Electronic Supplementary Material (ESI) for CrystEngComm
This journal is © The Royal Society of Chemistry 2012

SI: "Acid-Base Crystalline Complexes and the pK_a Rule"

Aurora J. Cruz-Cabeza[a,b*]

[a]*Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK.*
[b]*Van 't Hoff Institute for Molecular Sciences, University of Amsterdam, Science Park 904, 1098 XH Amsterdam, The Netherlands. Tel: (+31) 20 525 8264; E-mail: aurorajosecruz@gmail.com*

1. CSD search and filtering of the dataset

A detailed workflow of the methodology followed is given in figure S1. The first filtering criteria were applied using queries or search options with the software Conquest.[1] Conquest was also used to separate between multicomponent crystal structures with no ions and with ions. This resulted in 12 506 AB complexes and 11 421 A^-B^+ complexes.

Some extra filtering criteria were then applied using Pipeline Pilot (Figure S1).[2] At this stage, the reduction in numbers of AB complexes was mainly due to the removal of multicomponent complexes that did not contain an acid-base pair whilst the reduction in numbers of A^-B^+ complexes was mainly due to the removal of salts with cations/anions with charges >+1 & < -1.

It is important to highlight that the acids and bases identified correspond to queries predefined within the pK_a component of Pipeline Pilot[2] only plus the user set defined inorganic acids and water. Calculations of pK_a values were done with Pipeline Pilot (PP)[2] and ChemAxon (CA)[3] pKa calculators on 5258 and 8055 AB and A^-B^+ complexes respectively.

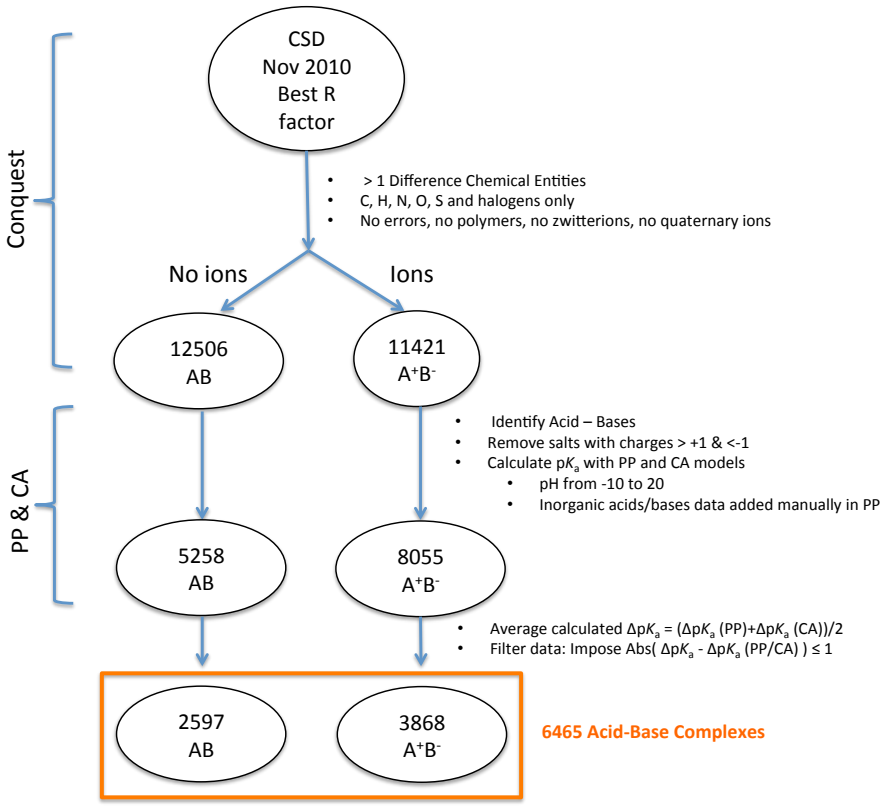

Figure S1. Schematic representation of the workflow.

A filtering criterion was then applied in order to minimise the errors due to the pK_a calculations. Because the PP and CA pK_a calculators are derived differently, it is expected that if both models predict similar values, their average prediction will be reasonably close to the experimental value. Note that the PP pK_a calculator estimates the pK_a data of a given molecule by calculating its molecular

fingerprint and then applying different learning models derived from experimental pK_a values whilst the CA pK_a calculator predicts pK_a data based on calculations of partial charges of atoms in the molecules.

Figure S2 depitcs a histogram of the difference between the final averaged ΔpK_a (the middle value between the PP and the CA prediction) and the pK_a value calculated with PP only. A ΔpK_a cut-off of 1 ΔpK_a only was applied to ensure high quality data in the final dataset.

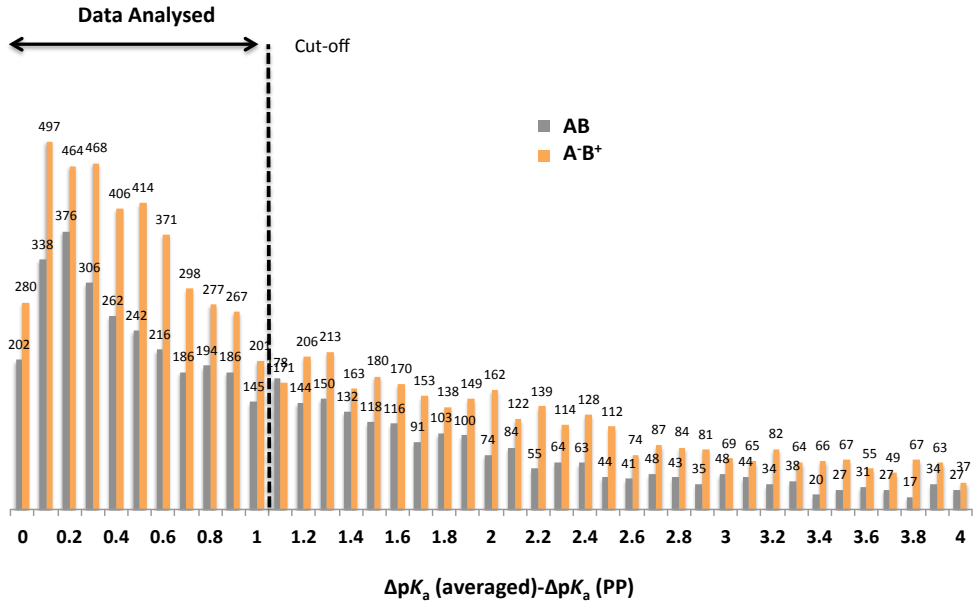

Figure S2. Histogram of the difference between the averaged ΔpK_a and the predicted ΔpK_a with the PP model.

Assuming that the averaged ΔpK_a value is the correct value (or at least closest to the real experimental value), the standard deviation of the final dataset was calculated to be 0.45 only.

2. Exceptions in zones 1 and 3

The refcodes of the 18 A^-B^+ complexes found in zone 1 and the 26 AB complexes found in zone 3 are given bellow.

<u>18 ionised A^-B^+ complexes in zone 1:</u>

LOLDOG, HYCYAN, HETCOZ, QESYAP, KIGJUG, SAJJET, RABCOI, HOTYOF, BUBZED, DIZTUC, DEWYOU, EMOZOV, XUPVIN, CUQHAX, LEZSAL, SOMDUU, POVZOQ and EYIXAL

<u>26 non-ionised AB complexes in zone 3:</u>

GOYHEH, VENLUV, EJUKEZ, DEBTIN, EBIBEW, CAYGUF, PUGGAA, GIQRON, VEKTAG, MICFEJ, CONYAF10, GADQEH, GADQIL, FOCXAX, GADQAD, TIHPAC. YUFQUM, NEBFUV, BEXVOP, XEFDAN, FOGWIH, EVAMIX, QEDBUW, MUFKON, XOTYOV and NIMHEX

3. Acid-Base pKa histograms for complexes in zones 1 and 3

The histogram in figure S3 clearly illustrates the dominance of very weak acids and bases in zone 1 (upper) and very strong acids and bases in zone 3 (lower).

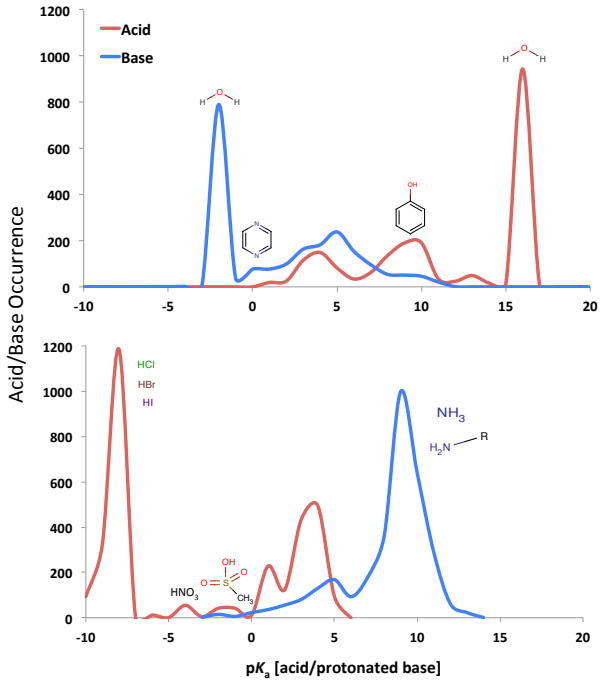

Figure S3. Occurrence of acids (red) and bases (blue) of different strength (pK_a values) in ΔpK_a zones 1 (upper) & 3 (lower).

4. Occurrences of Acids and Bases within the AB and A^-B^+ subsets in zone 2

Table S1 provides information on the number of acids and bases containing a particular substructure found in the AB and A^-B^+ subsets in zone 2. Benzoic acid derivatives are the most commonly found acids together with aliphatic dicarboxylic acids such as fumaric, maleic and tartaric acids. In what concerns to bases, pyridine bases are most common and seem to be found in comparable numbers in both AB and A^-B^+ subsets.

Table S1. Number of crystal structures of AB and A^-B^+ containing the most common acid/base substructures in the zone 2 range of ΔpK_a values.

Acid/Base Substructure	AB *Cocrystals/Solvates*	A^-B^+ *Salts*
benzoic acid	198	230
aliphatic dicarboxylic acid	152	114
aliphatic monocarboxylic acid	74	162
phenol	40	89
pyridine	314	257
amine	25	172
2/5-aminopyrimidine	56	67
imidazole	32	63
aniline	4	73

References:

1. I. J. Bruno, J. C. Cole, P. R. Edgington, M. Kessler, C. F. Macrae, P. McCabe, J. Pearson, R. Taylor, *Acta Crystallogr. B* 2002, **58**, 389-397.
2. SciTegic Pipeline Pilot, Accelrys *Inc.: San Diego, CA* 2008.
3. Marvin 5.10.1, 2012, ChemAxon (http://www.chemaxon.com)