Electronic Supplementary Material (ESI) for Digital Discovery.

Supporting Information - Chemical Space Analysis and Property Prediction for Carbon Capture Amine molecules

James L. McDonagh, Stamatia Zavitsanou, Alexander Harrison, Dimitry Zubarev, Theordore van Kessel, Benjamin H. Wunsch and Flaviu Cipcigan
flaviu.cipcigan@ibm.com, james.mcdonagh@serna.bio

December 29, 2023

1 Unit Conversions

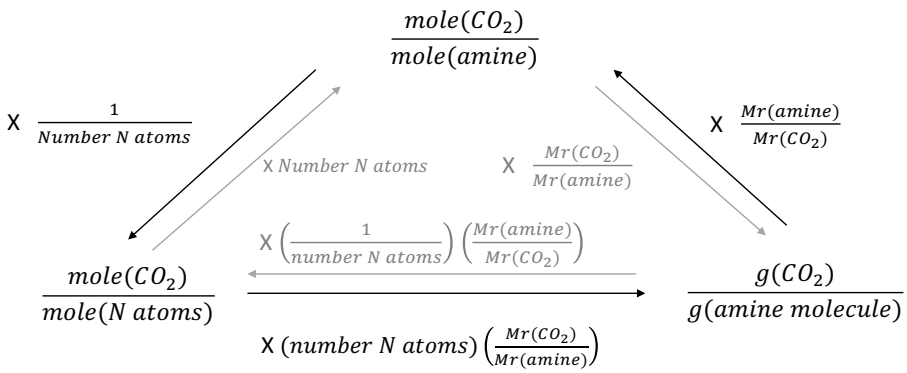


Figure 1: The conversion factors used in this work to move between common capacity measurement units.

Where units were given that need knowledge of the solution density for converting, such as $\frac{g(\mathrm{CO_2})}{L(\mathrm{solution})}$, the following was used assuming the density (were not given) was that of water, and therefore has a density of 1 g/ml.

$$\text{capacity} \times \frac{1}{\text{amine weight fraction} \times 1000} \times \frac{\text{Mr(amine)}}{\text{Mr}(\text{CO}_2)} = \frac{mol(\text{CO}_2)}{mol(\text{amine})} \tag{1}$$

2 Data set

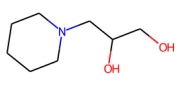
Table 1: Experimental data and data set for 98 amine molecules tested for carbon capture

Index	IUPAC Name	InChIKey	Absorption capacity $\left(\frac{\text{moles}_{CO_2}}{\text{moles}_{amine}}\right)$	Observed initial rate $\left(\frac{\text{moles}_{CO_2}}{\text{moles}_{amine}.\text{sec}}\right)$
0	3-piperidin-1-ylpropane-1,2-diol	MECNWXGGNCJFQJ-UHFFFAOYSA-N	1.32	0.00
1	piperidin-2-ylmethanol	PRAYXGYYVXRDDW-UHFFFAOYSA-N	1.30	0.10
2	1-azabicyclo[2.2.2]octan-3-ol	IVLICPVPXWEGCA-UHFFFAOYSA-N	1.24	0.00
3	2-(diethylamino)ethanol	BFSVOASYOCHEOV-UHFFFAOYSA-N	1.14	0.09
4	2-(tert-butylamino)ethanol	IUXYVKZUDNLISR-UHFFFAOYSA-N	1.09	0.08
5	2-piperidin-2-ylethanol	PTHDBHDZSMGHKF-UHFFFAOYSA-N	1.06	0.14
6	2-(dimethylamino)-2-methylpropan-1-ol	XRIBIDPMFSLGFS-UHFFFAOYSA-N	1.03	0.00
7	2-pyrrolidin-1-ylethanol	XBRDBODLCHKXHI-UHFFFAOYSA-N	1.02	0.00
8	2-(dimethylamino)ethanol	UEEJHVSXFDXPFK-UHFFFAOYSA-N	0.93	0.11
9	piperidine	NQRYJNQNLNOLGT-UHFFFAOYSA-N	0.93	0.08
10	2-amino-2-methylpropan-1-ol	CBTVGIZVANVGBH-UHFFFAOYSA-N	0.90	0.08
11	2-aminocyclohexan-1-ol	PQMCFTMVQORYJC-UHFFFAOYSA-N	0.74	0.35
12	2-(ethylamino)ethanol	MIJDSYMOBYNHOT-UHFFFAOYSA-N	0.71	0.22
13	(2R,3R,4R,5S)-6-(methylamino)hexane-1,2,3,4,5-pentol	MBBZMMPHUWSWHV-BDVNFPICSA-N	0.71	0.20
14	(2R)-2-aminopropan-1-ol	BKMMTJMQCTUHRP-GSVOUGTGSA-N	0.67	0.10
15	N,N,N',N'-tetramethylpropane-1,3-diamine	DMQSHEKGGUOYJS-UHFFFAOYSA-N	0.66	0.00
16	2-aminopropane-1,3-diol	KJJPLEZQSCZCKE-UHFFFAOYSA-N	0.65	0.10
17	2-(2-hydroxyethylamino)ethanol	ZBCBWPMODOFKDW-UHFFFAOYSA-N	0.65	0.11
18	N'-[3-(dimethylamino)propyl]propane-1,3-diamine	OMKZWUPRGQMQJC-UHFFFAOYSA-N	0.62	0.20
19	N',N'-dimethylpropane-1,3-diamine	IUNMPGNGSSIWFP-UHFFFAOYSA-N	0.61	0.15
20	2-[2-(dimethylamino)ethoxy]ethanol	YSAANLSYLSUVHB-UHFFFAOYSA-N	0.61	0.00
21	pentan-2-amine	IGEIPFLJVCPEKU-UHFFFAOYSA-N	0.59	0.14
22	N,N',N'-trimethylpropane-1,3-diamine	SORARJZLMNRBAQ-UHFFFAOYSA-N	0.58	0.14
23	4-aminobutan-1-ol	BLFRQYKZFKYQLO-UHFFFAOYSA-N	0.56	0.20
24	(1S,2R)-cyclohexane-1,2-diamine	SSJXIUAHEKJCMH-OLQVQODUSA-N	0.56	1.69
25	2-aminoethanol	HZAXFHJVJLSVMW-UHFFFAOYSA-N	0.55	0.13
26	morpholine	YNAVUWVOSKDBBP-UHFFFAOYSA-N	0.55	0.09
27	N'-(3-aminopropyl)propane-1,3-diamine	OTBHHUPVCYLGQO-UHFFFAOYSA-N	0.54	0.38
28	1,3-diaminopropan-2-ol	UYBWIEGTWASWSR-UHFFFAOYSA-N	0.54	0.14
29	2-amino-2-methylpropane-1,3-diol	UXFQFBNBSPQBJW-UHFFFAOYSA-N	0.54	0.00
30	N,N'-dimethylethane-1,2-diamine	KVKFRMCSXWQSNT-UHFFFAOYSA-N	0.53	0.14
31	N',N'-diethylethane-1,2-diamine	UDGSVBYJWHOHNN-UHFFFAOYSA-N	0.52	0.30
32	2-amino-3-methylbutan-1-ol	NWYYWIJOWOLJNR-UHFFFAOYSA-N	0.52	0.16
33	piperidin-2-ylmethanamine	RHPBLLCTOLJFPH-UHFFFAOYSA-N	0.52	0.42
34	azetidine	HONIICLYMWZJFZ-UHFFFAOYSA-N	0.51	0.18
35	1-aminobutan-2-ol	KODLUXHSIZOKTG-UHFFFAOYSA-N	0.51	0.24
36	N,N-diethylpiperidin-4-amine	OUXRMEUJNPVXMM-UHFFFAOYSA-N	0.51	0.15
37	butan-1-amine	HQABUPZFAYXKJW-UHFFFAOYSA-N	0.51	0.18
38	pentan-1-amine	DPBLXKKOBLCELK-UHFFFAOYSA-N	0.51	0.29
39	(2R)-3-aminopropane-1,2-diol	KQIGMPWTAHJUMN-GSVOUGTGSA-N	0.50	0.08
40	N',N'-diethylpropane-1,3-diamine	QOHMWDJIBGVPIF-UHFFFAOYSA-N	0.49	0.15
41	(2R)-2-amino-4-methylpentan-1-ol	VPSSPAXIFBTOHY-ZCFIWIBFSA-N	0.48	0.19
42	N,N'-dimethylpropane-1,3-diamine	UQUPIHHYKUEXQD-UHFFFAOYSA-N	0.48	0.34
43	1-methyl-1,4-diazepane	FXHRAKUEZPSMLJ-UHFFFAOYSA-N	0.48	0.26
44	2-(methylamino)ethanol	OPKOKAMJFNKNAS-UHFFFAOYSA-N	0.47	0.08
45	piperazine	GLUUGHFHXGJENI-UHFFFAOYSA-N	0.47	0.18
46	pyrrolidin-3-amine	NGXSWUFDCSEIOO-UHFFFAOYSA-N	0.45	0.21
47	8-methyl-8-azabicyclo[3.2.1]octan-3-amine	HJGMRAKQWLKWMH-UHFFFAOYSA-N	0.44	0.25
48	N,N-dimethylpyrrolidin-3-amine	AVAWMINJNRAQFS-UHFFFAOYSA-N	0.43	0.16
49	3-[4-(3-aminopropyl)piperazin-1-yl]propan-1-amine	XUSNPFGLKGCWGN-UHFFFAOYSA-N	0.42	0.21
50	N,N-dimethylpiperidin-4-amine	YFJAIURZMRJPDB-UHFFFAOYSA-N	0.41	0.14
51	1-(2-hydroxypropylamino)propan-2-ol	LVTYICIALWPMFW-UHFFFAOYSA-N	0.41	0.44
52	N'-[2-[2-(2-aminoethylamino)ethylamino]ethyl]ethane-1,2-diamine	FAGUFWYHJQFNRV-UHFFFAOYSA-N	0.39	0.00
53	N'-(2-aminoethyl)ethane-1,2-diamine	RPNUMPOLZDHAAY-UHFFFAOYSA-N	0.38	0.12
54	(1R,2R)-cyclohexane-1,2-diamine	SSJXIUAHEKJCMH-PHDIDXHHSA-N	0.38	1.20
55	2-(2-aminoethoxy)ethanamine	GXVUZYLYWKWJIM-UHFFFAOYSA-N	0.37	0.13
56	2-aminobutan-1-ol	JCBPETKZIGVZRE-UHFFFAOYSA-N	0.35	0.27
57	1-[2-[bis(2-hydroxypropyl)amino]ethyl-(2-hydroxypropyl)amino]propan-2-ol	NSOXQYCFHDMMGV-UHFFFAOYSA-N	0.35	0.00
58	3-piperidin-1-ylpropan-1-amine	JMUCXULQKPWSTJ-UHFFFAOYSA-N	0.35	0.13
59	2-(2-aminoethylamino)ethanol	LHIJANUOQQMGNT-UHFFFAOYSA-N	0.32	0.15
60	(2R)-1-aminopropan-2-ol	HXKKHQJGJAFBHI-GSVOUGTGSA-N	0.32	0.24
61	2,3,4,6,7,8-hexahydropyrrolo[1,2-a]pyrimidine	SGUVLZREKBPKCE-UHFFFAOYSA-N	0.32	0.86
62	3-aminocyclohexan-1-ol	NIQIPYGXPZUDDP-UHFFFAOYSA-N	0.31	0.10
63	2,3,4,6,7,8,9,10-octahydropyrimido[1,2-a]azepine	GQHTUMJGOHRCHB-UHFFFAOYSA-N	0.30	0.41
64	3,4,6,7,8,9-hexahydro-2H-pyrimido[1,2-a]pyrimidine	FVKFHMNJTHKMRX-UHFFFAOYSA-N	0.30	0.35
65	[(2S)-pyrrolidin-2-yl]methanol	HVVNJUAVDAZWCB-YFKPBYRVSA-N	0.30	0.24
66	(2S)-1-aminopropan-2-ol	HXKKHQJGJAFBHI-VKHMYHEASA-N	0.28	0.23
67	N'-[2-[2-[2-(2-aminoethylamino)ethyl-[2-[bis(2-aminoethyl)amino]ethyl]amino]ethyl-[2-[2-[bis(2-aminoethyl)amino]ethylamino]ethyl]amino]ethyl]ethane-1,2-diamine	NA	0.28	0.15

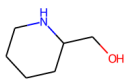
Continued on next page

Table 1: Experimental data and data set for 98 amine molecules tested for carbon capture

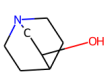
Index	IUPAC Name	InChIKey	Absorption capacity $\left(\frac{\text{moles}_{CO_2}}{\text{moles}_{amine}}\right)$	Observed initial rate $\left(\frac{\text{moles}_{CO_2}}{\text{moles}_{amine}.\text{sec}}\right)$
68	piperidin-3-ol	BIWOSRSKDCZIFM-UHFFFAOYSA-N	0.28	0.29
69	2-(aminomethyl)aniline	GVOYKJPMUUJXBS-UHFFFAOYSA-N	0.28	0.17
70	2-pyridin-4-ylethanamine	IDLHTECVNDEOIY-UHFFFAOYSA-N	0.26	0.27
71	1-amino-2-methylpropan-2-ol	LXQMHOKEXZETKB-UHFFFAOYSA-N	0.26	0.17
72	2-[2-hydroxyethyl(methyl)amino]ethanol	CRVGTESFCCXCTH-UHFFFAOYSA-N	0.24	0.00
73	1-[bis(2-hydroxyethyl)amino]propan-2-ol	ZFECCYLNALETDE-UHFFFAOYSA-N	0.22	0.00
74	2-[bis(2-hydroxyethyl)amino]ethanol	GSEJCLTVZPLZKY-UHFFFAOYSA-N	0.20	0.05
75	pyrazin-2-ylmethanamine	HQIBSDCOMQYSPF-UHFFFAOYSA-N	0.19	0.09
76	2-(1H-imidazol-5-yl)ethanamine	NTYJJOPFIAHURM-UHFFFAOYSA-N	0.17	0.24
77	2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanol	VARKIGWTYBUWNT-UHFFFAOYSA-N	0.11	0.03
78	(2S)-2-amino-3-methyl-3-sulfanylbutanoic acid	VVNCNSJFMMFHPL-VKHMYHEASA-N	0.09	0.00
79	piperidin-1-amine	LWMPFIOTEAXAGV-UHFFFAOYSA-N	0.07	0.09
80	1H-imidazol-5-ylmethanol	QDYTUZCWBJRHKK-UHFFFAOYSA-N	0.07	0.00
81	2-morpholin-4-ylethanol	KKFDCBRMNNSAAW-UHFFFAOYSA-N	0.07	0.04
82	2-[3-[[1,3-dihydroxy-2-(hydroxymethyl)propan-2-yl]amino]propylamino]-2-(hydroxymethyl)propane-1,3-diol	HHKZCCWKTZRCCL-UHFFFAOYSA-N	0.06	0.15
83	N-(2-hydroxyethyl)acetamide	PVCJKHHOXFKFRP-UHFFFAOYSA-N	0.05	0.23
84	1-pyridin-3-yl-N-(pyridin-3-ylmethyl)methanamine	FEBQXMFOLRVSGC-UHFFFAOYSA-N	0.05	0.17
85	tert-butyl N-(2-hydroxyethyl)carbamate	GPTXCAZYUMDUMN-UHFFFAOYSA-N	0.04	0.27
86	(2S)-2,5-diamino-5-oxopentanoic acid	ZDXPYRJPNDTMRX-VKHMYHEASA-N	0.03	0.00
87	pyridazin-3-amine	LETVJWLLIMJADE-UHFFFAOYSA-N	0.02	0.38
88	propanediamide	WRIRWRKPLXCTFD-UHFFFAOYSA-N	0.02	0.17
89	N-(2-sulfanylethyl)acetamide	AXFZADXWLMXITO-UHFFFAOYSA-N	0.02	0.00
90	1,3-oxazolidin-2-one	IZXIZTKNFFYFOF-UHFFFAOYSA-N	0.02	0.07
91	N,N-diethylhydroxylamine	FVCOIAYSJZGECG-UHFFFAOYSA-N	0.01	0.00
92	1,4-diazabicyclo[2.2.2]octane	IMNIMPAHZVJRPE-UHFFFAOYSA-N	0.01	0.00
93	pyridin-2-amine	ICSNLGPSRYBMBD-UHFFFAOYSA-N	0.01	0.08
94	1,3,5,7-tetrazatricyclo[3.3.1.13,7]decane	VKYKSIONXSXAKP-UHFFFAOYSA-N	0.01	0.08
95	2-aminoethyl dihydrogen phosphate	SUHOOTKUPISOBE-UHFFFAOYSA-N	0.01	0.05
96	pyrrol-1-amine	YNZAFFFENDLJQG-UHFFFAOYSA-N	0.01	0.00
97	pyrazin-2-amine	XFTQRUTUGRCSGO-UHFFFAOYSA-N	0.01	0.00



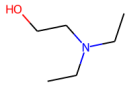
1) 3-piperidin-1-ylpropane-1,2-diol



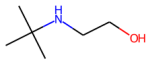
2) piperidin-2-ylmethanol



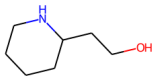
3) 1-azabicyclo[2.2.2]octan-3-ol



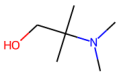
4) 2-(diethylamino)ethanol



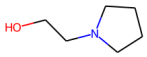
5) 2-(tert-butylamino)ethanol



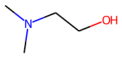
6) 2-piperidin-2-ylethanol



7) 2-(dimethylamino)-2-methylpropan-1-ol

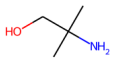


8) 2-pyrrolidin-1-ylethanol

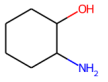


9) 2-(dimethylamino)ethanol

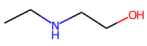
10) piperidine



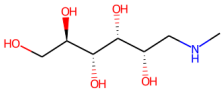
11) 2-amino-2-methylpropan-1-ol



12) 2-aminocyclohexan-1-ol

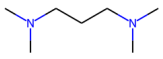


13) 2-(ethylamino)ethanol

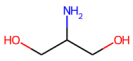


14) (2R,3R,4R,5S)-6-(methylamino)hexane-1,2,3,4,5-pentol

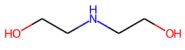
15) (2R)-2-aminopropan-1-ol



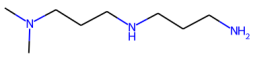
16) N,N,N',N'-tetramethylpropane-1,3-diamine



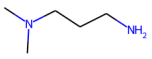
17) 2-aminopropane-1,3-diol



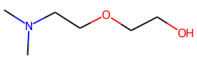
18) 2-(2-hydroxyethylamino)ethanol



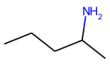
19) N'-[3-(dimethylamino)propyl]propane-1,3-diamine



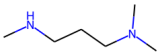
20) N,N'-dimethylpropane-1,3-diamine



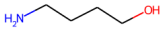
21) 2-[2-(dimethylamino)ethoxy]ethanol



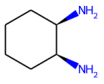
22) pentan-2-amine



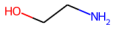
23) N,N',N'-trimethylpropane-1,3-diamine



24) 4-aminobutan-1-ol

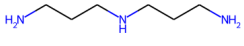


25) (1S,2R)-cyclohexane-1,2-diamine

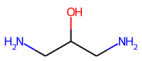


26) 2-aminoethanol

27) morpholine

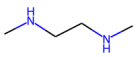


28) N'-(3-aminopropyl)propane-1,3-diamine

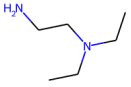


29) 1,3-diaminopropan-2-ol

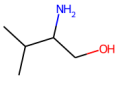
30) 2-amino-2-methylpropane-1,3-diol



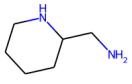
31) N,N'-dimethylethane-1,2-diamine



32) N',N'-diethylethane-1,2-diamine



33) 2-amino-3-methylbutan-1-ol



34) piperidin-2-ylmethanamine

35) azetidine

36) 1-aminobutan-2-ol

37) N,N-diethylpiperidin-4-amine

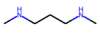
38) butan-1-amine

39) pentan-1-amine

40) (2R)-3-aminopropane-1,2-diol

41) N'N'-diethylpropane-1,3-diamine

42) (2R)-2-amino-4-methylpentan-1-ol



43) N,N'-dimethylpropane-1,3-diamine

44) 1-methyl-1,4-diazepane

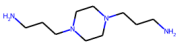
45) 2-(methylamino)ethanol

46) piperazine

47) pyrrolidin-3-amine

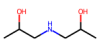
48) 8-methyl-8-azabicyclo[3.2.1]octan-3-amine

49) N,N-dimethylpyrrolidin-3-amine

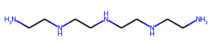


50) 3-[4-(3-aminopropyl)piperazin-1-yl]propan-1-amine

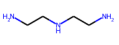
51) N,N-dimethylpiperidin-4-amine



52) 1-(2-hydroxypropylamino)propan-2-ol



53) N'-[2-[2-(2-aminoethylamino)ethylamino]ethyl]ethane-1,2-diamine

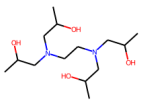


54) N'-(2-aminoethyl)ethane-1,2-diamine

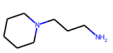
55) (1R,2R)-cyclohexane-1,2-diamine

56) 2-(2-aminoethoxy)ethanamine

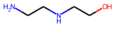
57) 2-aminobutan-1-ol



58) 1-[2-[bis(2-hydroxypropyl)amino]ethyl-(2-hydroxypropyl)amino]propan-2-ol



59) 3-piperidin-1-ylpropan-1-amine



60) 2-(2-aminoethylamino)ethanol

61) (2R)-1-aminopropan-2-ol

62) 2,3,4,6,7,8-hexahydropyrrolo[1,2-a]pyrimidine

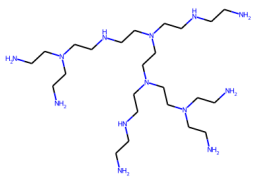
63) 3-aminocyclohexan-1-ol

64) 2,3,4,6,7,8,9,10-octahydropyrimido[1,2-a]azepine

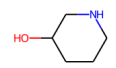
65) 3,4,6,7,8,9-hexahydro-2H-pyrimido[1,2-a]pyrimidine

66) [(2S)-pyrrolidin-2-yl]methanol

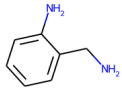
67) (2S)-1-aminopropan-2-ol



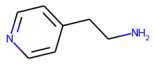
68) N'-[2-[[2-[[2-(2-aminoethylamino)ethyl]-[2-[bis(2-aminoethyl)amino]ethyl]amino]ethyl-[2-[bis(2-aminoethyl)amino]ethyl]amino]ethyl]ethane-1,2-diamine



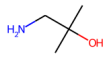
69) piperidin-3-ol



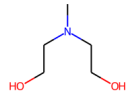
70) 2-(aminomethyl)aniline



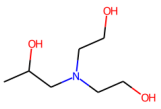
71) 2-pyridin-4-ylethanamine



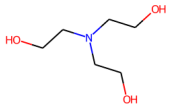
72) 1-amino-2-methylpropan-2-ol



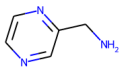
73) 2-[2-hydroxyethyl(methyl)amino]ethanol



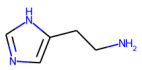
74) 1-[bis(2-hydroxyethyl)amino]propan-2-ol



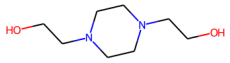
75) 2-[bis(2-hydroxyethyl)amino]ethanol



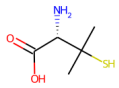
76) pyrazin-2-ylmethanamine



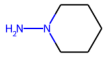
77) 2-(1H-imidazol-5-yl)ethanamine



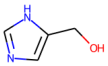
78) 2-[4-(2-hydroxyethyl)piperazin-1-yl]ethanol



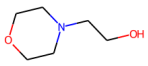
79) (2S)-2-amino-3-methyl-3-sulfanylbutanoic acid



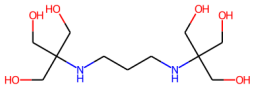
80) piperidin-1-amine



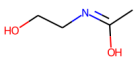
81) 1H-imidazol-5-ylmethanol



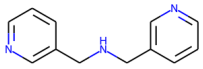
82) 2-morpholin-4-ylethanol



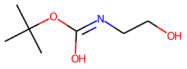
83) 2-[3-[[3-dihydroxy-2-(hydroxymethyl)propan-2-yl]amino]propylamino]-2-(hydroxymethyl)propane-1,3-diol



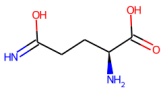
84) N-(2-hydroxyethyl)acetamide



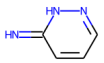
85) 1-pyridin-3-yl-N-(pyridin-3-ylmethyl)methanamine



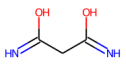
86) tert-butyl N-(2-hydroxyethyl)carbamate



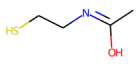
87) (2S)-2,5-diamino-5-oxopentanoic acid



88) pyridazin-3-amine

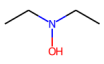


89) propanediamide



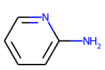
90) N-(2-sulfanylethyl)acetamide

91) 1,3-oxazolidin-2-one



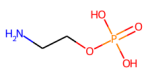
92) N,N-diethylhydroxylamine

93) 1,4-diazabicyclo[2.2.2]octane



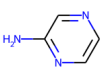
94) pyridin-2-amine

95) 1,3,5,7-tetrazatricyclo[3.3.1.13,7]decane



96) 2-aminoethyl dihydrogen phosphate

97) pyrrol-1-amine



98) pyrazin-2-amine

3 Metrics for classification

The false positive rate (false alarm rate) is calculated as the number of false positives divided by the sum of the number of false positives and the number of true negatives, (Eq.2) and it summarizes how often a positive class is predicted when the actual outcome is negative.

$$\text{False Positive Rate} = \frac{\text{False Positives}}{(\text{False Positives} + \text{True Negatives})} \tag{2}$$

The ROC plot is a useful tool because the curves of different models can be compared directly or indirectly for different thresholds. Also, the area under the curve (AUC) can be used as a summary of the model skill. The shape of the curve contains information about the expected false positive rate, and the false negative rate. Smaller values on the x-axis of the plot indicate lower false positives and higher true negatives, and larger values on the y-axis of the plot indicate higher true positives and lower false negatives. Generally, a model is considered predictive when the curves are shown on the top left of the plot. A perfect model is represented by a line that travels from the bottom left of the plot to the top left and then across the top to the top right and has an AUC of 1. A less predictive classifier is one that cannot discriminate between the classes and would predict a random class or a constant class in all cases. Such a model is represented by a diagonal line from the bottom left of the plot to the top right and has an AUC of 0.5.

Sensitivity is calculated as the ratio of the number of true positives divided by the sum of the true positives and the false negatives (Eq.3), and specificity is calculated as the ratio of the number of true negatives divided by the sum of true negatives and false positives (Eq.4).

$$\text{Sensitivity} = \frac{\text{True Positives}}{(\text{True Positives} + \text{False Negatives})} \tag{3}$$

$$\text{Specificity} = \frac{\text{True Negatives}}{(\text{True Negatives} + \text{False Positives})} \tag{4}$$

Matthews Correlation Coefficient shows the correlation between observed and predicted values and is calculated from Eq.5. It returns a value between -1 and 1. A coefficient of 1 represents a perfect prediction and 0 no better than random prediction. Everything above 0.5 is considered a good correlation.

$$MCC = \frac{(TP \times TN – FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{5}$$

Lastly, geometric mean (g-mean) is the geometric mean of the sensitivity and specificity, [1] this provides a measure of the combined performance of a model for both sensitivity and specificity (Eq.6).

$$\text{g-mean} = \sqrt{\text{sensitivity} \times \text{specificity}} \tag{6}$$

4 CCS fingerprint

Table 2: Substructure in the ccs fingerprint

Indexes	Search
0	[NH3]
1	[NX3;H2][C;!$(C=[#7,#8])]
2	[NX3;H1][C;!$(C=[#7,#8])][C;!$(C=[#7,#8])]
3	[NX3]([C;!$(C=[#7,#8])])([C;!$(C=[#7,#8])])[C;!$(C=[#7,#8])]
4	[NX4+]
5	[N]=[C]
6	[#6] [#7]
7	[a]:[nX3,X2]:[a]
8	[CX3;$([#6]),$([O;H1])](=[OX1])[$([O])]
9	[#6][#6;!$(C(=O)[OH])][OH]
10	[#6][#6]([#6])[OH]
11	[#6][#6]([#6])([#6])[OH]
12	[#6]C([CH3])([CH3])([CH3])
13	[CX3]=[O;!$(O*)]
14	[#6] [F,Cl,Br,I]
15	c1ccccc1
16	[c,n]1[c,n][c,n][c,n][c,n][c,n]1
17	[#6,#8]1 [#6,#8] [#6,#8] [#6,#8] [#6,#8] 1
18	[#6]1 [#6] [#6] [#6] [#6] 1
19	[c,n]1[c,n][c,n][c,n][c,n]1
20	[#6,#8]1 [#6,#8] [#6,#8] [#6,#8] [#6,#8] 1
21	C1CCCCC1
22	[NX3;H2,H1][#6]1 [#6] [#6] [#6] [#6] [#6] 1
23	[NH2]c1ccccc1
24	c1ccccc1[CH2][NH2]
25	C1N([#1])CCCC1
26	c1ncccc1
27	c1n([H])ccc1
28	[$([#6]([OH])[#6][#7H2]);!$([#6]([OH])(=O)[#6][#7H2])]
29	[$([#6]([OH])[#6][#7H]([#6]));!$([#6]([OH])(=O)[#6][#7H]([#6]))]
30	[$([#6]([OH])[#6][#7]([#6])([#6]));!$([#6]([OH])(=O)[#6][#7]([#6])([#6]))]
31	[$([#6]([OH])[#6][#6][#7H2]);!$([#6]([OH])(=O)[#6][#6][#7H2])]
32	[$([#6]([OH])[#6][#6][#7H]([#6]));!$([#6]([OH])(=O)[#6][#6][#7H]([#6]))]
33	[$([#6]([OH])[#6][#6][#7]([#6])([#6]));!$([#6]([OH])(=O)[#6][#6][#7]([#6])([#6]))]
34	C([#6,#1])([#6,#1])([OH])C([#6,#1])([#6,#1])[#7H2]
35	C([#6,#1])([#6,#1])([OH])C([#6,#1])([#6,#1])[#7H]([CX4])
36	C([#6,#1])([#6,#1])([OH])C([#6,#1])([#6,#1])[#7]([CX4])([CX4])
37	C([#6,#1])([#6,#1])([OH])C([#6,#1])([#6,#1])C([#6,#1])([#6,#1])[NH2]
38	C([#6,#1])([#6,#1])([OH])C([#6,#1])([#6,#1])C([#6,#1])([#6,#1])[NH]([CX4])
39	C([#6,#1])([#6,#1])([OH])C([#6,#1])([#6,#1])C([#6,#1])([#6,#1])[N]([CX4])([CX4])
40	[a][C][#7H2]
41	[a][C][C][#7H2]
42	[a][C][C][C][#7H2]
43	[a][C][#7H]([#6])
44	[a][C][C][#7H]([#6])
45	[a][C][C][C][#7H]([#6])
46	[a][C][#7]([#6])([#6])
47	[a][C][C][#7]([#6])([#6])
48	[a][C][C][C][#7]([#6])([#6])
49	[NH2][CX4]([CH3])
50	[NH2][CX4][CX4]([CH3])
51	[NH2][CX4][CX4][CX4]([CH3])
52	[NH2][CX4][CX4][CX4][CX4]([CH3])
53	[NH2][CX4][CX4][CX4][CX4][CX4]([CH3])
54	[NH2][CX4][CX4][CX4][CX4][CX4][CX4]([CH3])
55	[CX4;H2][CX4;H2]
56	[CX4;H2][CX4;H2][CX4;H2]
57	[CX4;H2][CX4;H2][CX4;H2][CX4;H2]
58	[CX4;H2][CX4;H2][CX4;H2][CX4;H2][CX4;H2]
59	[CX4;H2][CX4;H2][CX4;H2][CX4;H2][CX4;H2][CX4;H2]
60	[$([#7X3;H2][C;!$(C=[#7,#8])]),$([#7X3;H1]([C;!$(C=[#7,#8])])[C;!$(C=[#7,#8])]).[$([#7X3;H2][C;!$(C=[#7,#8])]),$([#7X3;H1]([C;!$(C=[#7,#8])])[C;!$(C=[#7,#8])])]
61	[$([#7X3;H2][C;!$(C=[#7,#8])]),$([#7X3;H1]([C;!$(C=[#7,#8])])[C;!$(C=[#7,#8])]),$([#7X3]([C;!$(C=[#7,#8])])([C;!$(C=[#7,#8])])[C;!$(C=[#7,#8])]).[$([#7X3;H2][C;!$(C=[#7,#8])]),$([#7X3;H1]([C;!$(C=[#7,#8])])[C;!$(C=[#7,#8])]),$([#7X3]([C;!$(C=[#7,#8])])([C;!$(C=[#7,#8])])[C;!$(C=[#7,#8])])]
62	[#6][O;H1].[#6][O;H1]
63	n1c([CX4])cnc([CX4])c1
64	n1c([CX4])ccc([CX4])c1
65	n1cccc([CX4])c1
66	[#5]
67	[#14]
68	[#15]

Continued on next page

Table 2: Substructure in the ccs fingerprint

Indexes	Search
69	[#16]
70	[+]
71	[-]

5 Occurrence of fingerprint bits in datasets (fingerprint version 1)

5.1 ZINC dataset

Table 3: Fingerprint SMARTS bit number of occurrence over the ZINC dataset.

Bit	Occurrences
ammonia	0
primary amine	1666
secondary amine	3061
tertiary amine	1908
quaternary N	3
imine	128
nitrogen bonded to carbon	11725
aromatic N sp2	7713
carboxylic acid	3131
primary alcohol	1424
secondary alcohol	756
tertiary alcohol	58
t butyl	85
carbonyl	4465
halocarbon	5095
benezene ring	5827
6 member aromatic c and n ring	9410
6 member c and o ring	272
5 c ring	121
5 member aromatic c and n ring	2434
5 member c and o ring	272
Cyclohexane	242
Cyclohexylamine	472
Aniline	164
benzylamine	99
piperidine	839
pyridine	2880
pyrrole	339
primary amino alcohol two carbon separation	58
secondary amino alcohol two carbon separation	100
tertiary amino alcohol two carbon separation	486
primary amino alcohol three carbon separation	37
secondary amino alcohol three carbon separation	58
tertiary amino alcohol three carbon separation	295
aliphatic primary amino alcohol two carbon separation	53
aliphatic secondary amino alcohol two carbon separation	75
aliphatic tertiary amino alcohol two carbon separation	310
aliphatic primary amino alcohol three carbon separation	22
aliphatic secondary amino alcohol three carbon separation	28

Continued on next page

Table 3: Fingerprint SMARTS bit number of occurrence over the ZINC dataset.

Bit	Occurrences
aliphatic tertiary amino alcohol three carbon separation	83
primary amine one carbon aromatic group	429
primary amine two carbon aromatic group	263
primary amine three carbon aromatic group	33
secondary amine one carbon aromatic group	1051
secondary amine two carbon aromatic group	249
secondary amine three carbon aromatic group	88
tertiary amine one carbon aromatic group	1578
tertiary amine two carbon aromatic group	119
tertiary amine three carbon aromatic group	40
methyl branch one carbon from a N atom	121
methyl branch two carbon from a N atom	82
methyl branch three carbon from a N atom	24
methyl branch four carbon from a N atom	11
methyl branch five carbon from a N atom	0
methyl branch six carbon from a N atom	2
ethyl chain	5605
propyl chain	2251
butyl chain	680
pentyl chain	200
hexyl chain	66
poly primary and or secondary amine	362
poly primary and or secondary and or tertiary amine	1154
poly alcohol	166
pyrazine aliphatic C 2 and 5 substitution	5
pyridine aliphatic C 2 and 5 substitution	40
pyridine aliphatic C 2 substitution	478
Presence of Boron	136
Presence of Silicon	26
Presence of Phosphurus	10
Presence of Sulphur	1708
positive charge group	805
negative charge group	562

5.2 CCS literature datasets

Table 4: Fingerprint SMARTS bit number of occurrence over the ZINC dataset.

Bit	Occurrences
ammonia	0
primary amine	60

Continued on next page

Table 4: Fingerprint SMARTS bit number of occurrences over the ZINC dataset

Bit	Occurrences
secondary amine	45
tertiary amine	41
quaternary N	0
imine	0
nitrogen bonded to carbon	128
aromatic N sp2	18
carboxylic acid	5
primary alcohol	60
secondary alcohol	14
tertiary alcohol	0
t butyl	0
carbonyl	5
halocarbon	0
benezene ring	0
6 member aromatic c and n ring	17
6 member c and o ring	1
5 c ring	1
5 member aromatic c and n ring	1
5 member c and o ring	1
Cyclohexane	4
Cyclohexylamine	4
Aniline	0
benzylamine	0
piperidine	6
pyridine	13
pyrrole	0
primary amino alcohol two carbon separation	9
secondary amino alcohol two carbon separation	14
tertiary amino alcohol two carbon separation	24
primary amino alcohol three carbon separation	2
secondary amino alcohol three carbon separation	2
tertiary amino alcohol three carbon separation	10
aliphatic primary amino alcohol two carbon separation	9
aliphatic secondary amino alcohol two carbon separation	14
aliphatic tertiary amino alcohol two carbon separation	24
aliphatic primary amino alcohol three carbon separation	1
aliphatic secondary amino alcohol three carbon separation	2
aliphatic tertiary amino alcohol three carbon separation	6
primary amine one carbon aromatic group	3
primary amine two carbon aromatic group	0
primary amine three carbon aromatic group	1
secondary amine one carbon aromatic group	6
secondary amine two carbon aromatic group	0

Continued on next page

Table 4: Fingerprint SMARTS bit number of occurrences over the ZINC dataset

Bit	Occurrences
secondary amine three carbon aromatic group	0
tertiary amine one carbon aromatic group	0
tertiary amine two carbon aromatic group	0
tertiary amine three carbon aromatic group	0
methyl branch one carbon from a N atom	6
methyl branch two carbon from a N atom	6
methyl branch three carbon from a N atom	1
methyl branch four carbon from a N atom	1
methyl branch five carbon from a N atom	2
methyl branch six carbon from a N atom	0
ethyl chain	89
propyl chain	44
butyl chain	24
pentyl chain	11
hexyl chain	2
poly primary and or secondary amine	36
poly primary and or secondary and or tertiary amine	49
poly alcohol	23
pyrazine aliphatic C 2 and 5 substitution	1
pyridine aliphatic C 2 and 5 substitution	1
pyridine aliphatic C 2 substitution	4
Presence of Boron	0
Presence of Silicon	0
Presence of Phosphurus	0
Presence of Sulphur	0
positive charge group	0
negative charge group	0

6 Mordred features

We find 35 correlating features which have a significant p-value at 95%, shown in table 5. Following feature generation, we apply one-hot encoding for categorical features and min-max scaling for continuous features. There were 6 features (nBondsM, nBondsKD, C1SP2, HybRatio, FCSP3, ETA_beta_ns) considered as categorical out of the 35. Categorical in this case includes features with specific increments such as counts. Following one hot encoding the feature set extends to 84 as every unique value of the categorical features becomes a binary feature array.

Table 5: Features used as descriptors in the Mordred models based on Spearman correlation of ≥ 0.5 cutoff and significance at 95% on a two tail p-test of the Spearman correlation.

Index	Descriptor name
0	ATS2dv
1	AATS0dv
2	AATS1dv
3	AATS2dv
4	AATS0Z
5	AATS1Z
6	AATS2Z
7	AATS0m
8	AATS1m
9	AATS2m
10	AATS0v
11	AATS1pe
12	AATS2pe
13	AATSC0dv
14	GATS1dv
15	GATS1Z
16	GATS1m
17	BCUTdv-1h
18	nBondsM
19	nBondsKD
20	C1SP2
21	HybRatio
22	FCSP3
23	AXp-2dv
24	MZ
25	Mm
26	Mv
27	AETA_beta
28	ETA_beta_ns
29	AETA_beta_ns
30	AETA_eta_F
31	ETA_epsilon_1
32	PEOE_VSA2
33	SlogP_VSA6
34	AMW

7 Extra Trees Accuracy

Table 6: This table shows the headline accuracy results for using the Extra trees classifier. We see that this classifier is consistently the worst performer in 2 of three feature sets hence not included in the main analysis where we have focused on the best performing models.

Features	Accuracy
Mordred	0.72
MACCS	0.79
CCS	0.80

8 SHAP

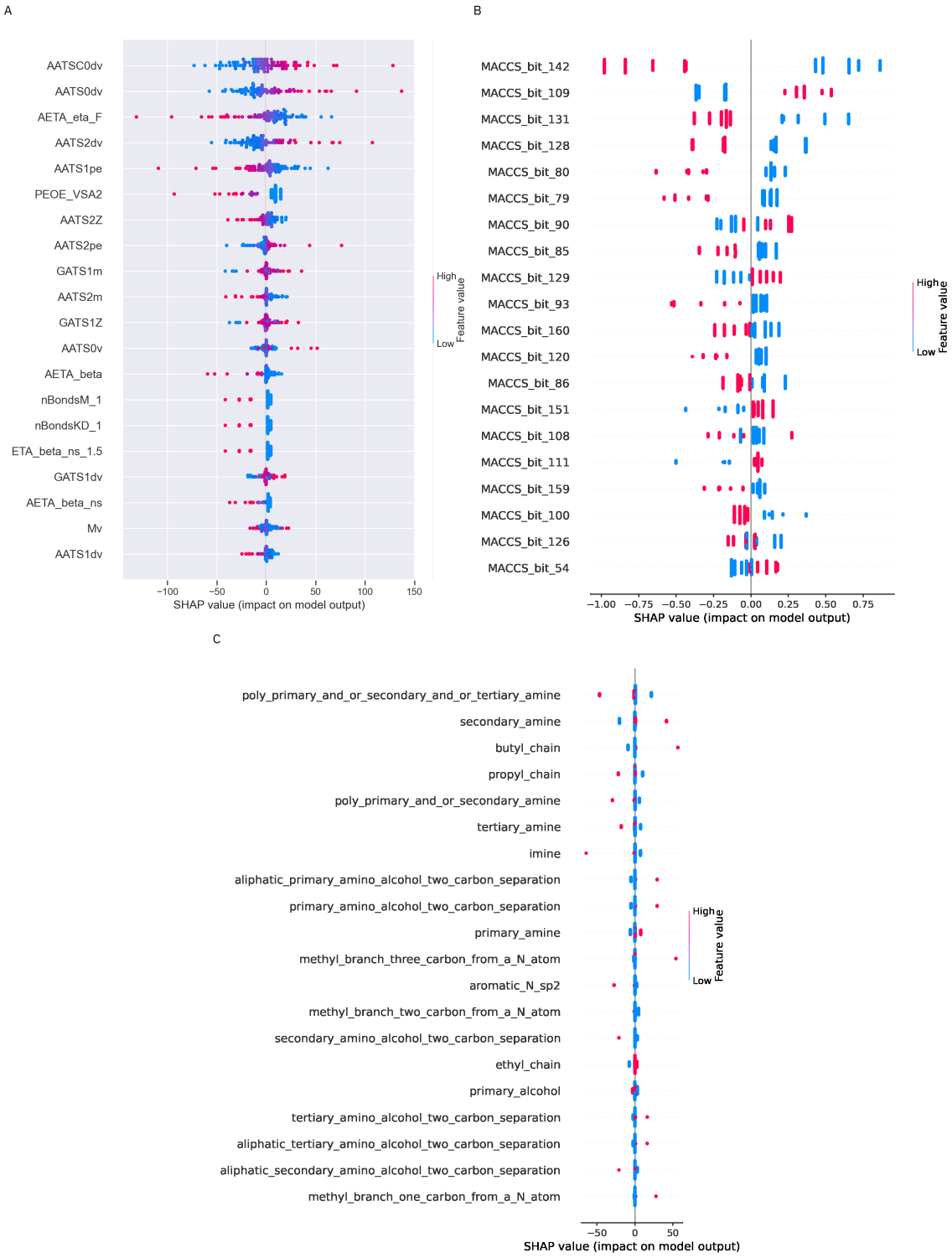


Figure 2: SHAP analysis over a 5 fold cross validation for the Logistic Regression models in this work. Keys can be found RDKit using Chem.MACCSkeys.smartsPatts. We used RDKit version 2022.03.2.

References

[1] Q. Zang, D. M. Rotroff and R. S. Judson, *Journal of chemical information and modeling*, 2013, **53**, 3244–3261.