaction	8	capable_undergoing	0.6278
chemical_reaction	9	reactions_occur	0.6275
chemical_reaction	10	physical_chemical	0.6267
magnetic_separation	1	using_antibody-coated_magnetic	0.6705
magnetic_separation	2	paramagnetic_beads	0.6437
magnetic_separation	3	magnetic_beads	0.6434
magnetic_separation	4	macs	0.6329
magnetic_separation	5	magnetic_separator	0.6304
magnetic_separation	6	paramagnetic_particles	0.6288
magnetic_separation	7	filtration_centrifugation	0.6272

magnetic_separation	8	centrifugation	0.6234
magnetic_separation	9	particle_concentrator	0.6213
magnetic_separation	10	centrifugation_filtration	0.6195
reactive_distillation	1	reactive_distillation_column	0.7872
reactive_distillation	2	distillation_column	0.7860
reactive_distillation	3	dialkyl_carbonate	0.7544
reactive_distillation	4	distillation_columns	0.7412
reactive_distillation	5	light_ends	0.7406
reactive_distillation	6	flash_distillation	0.7403
reactive_distillation	7	low_boilers	0.7389
reactive_distillation	8	rich_stream	0.7356
reactive_distillation	9	multi-stage_distillation	0.7345
reactive_distillation	10	distillation	0.7305
electrostatic_separation	1	fluid_porous_structure	0.7386
electrostatic_separation	2	desalting_process_may_include	0.6755
electrostatic_separation	3	filtered_diatomaceous	0.5619
electrostatic_separation	4	portion_milled_grinding	0.5240
electrostatic_separation	5	membrane_separation	0.5080
electrostatic_separation	6	hydrocyclones	0.5055
electrostatic_separation	7	filtration_sedimentation	0.5008
electrostatic_separation	8	liquid-liquid_extraction	0.4926
electrostatic_separation	9	chromatography_electrophoresis	0.4918
electrostatic_separation	10	separation	0.4911
affinity_chromatography	1	immunoaffinity_chromatography	0.8212
affinity_chromatography	2	ion-exchange_chromatography	0.8185
affinity_chromatography	3	gel_filtration	0.8173
affinity_chromatography	4	affinity_chromatography_using	0.8114
affinity_chromatography	5	affinity_purification	0.8031
affinity_chromatography	6	affinity_column	0.8007
affinity_chromatography	7	hydrophobic_interaction_chromatography	0.7985
affinity_chromatography	8	ion_exchange_chromatography	0.7914
affinity_chromatography	9	ammonium_sulfate_precipitation	0.7869
affinity_chromatography	10	affinity_columns	0.7794
gravity_separation	1	filtration_centrifugation	0.6646
gravity_separation	2	settling	0.6640
gravity_separation	3	flotation	0.6347
gravity_separation	4	separation_vessel	0.6347
gravity_separation	5	solids-liquid	0.6342
gravity_separation	6	decanting	0.6287
gravity_separation	7	liquid-liquid_extraction	0.6276
gravity_separation	8	liquid_phase	0.6266
gravity_separation	9	liquid_phases	0.6226
gravity_separation	10	membrane-assisted	0.6219

high-performance_liquid_chromatography	1	high_performance_liquid_chromatography	0.8702
high-performance_liquid_chromatography	2	liquid_chromatography	0.8325
high-performance_liquid_chromatography	3	high_pressure_liquid_chromatography	0.8260
high-performance_liquid_chromatography	4	chromatography_hplc	0.7680
high-performance_liquid_chromatography	5	high-pressure_liquid_chromatography	0.7645
high-performance_liquid_chromatography	6	liquid_chromatography_hplc	0.7634
high-performance_liquid_chromatography	7	reversed-phase_hplc	0.7621
high-performance_liquid_chromatography	8	liquid_chromatography_mass_spectrometry	0.7565
high-performance_liquid_chromatography	9	high_performance_liquid	0.7502
high-performance_liquid_chromatography	10	rp-hplc	0.7500

Dataset for Acronyms and Abbreviations Finding

Dataset used for acronyms and abbreviation discovering application. This dataset contains a total of 40 examples of acronyms and abbreviations extracted from a chapter of a visualizing chemistry book. There's one restriction applied here while extracting the examples, which is all the acronyms, abbreviations and their corresponding definitions should be included in the vocabulary of $\mathbf{W}_p$.

Acronyms and Abbreviations	Definitions
aes	atomic_emission_spectroscopy
afm	atomic_force_microscopy
ccd	charge-coupled_device
ct	computed_tomography
ctab	cetyltrimethylammonium_bromide
dic	differential_interference_contrast
dwt	discrete_wavelet_transform
em	electron_microscopy
fcs	fluorescence_correlation_spectroscopy
ft	fourier_transform
ftir	fourier_transform_infrared_spectroscopy
gfp	green_fluorescent_protein
gui	graphical_user_interface
hplc	high-performance_liquid_chromatography
ir	infrared
ldl	low-density_lipoprotein
led	light-emitting_diode
md	molecular_dynamics
mpm	multiphoton_microscopy
mri	magnetic_resonance_imaging
ms	mass_spectroscopy
mva	multivariate_analysis
nir	near_infrared
nmr	nuclear_magnetic_resonance
pca	principal_component_analysis
pdb	protein_data_bank
pet	positron_emission_tomography
prp	prion_protein
rfp	red_fluorescent_protein
sem	scanning_electron_microscopy
shg	second_harmonic_generation
sims	secondary_ion_mass_spectrometry
spm	scanning_probe_microscopy

stem	scanning_transmission_electron_microscopy
stm	scanning_tunneling_microscopy
tem	transmission_electron_microscope
thg	third_harmonic_generation
thz	terahertz
uv	ultraviolet
wt	wavelet_transform