Investigating the behavior of published PAINS alerts using a pharmaceutical company dataset: Supporting Information

Lewis R. Vidler*, Ian A. Watson, Brandon J. Margolis, David J. Cummins, and Michael Brunavs

Contents

Table S1. Promiscuity data on 62 PAINS alerts matching compounds with data.

PAINS Alert	Actives	Data Points	% Active	Unique Compounds	Assays Active[a]	Assays Tested	Gene Families Active[b]	Gene Families Tested	PAINS Activity Enrichment[c]
dyes3A	2	3	66.7	1	2	3	1	2	**4.9***
anil_di_alk_A	29846	114572	26	5291	870	1847	14	14	**1.9***
naphth_amino_B	3	15	20	3	3	15	2	5	**1.5**
imidazole_A	1612	9895	16.3	413	135	653	13	14	**1.2***
het_pyridiniums_A	277	1752	15.8	126	128	483	8	13	**1.2***
azo_A	514	3387	15.2	718	166	659	10	13	**1.1***
catechol_A	628	4527	13.9	448	268	874	11	13	**1**
ene_rhod_A	1501	10999	13.6	1033	296	874	12	15	**1**
anil_alk_ene	549	4067	13.5	430	166	570	11	13	**1**
thio_carbonate_A	4	30	13.3	3	4	28	2	7	**1**
quinone_A	774	5818	13.3	590	267	753	10	13	**1**
ene_five_het_E	12	97	12.4	8	12	74	3	10	**0.9**
dhp_bis_amino_CN	8	67	11.9	3	5	42	2	7	**0.9**
cyano_pyridone_B	228	1990	11.5	91	143	518	9	14	**0.8**
thiophene_amino_Aa	406	3632	11.2	223	142	508	10	13	**0.8**
pyrrole_B	222	2153	10.3	179	46	326	8	12	**0.8**
cyano_pyridone_A	205	2041	10	105	73	365	8	13	**0.7**
sulfonamide_A	32	319	10	17	32	265	4	10	**0.7**
amino_acridine_A	177	1812	9.8	128	61	380	7	12	**0.7**
ene_five_het_B	106	1118	9.5	75	100	467	7	13	**0.7**
rhod_sat_A	88	948	9.3	56	39	237	6	12	**0.7**
ene_five_het_F	18	201	9	9	16	115	5	12	**0.7**
sulfonamide_B	84	1049	8	104	35	229	8	13	**0.6**
anil_no_alk	544	7278	7.5	511	245	875	12	15	**0.5**
thio_dibenzo	54	768	7	49	40	227	5	12	**0.5**
naphth_amino_A	161	2315	7	100	81	353	9	14	**0.5**
anthranil_one_A	314	4569	6.9	308	166	751	11	14	**0.5**
anil_di_alk_C	2106	31649	6.7	2167	332	1215	15	16	**0.5**
anil_di_alk_B	300	4537	6.6	324	195	633	13	15	**0.5**
ene_one_hal	71	1092	6.5	105	49	275	9	13	**0.5**
pyrrole_A	662	10722	6.2	666	261	830	12	13	**0.4**
anil_di_alk_E	972	16696	5.8	1210	262	1016	15	16	**0.4**
hzone_acyl_naphthol	4	69	5.8	15	4	54	2	9	**0.4**
imine_one_fives	9	160	5.6	29	8	59	5	11	**0.4**
anil_di_alk_D	110	2239	4.9	170	40	341	11	14	**0.4**
keto_keto_beta_A	212	4842	4.4	285	122	469	10	13	**0.3**
ene_one_ene_A	4	93	4.3	15	3	66	2	8	**0.3**
ene_cyano_A	10	237	4.2	11	9	141	6	11	**0.3**
indol_3yl_alk	1124	27840	4	1866	258	804	14	16	**0.3**
mannich_A	801	20449	3.9	1336	246	873	15	16	**0.3**
ene_five_het_A	71	1908	3.7	195	53	351	9	13	**0.3**
ene_five_one_A	36	968	3.7	69	26	253	8	14	**0.3**
cyano_imine_A	3	86	3.5	10	3	35	2	6	**0.3**

imine_one_A	15	440	3.4	71	10	164	5	12	**0.2**
cyano_ene_amine_A	1	32	3.1	4	1	23	1	3	**0.2**
thiophene_hydroxy	7	229	3.1	19	7	87	4	11	**0.2**
het_6_tetrazine	35	1223	2.9	42	18	185	6	13	**0.2**
thiophene_amino_Ab	72	2622	2.7	206	50	345	10	14	**0.2**
hzone_phenol_A	37	1354	2.7	144	25	210	6	11	**0.2**
hzone_enamin	22	832	2.6	93	16	183	5	12	**0.2**
hzone_phenol_B	19	806	2.4	99	15	148	6	10	**0.2**
ene_five_het_C	19	840	2.3	79	12	168	5	12	**0.2**
diazox_sulfon_A	19	855	2.2	61	16	205	5	12	**0.2**
ene_six_het_A	52	2353	2.2	200	29	271	8	13	**0.2**
anthranil_acid_A	1	49	2	5	1	35	1	8	**0.1**
het_65_A	9	463	1.9	23	7	172	5	12	**0.1**
thiaz_ene_A	96	5667	1.7	288	58	561	11	13	**0.1**
thiaz_ene_B	5	331	1.5	10	5	102	3	11	**0.1**
ene_one_ester	3	237	1.3	14	2	84	2	10	**0.1**
imine_one_isatin	11	1867	0.6	116	10	243	5	8	**0**
hzone_pipzn	1	422	0.2	54	1	38	1	4	**0**
hzone_anil_di_alk	0	27	0	3	0	27	0	3	**0**

[a] Assays for which at least one molecule matching the rule is found active. [b] Gene families for which at least one molecule matching the rule is found active. [c] A star indicates statistical significance (FDR<0.1)

Table S2. Stability data on 47 PAINS alerts matching compounds with data.

PAINS Alert	Measured Impure	Data Points	% Impure	Unique Compounds	PAINS Instability Enrichment[a]
dhp_bis_amino_CN	1	1	100.0	1	**19.0**
ene_five_het_C	1	1	100.0	1	**19.0**
ene_five_one_A	1	1	100.0	1	**19.0**
ene_one_ene_A	1	1	100.0	1	**19.0**
thio_dibenzo	1	1	100.0	1	**19.0**
pyrrole_B	8	9	88.9	9	**16.9***
keto_keto_beta_A	3	5	60.0	5	**11.4***
anil_alk_ene	10	19	52.6	19	**10.0***
ene_five_het_E	1	2	50.0	2	**9.5**
quinone_A	29	60	48.3	60	**9.2***
cyano_pyridone_A	5	13	38.5	13	**7.3***
ene_one_hal	4	11	36.4	11	**6.9**
azo_A	103	291	35.4	291	**6.7***
anil_di_alk_B	7	21	33.3	21	**6.3***
het_pyridiniums_A	2	6	33.3	6	**6.3***
pyrrole_A	3	10	30.0	10	**5.7***
ene_rhod_A	7	25	28.0	25	**5.3***
catechol_A	6	29	20.7	29	**3.9***
anil_di_alk_D	2	10	20.0	10	**3.8**
anil_di_alk_C	37	203	18.2	203	**3.5***
anthranil_one_A	5	30	16.7	30	**3.2**
hzone_phenol_A	1	6	16.7	6	**3.2**
thiaz_ene_A	1	6	16.7	6	**3.2**
thiophene_amino_Aa	4	24	16.7	24	**3.2***
mannich_A	15	143	10.5	143	**2.0***
indol_3yl_alk	9	155	5.8	155	**1.1**
anil_no_alk	2	46	4.3	46	**0.8**
anil_di_alk_A	42	1326	3.2	1326	**0.6**
anil_di_alk_E	11	377	2.9	377	**0.6**
thiophene_amino_Ab	1	91	1.1	91	**0.2**
cyano_imine_B	0	1	0.0	1	**0.0**
cyano_pyridone_B	0	14	0.0	14	**0.0**
diazox_sulfon_A	0	4	0.0	4	**0.0**
ene_cyano_A	0	3	0.0	3	**0.0**
ene_five_het_A	0	1	0.0	1	**0.0**
ene_five_het_B	0	2	0.0	2	**0.0**
ene_six_het_A	0	1	0.0	1	**0.0**
het_65_A	0	1	0.0	1	**0.0**
het_6_tetrazine	0	2	0.0	2	**0.0**
hzone_enamin	0	3	0.0	3	**0.0**
hzone_phenol_B	0	2	0.0	2	**0.0**
imine_one_A	0	1	0.0	1	**0.0**
naphth_amino_A	0	1	0.0	1	**0.0**

rhod_sat_A	0	2	0.0	2	**0.0**
sulfonamide_A	0	1	0.0	1	**0.0**
sulfonamide_B	0	3	0.0	3	**0.0**

[a]A star indicates a statistically significant increase (FDR<0.1)

Table S3. Cytotoxicity data on 53 PAINS alerts matching compounds with data.

PAINS Alert	Measured Cytotoxic	Data Points	% Cytotoxic	Unique Compounds	Assays Cytotoxic[a]	Assays Tested	% Assays Cytotoxic	PAINS Cytotoxicity enrichment[b]
quinone_A	352	574	61.3	81	217	246	88.2	**4.0***
amino_acridine_A	44	85	51.8	17	22	42	52.4	**3.4***
ene_cyano_A	13	28	46.4	2	13	28	46.4	**3.0**
anil_alk_ene	44	109	40.4	42	33	69	47.8	**2.6***
ene_one_ene_A	4	10	40.0	2	4	10	40.0	**2.6**
ene_five_het_B	3	8	37.5	5	3	6	50.0	**2.5**
anthranil_one_A	71	190	37.4	53	64	100	64.0	**2.5***
het_pyridiniums_A	237	646	36.7	64	42	58	72.4	**2.4***
catechol_A	279	779	35.8	93	57	144	39.6	**2.3***
ene_five_het_A	91	288	31.6	14	91	272	33.5	**2.1***
imidazole_A	70	227	30.8	87	19	44	43.2	**2.0***
ene_one_hal	11	39	28.2	12	10	26	38.5	**1.8**
anil_di_alk_A	1896	6939	27.3	1850	666	755	88.2	**1.8***
imine_one_isatin	2	8	25.0	1	2	8	25.0	**1.6**
hzone_phenol_A	7	30	23.3	13	6	18	33.3	**1.5**
ene_rhod_A	40	192	20.8	91	20	51	39.2	**1.4**
naphth_amino_A	9	44	20.5	25	9	22	40.9	**1.3**
mannich_A	319	1581	20.2	409	241	306	78.8	**1.3***
hzone_enamin	5	25	20.0	10	5	17	29.4	**1.3**
sulfonamide_B	22	113	19.5	36	8	35	22.9	**1.3**
azo_A	34	214	15.9	80	16	72	22.2	**1.0**
anil_no_alk	20	135	14.8	99	12	34	35.3	**1.0**
sulfonamide_A	18	141	12.8	1	18	141	12.8	**0.8**
thiophene_amino_Aa	7	55	12.7	34	5	20	25.0	**0.8**
cyano_imine_B	2	20	10.0	3	2	12	16.7	**0.7**
pyrrole_B	3	30	10.0	22	2	11	18.2	**0.7**
anil_di_alk_E	82	1028	8.0	303	31	84	36.9	**0.5**
keto_keto_beta_A	18	250	7.2	80	13	49	26.5	**0.5**
thio_dibenzo	1	16	6.3	8	1	11	9.1	**0.4**
ene_five_het_E	1	19	5.3	3	1	11	9.1	**0.3**
het_6_tetrazine	1	20	5.0	15	1	8	12.5	**0.3**
anil_di_alk_B	4	83	4.8	54	3	31	9.7	**0.3**
cyano_pyridone_A	1	21	4.8	13	1	6	16.7	**0.3**
pyrrole_A	26	548	4.7	115	21	74	28.4	**0.3**
anil_di_alk_C	92	2327	4.0	685	55	142	38.7	**0.3**
indol_3yl_alk	77	1971	3.9	473	38	99	38.4	**0.3**
cyano_pyridone_B	1	31	3.2	26	1	6	16.7	**0.2**
diazox_sulfon_A	1	32	3.1	15	1	20	5.0	**0.2**

thiaz_ene_A	3	131	2.3	71	3	38	7.9	**0.2**
thiophene_amino_Ab	2	342	0.6	54	2	22	9.1	**0.0**
anil_di_alk_D	0	61	0.0	36	0	24	0.0	**0.0**
anthranil_acid_A	0	1	0.0	1	0	1	0.0	**0.0**
cyano_ene_amine_A	0	1	0.0	1	0	1	0.0	**0.0**
ene_five_het_C	0	11	0.0	7	0	4	0.0	**0.0**
ene_five_het_F	0	5	0.0	4	0	2	0.0	**0.0**
ene_five_one_A	0	16	0.0	13	0	8	0.0	**0.0**
ene_one_ester	0	15	0.0	2	0	15	0.0	**0.0**
ene_six_het_A	0	14	0.0	11	0	3	0.0	**0.0**
het_65_A	0	6	0.0	5	0	4	0.0	**0.0**
hzone_acyl_naphthol	0	1	0.0	1	0	1	0.0	**0.0**
hzone_phenol_B	0	9	0.0	7	0	5	0.0	**0.0**
imine_one_A	0	6	0.0	5	0	5	0.0	**0.0**
imine_one_fives	0	6	0.0	2	0	5	0.0	**0.0**

[a] Assays for which at least one molecule matching the rule is found cytotoxic. [b] A star indicates statistical significance (FDR<0.1)

Table S4. Summary of all hill slope data included in analysis of PAINS alerts

Assay Format	High Hill Slope	Data Points	% High Hill Slope	Unique Compounds	Unique Assays	Unique Gene Families
AS	13,901	62,011	22.4	43,623	159	7
ELISA	16,177	56,304	28.7	38,073	154	6
FB	10,431	240,900	4.3	106,818	471	8
FP	16,956	117,026	14.4	56,478	116	6
FRET	23,612	752,537	3.3	103,445	601	10
SPA	18,449	219,108	8.4	95,649	344	11
Overall	**99,526**	**1,447,886**	**6.9**	**281,315**	**1,845**	**16**

Table S5. Summary of hill slope data for compounds matching the PAINS alerts

Assay Format	High Hill Slope	Data Points	% High Hill Slope	Unique Compounds	Unique Assays	Unique Gene Families	PAINS Hill Slope enrichment (odds ratio)
AS	476	2,682	17.7	1,789	106	7	**0.8 (0.72)**
ELISA	268	818	32.8	569	65	4	**1.1 (1.2)**
FB	279	5,723	4.9	2,187	249	8	**1.1 (1.1)**
FP	684	4,022	17.0	2,024	94	5	**1.2 (1.2)**
FRET	715	27,014	2.6	2,569	393	9	**0.8 (0.9)**
SPA	510	4,270	11.9	2,373	182	9	**1.4 (1.4)**
Overall	**2,858**	**44,332**	**6.4**	**6877**	**1080**	**15**	**0.9 (1.0)**

Table S6. Hill slope data on 61 PAINS alerts matching compounds with data.

PAINS Alert	High Hill Slope	Data Points	% High Hill Slope	Unique Compounds	Assays With High Hill Slope[a]	Assays Tested	Gene Families With High Hill Slope[b]	Gene Families Tested	PAINS Hill Slope Enrichment[c]
cyano_imine_A	1	1	100	1	1	1	1	1	**15.1**
cyano_ene_amine_A	1	2	50	1	1	2	1	1	**7.6**
dyes3A	1	2	50	1	1	2	1	1	**7.6**
thiaz_ene_B	3	6	50	3	3	6	2	4	**7.6***
thio_carbonate_A	1	2	50	2	1	2	1	1	**7.6**
thiophene_hydroxy	1	2	50	2	1	2	1	2	**7.6**
het_65_A	3	9	33.3	8	2	7	2	5	**5**
ene_five_one_A	9	28	32.1	11	9	26	3	6	**4.9***
het_6_tetrazine	6	21	28.6	14	4	12	3	4	**4.3***
cyano_pyridone_A	44	157	28	62	24	70	6	7	**4.2***
ene_one_hal	17	61	27.9	27	16	44	5	9	**4.2***
hzone_acyl_naphthol	1	4	25	1	1	4	1	2	**3.8**
imine_one_A	3	12	25	6	2	6	2	3	**3.8**
rhod_sat_A	18	75	24	32	15	39	4	6	**3.6***
thiophene_amino_Aa	66	286	23.1	107	37	108	7	9	**3.5***
amino_acridine_A	27	118	22.9	67	15	39	6	7	**3.5***
thio_dibenzo	8	37	21.6	13	7	28	3	5	**3.3***
ene_five_het_F	3	14	21.4	6	3	14	2	5	**3.2***
diazox_sulfon_A	3	15	20	6	3	14	2	6	**3**
anil_alk_ene	91	460	19.8	149	42	187	7	9	**3***
naphth_amino_A	21	107	19.6	38	20	56	6	8	**3***
thiophene_amino_Ab	13	68	19.1	28	13	45	6	9	**2.9***
quinone_A	109	639	17.1	194	53	204	8	10	**2.6***
dhp_bis_amino_CN	1	6	16.7	3	1	5	1	2	**2.5**
ene_five_het_C	3	18	16.7	14	3	11	2	5	**2.5**
indol_3yl_alk	145	871	16.6	355	87	211	11	13	**2.5***
anil_di_alk_D	17	105	16.2	53	9	32	3	9	**2.4***
mannich_A	105	651	16.1	326	65	234	10	13	**2.4***
het_pyridiniums_A	39	252	15.5	58	17	154	6	8	**2.3***
ene_rhod_A	205	1345	15.2	416	63	285	8	11	**2.3***
sulfonamide_B	12	80	15	34	12	39	4	6	**2.3**
keto_keto_beta_A	23	156	14.7	52	19	79	6	9	**2.2***
ene_six_het_A	6	42	14.3	34	5	24	3	8	**2.2**
pyrrole_A	81	575	14.1	158	46	279	5	12	**2.1***
anil_di_alk_C	283	2053	13.8	604	74	374	9	14	**2.1***
azo_A	60	443	13.5	189	29	166	7	10	**2***
ene_cyano_A	1	8	12.5	4	1	7	1	5	**1.9**
pyrrole_B	25	212	11.8	89	12	56	6	7	**1.8***
anil_no_alk	51	459	11.1	117	36	271	8	11	**1.7***
anthranil_one_A	33	301	11	80	31	204	7	10	**1.7**
imidazole_A	48	446	10.8	145	24	165	9	11	**1.6***

ene_five_het_A	7	74	9.5	28	6	36	5	7	**1.4**
anil_di_alk_E	68	842	8.1	411	48	311	10	15	**1.2**
anil_di_alk_B	19	256	7.4	74	16	164	7	11	**1.1**
thiaz_ene_A	12	171	7	43	10	150	9	12	**1.1**
catechol_A	40	699	5.7	161	32	290	10	10	**0.9**
cyano_pyridone_B	14	257	5.4	29	13	142	4	9	**0.8**
anil_di_alk_A	1215	32029	3.8	2662	297	791	12	14	**0.6**
hzone_phenol_A	3	84	3.6	19	3	29	3	6	**0.5**
imine_one_isatin	1	30	3.3	9	1	14	1	5	**0.5**
ene_five_het_B	5	157	3.2	16	5	152	2	8	**0.5**
hzone_enamin	1	51	2	9	1	21	1	3	**0.3**
sulfonamide_A	2	105	1.9	3	2	105	2	5	**0.3**
ene_five_het_E	0	9	0	2	0	9	0	2	**0**
ene_one_ene_A	0	4	0	3	0	3	0	2	**0**
ene_one_ester	0	2	0	1	0	1	0	1	**0**
hzone_anil_di_alk	0	8	0	1	0	8	0	1	**0**
hzone_phenol_B	0	34	0	14	0	31	0	7	**0**
hzone_pipzn	0	1	0	1	0	1	0	1	**0**

[a]Assays for which at least one molecule matching the rule is found to have a high hill slope. [b]Gene families for which at least one molecule matching the rule is found to have a high hill slope. [c]A star indicates statistical significance (FDR<0.1)

Table S7. Summary of findings across all issues by alert

Enrichments for each factor considered for each alert are shown, calculated from the raw data. Those alerts showing enrichment in the raw data and statistically significant changes (FDR<0.1) once the data has been normalized by target and stats corrected for multiple hypothesis testing are highlighted in green (as described below). Totals per rule and per factor are only those showing statistically significant increases. Blank cells indicate no data available.

PAINS Alert	Overall	AS	ELISA	FB	FP	FRET	SPA	Overall HS	AS HS	ELISA HS	FB HS	FP HS	FRET HS	SPA HS	QC	Cytotox	Total demerits
anil_di_alk_A	1.9	3.4	1.6	2	4.1	3.1	1.3	0.6	0.6	1.5	1.1	0.8	0.4	0.8	0.6	1.8	9
het_pyridiniums_A	1.2	1.3	0.6	0.6	0.7	1.5	4.2	2.3	3.2	2.3	4.7	2.3	1.8	1.9	6.3	2.4	5
azo_A	1.1	4.5	1.9	0.9	1.8	1.8	1.9	2	0.7	1.1	0.3	2.7	1.8	0.8	6.7	1	5
catechol_A	1	2.9	0.7	1.2	2	2	2	0.9	1	0.9	0.8	1	0.8	1.3	3.9	2.3	2
ene_rhod_A	1	3.2	1.3	0.6	1.8	1.5	1.6	2.3	1.2	0.2	2.4	2.4	1.5	2.6	5.3	1.4	9
anil_alk_ene	1	3	0.4	0.4	2.6	2.1	0.9	3	0.4	0	0	2.7	3.9	2	10	2.6	8
quinone_A	1	4.4	1.8	1	2.1	0.8	2.2	2.6	0.8	1.1	2.1	1.6	2.7	2.9	9.2	4	7
ene_five_het_E	0.9	3.3	0	0	0	2.3	1	0	0				0	0	9.5	0.3	1
cyano_pyridone_B	0.8	0.3	0	0.3	0.2	1.9	0.7	0.8	4.5		0	0	1.1	2.2	0	0.2	0
thiophene_amino_Aa	0.8	3.3	0.5	0.7	1.7	0.7	1.6	3.5	1.6	0.5	4.9	2.2	5.9	2.1	3.2	0.8	9
pyrrole_B	0.8	7.9	0.4	0.1	0.3	0.6	0.3	1.8	0.2	3.5	3.9	1.7	9.3	0.9	16.9	0.7	4
cyano_pyridone_A	0.7	2.5	0.8	0.1	1	0.7	1.6	4.2	2.6	1.2	0	0.6	11.7	2.2	7.3	0.3	6
sulfonamide_A	0.7	0	2.5	0.1	1.4	2.4	0	0.3	4.5	0	0	6.9	0	0	0	0.8	0
ene_five_het_B	0.7	1.9	1.5	0.4	0.6	1.9	0.6	0.5	0.9	0.9	0	0	0	5.1	0	2.5	2
rhod_sat_A	0.7	2.4	1.7	0.1	1.2	0.6	1	3.6	2.7	0.2	0	1.9	10.7	2.5	0	0	3
sulfonamide_B	0.6	5.2	0.2	0.6	1	0.2	0.3	2.3	0.2	3.5	0	1.9	21.6	1.7	0	1.3	2
anil_no_alk	0.5	1.7	0.5	1.2	0.6	0.4	0.7	1.7	0.7	1.1	1.2	1.4	2.7	2.3	0.8	1	2
thio_dibenzo	0.5	1.8	1.8	0.2	1.2	0.5	0.6	3.3	1.8	0	0	1.2	7.6	4.8	19	0.4	1
naphth_amino_A	0.5	1.9	2	0.2	1.7	0.4	1	3	1.7	1.2	0	1.4	5.4	1	0	1.3	4
anthranil_one_A	0.5	1.5	0.8	0.6	1.6	0.4	0.8	1.7	0.9	0.6	2.3	1.3	2.2	0.6	3.2	2.4	2
anil_di_alk_C	0.5	1.8	0.3	0.3	0.4	0.3	1.4	2.1	1.5	0.6	1.3	2.5	2.3	1	3.5	0.3	7
anil_di_alk_B	0.5	1.7	0.2	0.3	0.5	1.1	0.5	1.1	0.5	0.7	1.5	1.1	0.7	2.2	6.3	0.3	2
ene_one_hal	0.5	1.4	2	0.3	0.7	0.4	0.9	4.2	0.6	0.3	4.7	2.6	15.2	4	6.9	1.8	2
pyrrole_A	0.4	1.2	0.4	0.4	0.9	0.7	0.6	2.1	0.6	0.4	0.5	2.1	3.8	2.1	5.7	0.3	2
anil_di_alk_E	0.4	1.9	0.4	0.8	0.1	0.3	0.7	1.2	0.2	0.3	0.9	2.6	2.6	1.7	0.6	0.5	1

anil_di_alk_D	0.4	3.1	0.5	0.2	0.1	0.2	0.5	2.4	1.2	0	0	0	3.4	0.8	3.8	0	2
keto_keto_beta_A	0.3	1	0.6	0.7	0.5	0.3	0.3	2.2	1.3	0	0	2.1	4.7	1.6	11.4	0.5	2
ene_cyano_A	0.3	4.2	1.7	0.2	0.5	0.5	1.3	1.9	4.5	0	0	0	0	0	0	3	1
indol_3yl_alk	0.3	0.9	0.2	0.4	0.4	0.4	0.5	2.5	0.8	1	1.3	2.1	5.6	1.9	1.1	0.3	4
mannich_A	0.3	0.9	0.3	0.5	0.2	0.4	0.4	2.4	1.1	0.8	0.7	2.7	7.7	1	2	1.3	4
ene_five_het_A	0.3	2.6	0	0.1	0.1	1.3	0.5	1.4	0.6		0.6	2.3	5.5	1.7	0	2.1	1
ene_five_one_A	0.3	1.9	0	0.1	0.7	0.2	0.4	4.9	0		0	2.3	15.2	3	19	0	2
imine_one_A	0.2	3.5	0	0.4	0	0.4	0.2	3.8	0		10		0		0	0	0
het_6_tetrazine	0.2	1.4	0.7	0.1	0	0.2	0.3	4.3	0	0.9	23.3		4.3	7.1	0	0.3	2
thiophene_amino_Ab	0.2	0.3	0.4	0.2	0.3	0.3	0.3	2.9	1.5		0	1.1	6.5	3.2	0.2	0	2
hzone_phenol_A	0.2	1	1.4	0.3	0.2	0.2	0.3	0.5	0	1.7	0	2.3	0	4	3.2	1.5	0
hzone_enamin	0.2	1.5	0.9	0.3	0.1	0.1	0.1	0.3	0	1.7	0			0	0	1.3	0
hzone_phenol_B	0.2	1.9	0	0.2	0.2	0.6	0.1	0	0		0	0	0	0	0	0	0
ene_five_het_C	0.2	2.6	1	0.1	0.3	0	0.4	2.5	0.6	3.5	0	3.5		0	19	0	0
diazox_sulfon_A	0.2	1.1	0.8	0.1	0.1	0.2	0.2	3	1.3	0	0	3.5	0	0	0	0.2	0
ene_six_het_A	0.2	1	0	0.1	0.9	0.3	0.9	2.2	1.1		0	2.6	10.1	0	0	0	1
het_65_A	0.1	0.5	0.8	0.1	0	0	0.6	5	0		23.3	0		4	0	0	0
thiaz_ene_A	0.1	0.6	0.1	0.1	0.1	0.1	0.2	1.1	0.3	0	1.2	4.2	1.4	1.1	3.2	0.2	0
thiaz_ene_B	0.1	0.3	0.4	0	0.2	0.1	0	7.6	4.5	0		3.5	0		0	0	1
ene_one_ene_A	0.3	2.6		0	0	2.3	0	0	0				0		19	2.6	0
imidazole_A	1.2	3.5	3.7	2.1	0.5	0.8	0.3	1.6	0.8	0.2	1.5	3.8	1.6	1.5		2	7
amino_acridine_A	0.7	0.8	0.8	0.5	1.2	1.1	4.1	3.5	2.7	3.5	1.9	6.3	5.1	1.6		3.4	4
ene_five_het_F	0.7	4.3	0	0.8	1.1	0.3	0.7	3.2	0		11.6	0	15.2	6		0	1
hzone_acyl_naphthol	0.4	8.9	0	0	0	0	1.2	3.8	0					11.9		0	0
imine_one_fives	0.4	6.9	0	0	0	0	2	0	0					0		0	0
thiophene_hydroxy	0.2	0.9	0	0.7	0	0.3	0.3	7.6					15.2			0	0
ene_one_ester	0.1	0	0.6	0	0.3	0.1	0	0				0	0			0	0
cyano_ene_amine_A	0.2		0	0	0	0	1.8	7.6						6		0	0
imine_one_isatin	0	0		0.1	0	0.5	0.4	0.5			0		0	11.9		1.6	0
anthranil_acid_A	0.1		0	0	0	0	1.4									0	0
dhp_bis_amino_CN	0.9	3		1.4	5.2	0	1	2.5	0		0	3.5		0	19		0
thio_carbonate_A	1	0	0	1	5.2	0	7	7.6			0	6.9					0
cyano_imine_A	0.3			0.2	0	0	1.6	15.1			23.3						0

hzone_pipzn	0			0	0	3.1	0	0					0				0
hzone_anil_di_alk	0	0		0	0		0	0			0						0
dyes3A	4.9						4.7	7.6						6			1
naphth_amino_B	1.5	8.3	0	0	13	0		0				0					0
cyano_imine_B															0	0.7	0
Totals	**5**	**11**	**5**	**2**	**8**	**6**	**9**	**27**	**2**	**1**	**1**	**10**	**15**	**5**	**13**	**11**	

PAINS SMARTS Queries discussion

A challenge we faced with utilizing the queries as described in the BH2010 paper was that we found that the aromaticity definitions implemented in SLN/SYBYL were frequently inconsistent with what we observed in other software, nor with our experience and expectations (e.g. 5-membered heterocycles not being classed as aromatic). Since the SYBYL implementation seems to take a very restrictive view of aromaticity, when the queries were directly implemented outside SYBYL, we frequently found many more matches for the queries requiring aromatic atoms and fewer when double bonds were required.

Having translated the alerts, we ran both the SLN queries, and the new implementation against the Lilly collection (~2 million molecules) and manually inspected differences in the matches from both. We iteratively modified the alerts to try to maximally match what we believed was originally intended. If a query depended on aromaticity definitions that differed between implementations, we decided whether to make the query behave more like the original implementation, or more in line with our own expectations of aromaticity. For those alerts not strongly dependent on aromaticity perception, we were able to make very direct translations from the SLN representation, and found excellent concordance between our implementation and what was observed with SLN.

Statistical assessments of the different datasets

Treatment of activity and hill slope data

We wanted to normalize the data by target, to achieve this, for each assay format we reduced the data to compound-target pairs (active or inactive). If one compound had been tested in multiple assays against the same target and had always been active or inactive this value was assigned to the compound-target pair. If compounds had been tested in multiple assays against the same target and found both active and inactive these compounds were excluded (this could happen if different assays had a different top concentration for example and happened less than 1% of the time). If a target had no PAINS matching compounds present it was excluded. The percentage of molecules containing PAINS for each target was then calculated, only those between the 5th and 95th centiles here were kept. A similar approach was taken when looking at the hit rates for each target. Those targets with hit rates between the 5th and 95th centiles here were kept.

In addition to the raw enrichment values, we have calculated odds ratios which is the ratio of the odds of being active given the presence of a PAINS alert compared to the odds of being active in the absence of a PAINS alert. This was done on the normalized set and provides a complementary value to the enrichment values calculated on the raw data.

Assessment of statistical significance

With hundreds to thousands of comparisons, using a p-value alone is not stringent enough. The power to detect real effects is great, but the number of false positives (statistically significant effect that turns out to be due to chance alone and not real) is unacceptably high.

Bonferroni adjusted p-values are a way to account for multiple hypothesis testing, however, it freqently gives results that are too stringent and can result in missing real effects with this method. In Statistical terms, the concept of "power" refers to the probability of capturing a real effect. With the Bonferroni approach, our power to capture real effects goes down.[1]

False Discovery Rate (or FDR) strikes a good middle ground of high power and low incidence of false discoveries. Developed by Benjamini and Hochberg, and this is the method we have used here.

Improvements to the FDR have been developed since B&H's original development. B&H method assumed the null prevalence rate was 100%. By this we mean that the method assumes that every alert has no effect on hit rate (or hill slope, etc), and then challenges the p-value distribution to deny that starting assumption. Recent advances by Brad Efron,[1] and independently by John Storey,[3] attempt to estimate the prevalence rate of the alternative hypothesis by using a smoothed distribution estimate of the density of the p-values input stream. This results in a slight improvement in power.

We used John Storey's Q value implemented in the R package, "qvalue," to compute the local FDR.

Experimental: LCMS purity method

QC was performed using a tiered set of LCMS conditions employing reversed-phase Water/Acetonitrile ballistic gradients in either low or high pH on C18 or polar C18 columns. Purity was determined by UV area percent with MS-based peak identification.

References

(1) Efron, B. Size, Power and False Discovery Rates *The Annals of Statistics* **2007**, 35, 4, 1351-1377

(2) Benjamini, Y.; Hochberg, Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing *J R Stat Soc Series B Stat Methodol* **1995**, 57, 1, 289-300

(3) Storey, J. D. A direct approach to false discovery rates *J. R. Statist. Soc. B* **2001**, 64, Part 3, 479-498