

pK_a Prediction from an *ab initio* bond length

Alex Harding

26th May 2011

Contents

- Background
- Predicting:
 - pK_a Using QCT Descriptors*
 - pK_a From One ab initio Bond Length**
- Methods
- Results
- Conclusions

* Harding et al., J. Chem. Inf. Model, 2009, **49**, 1914

** Harding and Popelier, PCCP, DOI: 10.1039/C1CP20379G (Part 2)

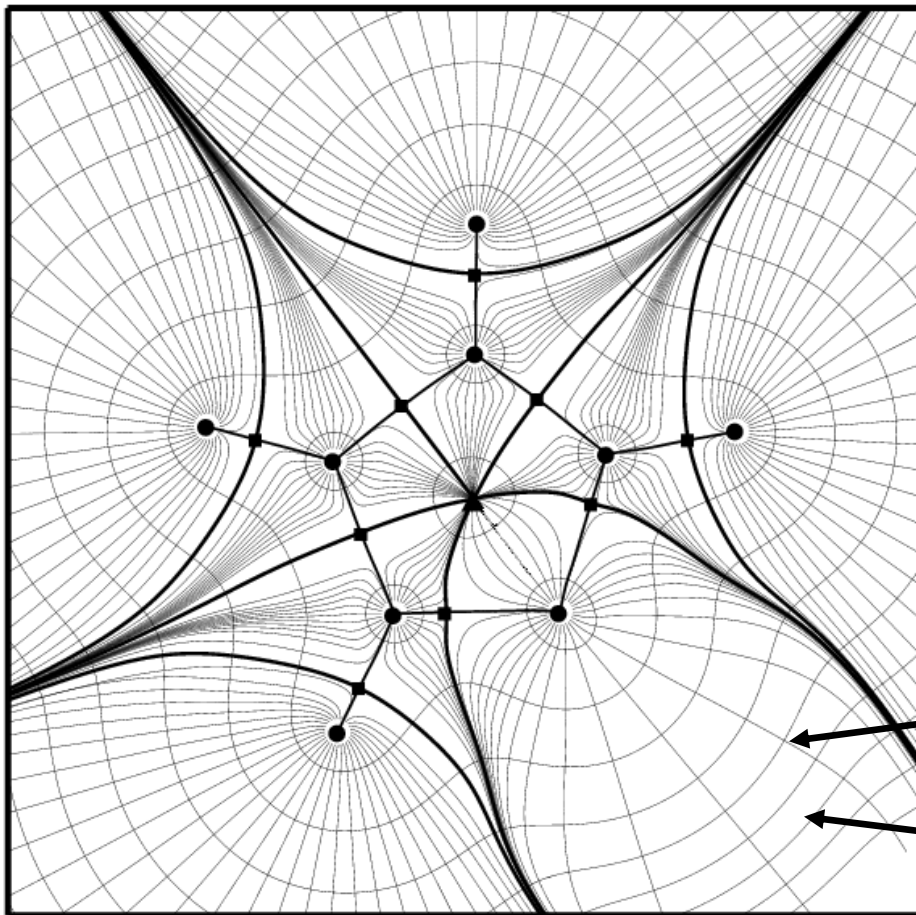
Harding and Popelier, PCCP, DOI: 10.1039/C1CP20380K (Part 3)

Quantum Chemical Topology*

- Chemical insight can be gained from *ab initio* wave functions.
- The electron density “ ρ ” is a source of information.
- The theory forms a bridge between quantum mechanics and working chemical concepts
 - chemical bonding, bioisosterism and force field design

* Popelier, P. L. A, Atoms in Molecules. An Introduction. Pearson Education: London, **2000**.

Partitioning the Electron Density

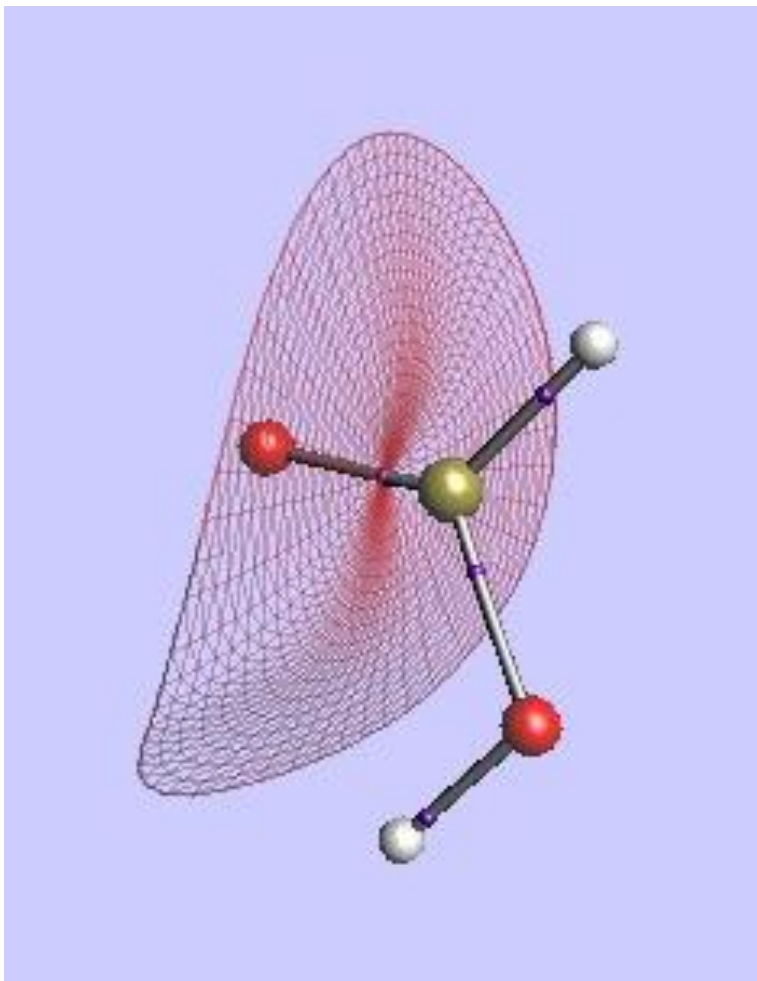


- Gradient vector field superimposed on a contour map of ρ in the symmetry plane of furan

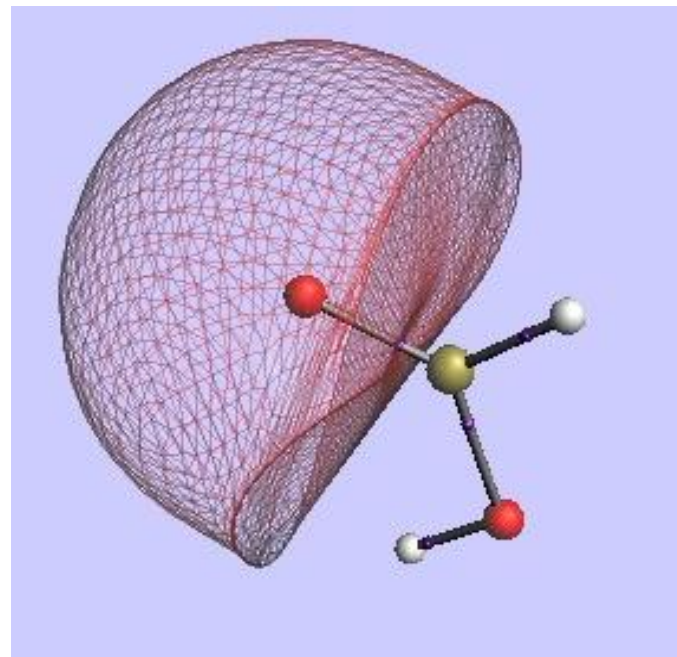
Gradient vector path

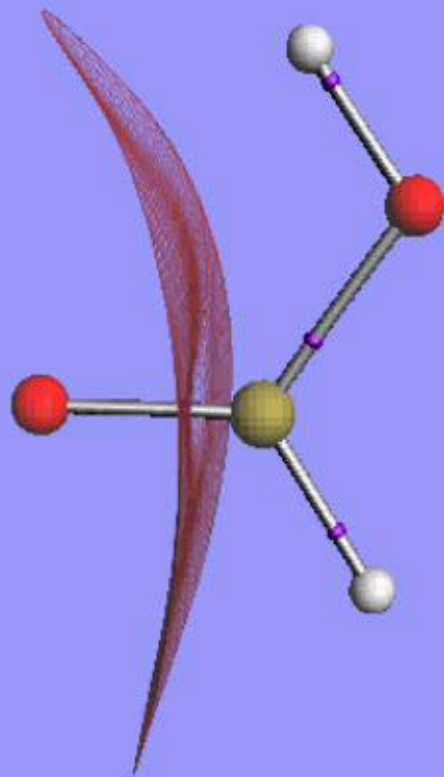
Contour line

Partitioning the Electron Density



- A collection of gradient paths that terminate at the **Bond Critical Point (BCP)** defines the interatomic surface.

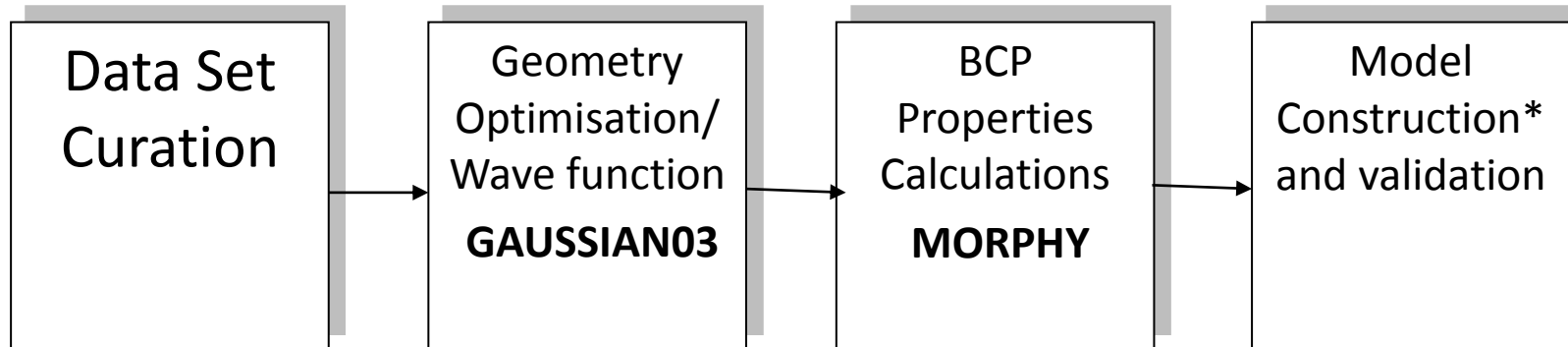




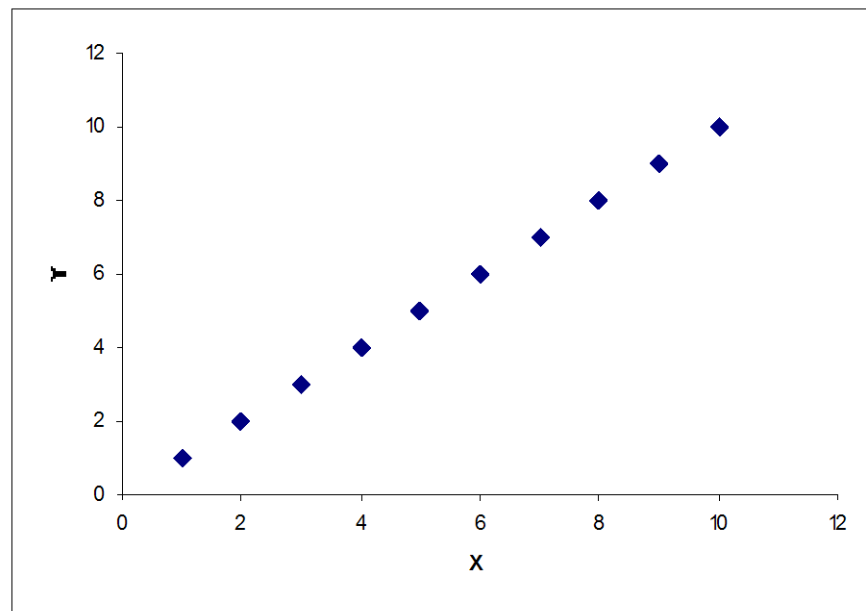
The Bond Critical Point

- A BCP forms when the gradient paths terminate in space between two bonded nuclei ($\nabla \rho = \mathbf{0}$)
 - The BCP is the point where ρ reaches a maximum on the IAS but is a minimum in all other directions.
 - Properties calculated at BCPs include:
 - ρ_b (bond order)
 - Laplacian of ρ " $\nabla^2 \rho_b$ " (ionicity)
 - Ellipticity: $\varepsilon_b = \lambda_1 / \lambda_2 - 1$ (π -character)
 - Two types of kinetic energy $K(r)$ and $G(r)$
 - R_e
 - BCP properties are QCT descriptors
-

Quantum Topological Molecular Similarity (QTMS)* Workflow



- PLS (SIMCA-P+)
- SVM (R)
- RBFNN (MATLAB)



*O'Brien, S. E.; Popelier, P. L. A., JCICS, 2001, **41**, 764.

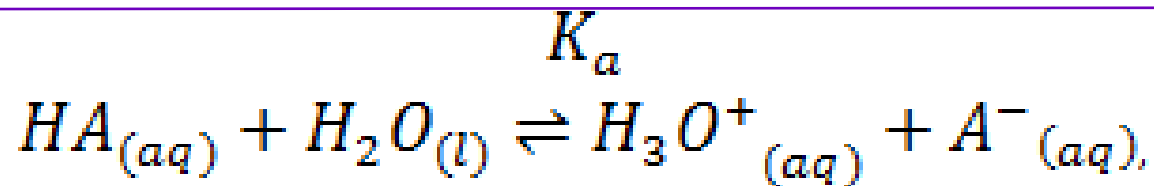
Statistics

$$r^2 = \frac{\sum_{i=1}^n (y_{calc,i} - \bar{y})^2}{\sum_{i=1}^n (y_{obs,i} - \bar{y})^2} \quad \text{RMSEE} = \sqrt{\frac{\sum_{i=1}^n (y_{obs,i} - y_{calc,i})^2}{n - 1 - a}}$$

$$q^2 = 1 - \frac{\sum_{i=1}^n (y_{obs,i} - \hat{y}_{pred,i})^2}{\sum_{i=1}^n (y_{obs,i} - \bar{y})^2}$$

$$\text{RMSEP} = \sqrt{\frac{\sum_{i=1}^n (y_{obs,i} - \hat{y}_{pred,i})^2}{n}} \quad r_m^2 = r^2 \times (1 - \sqrt{r^2 - r_0^2})$$

Predicting pK_a



$$K_a = \frac{[H_3O^+][A^-]}{[HA]}.$$

$$pK_a = -\log_{10}K_a.$$

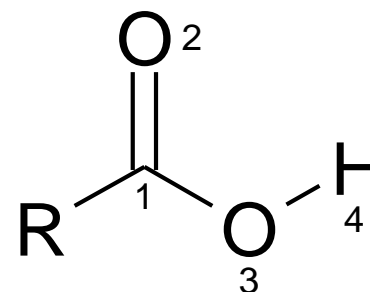
- Why?
 - Chemical and biological processes (ADMET)
 - Established experimental techniques, but no compound required
 - Current methods
-

Using QCT Descriptors

- 228 carboxylic acids of pharmaceutical relevance*.
 - aliphatic and benzoic acids.

- 5 levels of theory:

- A = AM1
- B = HF/3-21G(d)
- C = HF/6-31G(d)
- D = B3LYP/6-31+G(d,p)
- E = B3LYP/6-311+G(2d,p)



Common Skeleton

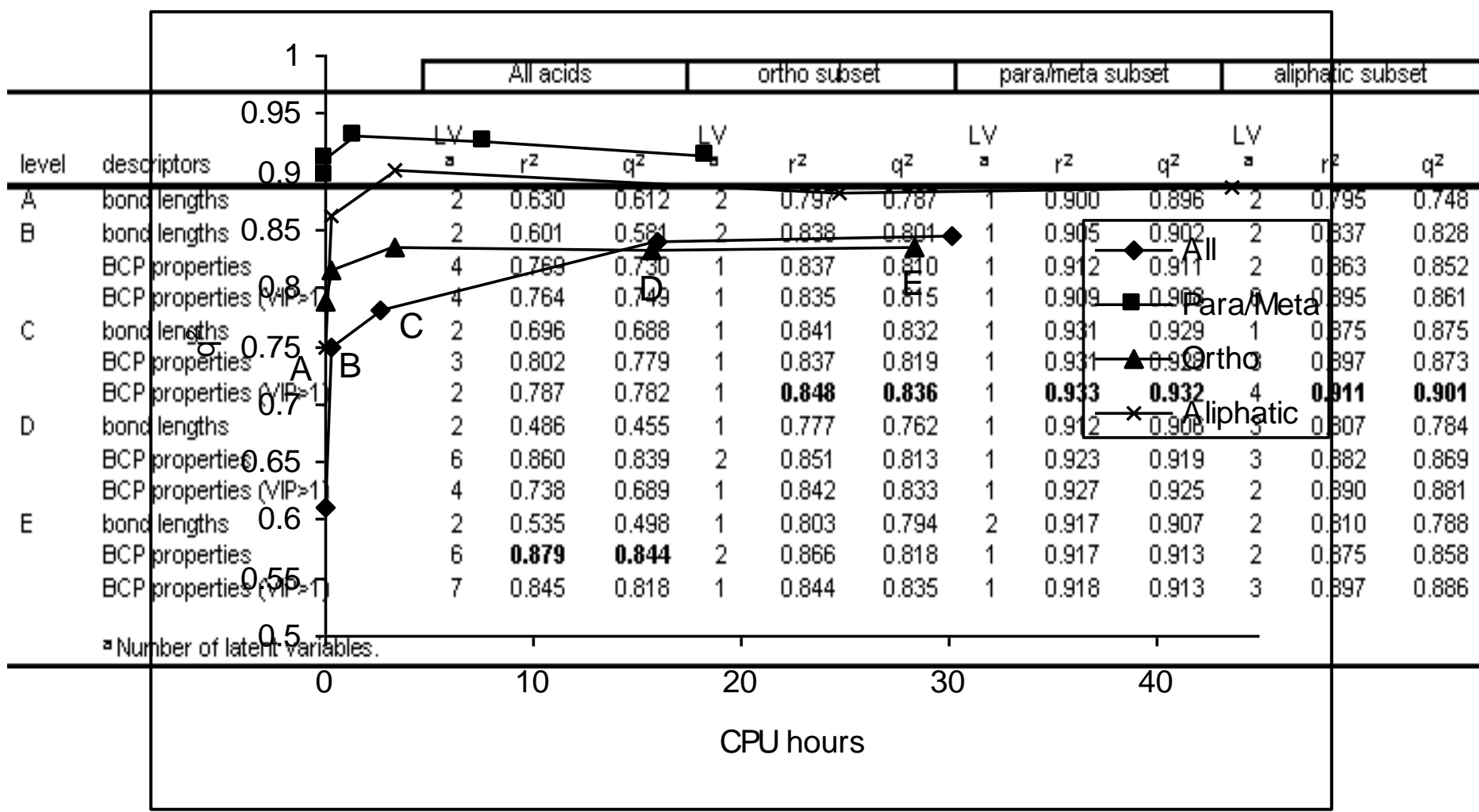
- 7 descriptors for each BCP

* Tehan *et al.*, QSAR, 2001, **21**, 457

Results

- Compare PLS, SVM and RBFNN as learning methods.
 - Validate using 4-, 7-, and 10-fold cross validation.
 - Selection of validation groups based on random and systematic sampling.
 - Comparison of HF/6-31G(d) (level C) and B3LYP/6-311+G(2d,p) (level E).
-

Level of Theory



* Wold, S. PLS in Chemistry, *Encycl. Of Comp. Chem.* 1998

Averaged Results

SVM better

Level C
(HF/6-31G(d))
better

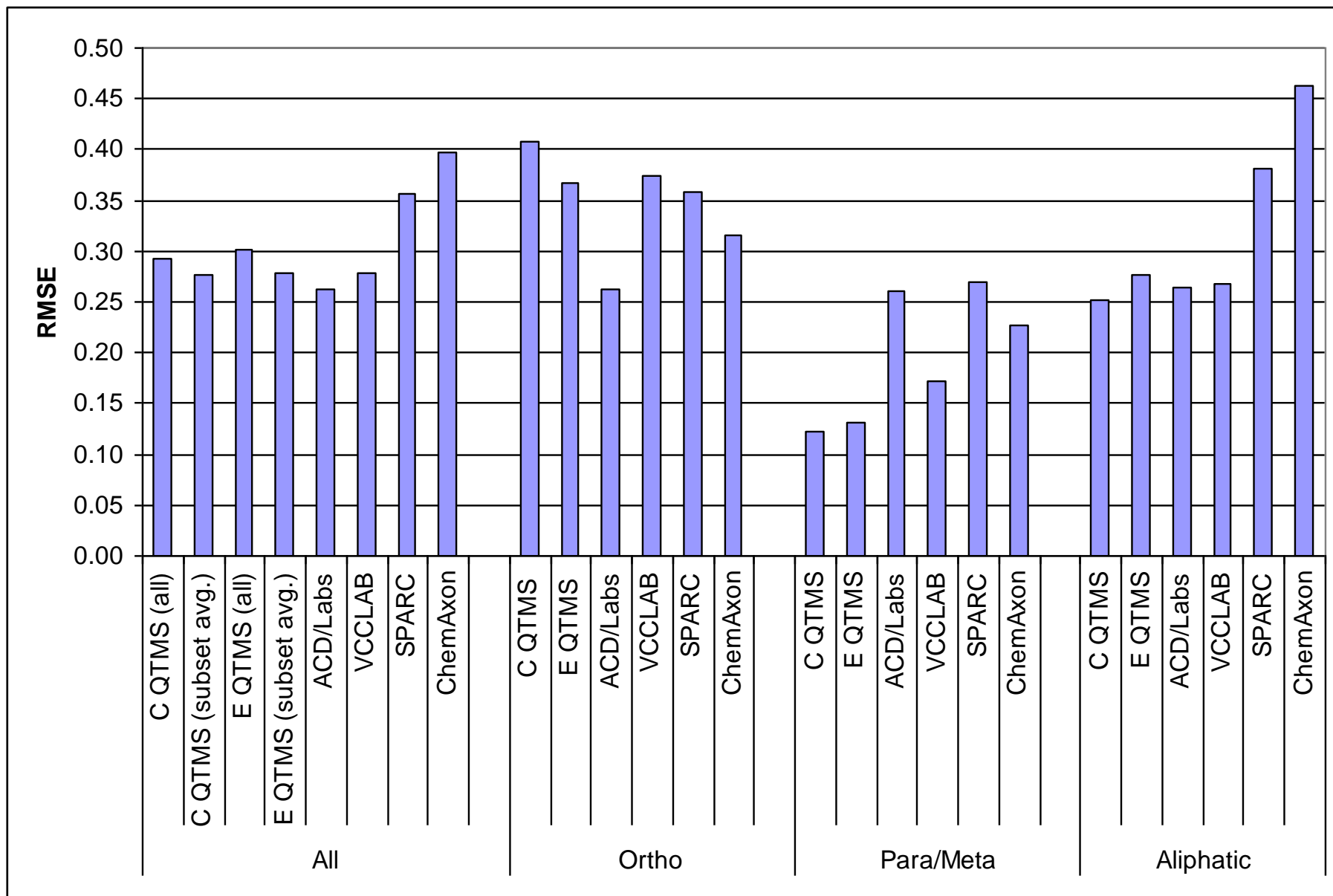
| | | PLS | | SVM | | RBFNN | |
|---|-----------|----------------|-------|----------------|-------|----------------|-------|
| Level C | Set | q ² | RMSE | q ² | RMSE | q ² | RMSE |
| Systematic Sampling | All | 0.780 | 0.422 | 0.893 | 0.291 | 0.897 | 0.293 |
| | Ortho | 0.834 | 0.375 | 0.821 | 0.380 | 0.784 | 0.427 |
| | Para/Meta | 0.930 | 0.109 | 0.924 | 0.113 | 0.896 | 0.128 |
| | Aliphatic | 0.906 | 0.296 | 0.903 | 0.271 | 0.891 | 0.294 |
| Random Sampling | All | 0.777 | 0.430 | 0.879 | 0.295 | 0.875 | 0.322 |
| | Ortho | 0.821 | 0.377 | 0.829 | 0.375 | 0.774 | 0.426 |
| | Para/Meta | 0.928 | 0.110 | 0.922 | 0.110 | 0.634 | 0.212 |
| | Aliphatic | 0.904 | 0.270 | 0.909 | 0.265 | 0.885 | 0.303 |
| Average of systematic and random sampling | All | 0.778 | 0.426 | 0.886 | 0.293 | 0.886 | 0.307 |
| | Ortho | 0.828 | 0.376 | 0.825 | 0.378 | 0.779 | 0.427 |
| | Para/Meta | 0.929 | 0.109 | 0.923 | 0.112 | 0.765 | 0.170 |
| | Aliphatic | 0.905 | 0.283 | 0.906 | 0.268 | 0.888 | 0.298 |
| Level E | | | | | | | |
| Systematic Sampling | All | 0.817 | 0.381 | 0.880 | 0.291 | 0.872 | 0.327 |
| | Ortho | 0.825 | 0.386 | 0.851 | 0.356 | 0.763 | 0.431 |
| | Para/Meta | 0.913 | 0.129 | 0.916 | 0.125 | 0.893 | 0.144 |
| | Aliphatic | 0.884 | 0.299 | 0.905 | 0.267 | 0.852 | 0.336 |
| Random Sampling | All | 0.805 | 0.392 | 0.887 | 0.295 | 0.881 | 0.316 |
| | Ortho | 0.861 | 0.389 | 0.822 | 0.380 | 0.686 | 0.508 |
| | Para/Meta | 0.896 | 0.136 | 0.912 | 0.127 | 0.820 | 0.176 |
| | Aliphatic | 0.881 | 0.300 | 0.903 | 0.273 | 0.886 | 0.301 |
| Average of systematic and random sampling | All | 0.811 | 0.387 | 0.883 | 0.293 | 0.876 | 0.322 |
| | Ortho | 0.843 | 0.388 | 0.837 | 0.368 | 0.724 | 0.469 |
| | Para/Meta | 0.905 | 0.132 | 0.914 | 0.126 | 0.857 | 0.160 |
| | Aliphatic | 0.883 | 0.300 | 0.904 | 0.270 | 0.869 | 0.319 |

Results

- Improved models were generated when the dataset was split.
 - $RMSE < 0.5$
 - SVM generally provides the best model.
 - The models are statistically robust using both random and systematic sampling methods to select the CV groups.
 - There is little different between the statistics for level C and E
 - level C should be used when considering the large increase in CPU times associated with level E
-

Comparison to Other Tools

- Based on the LOO using the RMSE.
 - Compared to ACD/Labs, VCCLAB, SPARC and ChemAxon.
 - Using the SVM models at HF/3-21G(d) and B3LYP/6-311+G(2d,p).
-



Conclusions

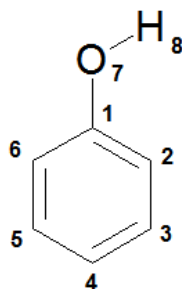
- Robust models have been created for the largest set of molecules studied using QTMS.
 - SVM generally proved to be the best learning method.
 - The predictions are comparable to pK_a tools frequently being used.
 - Harding *et al.*, J. Chem. Inf. Model, 2009, **49**, 1914.
-

Using a Single Bond Length

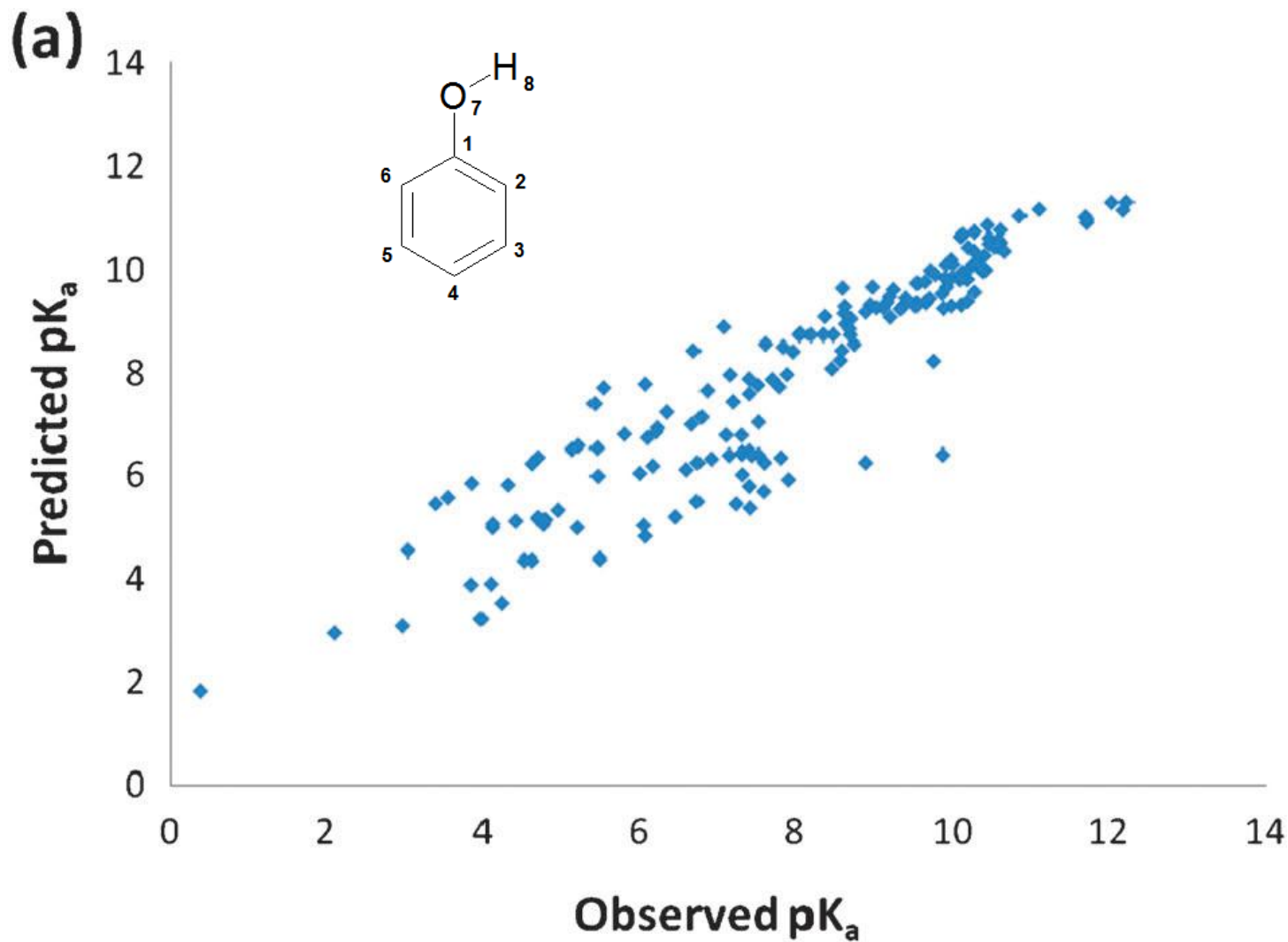
- Focus
 - Accuracy (<0.5 pK_a units)
 - Low level of theory (HF/6-31G(d))
 - Simplicity
 - Interpretation
 - Previous work
 - bromo-, chloro-, fluorophenols
 - Mono-substituted Carboxylic acids
 - All use an ammonia probe
-

Example - Phenols

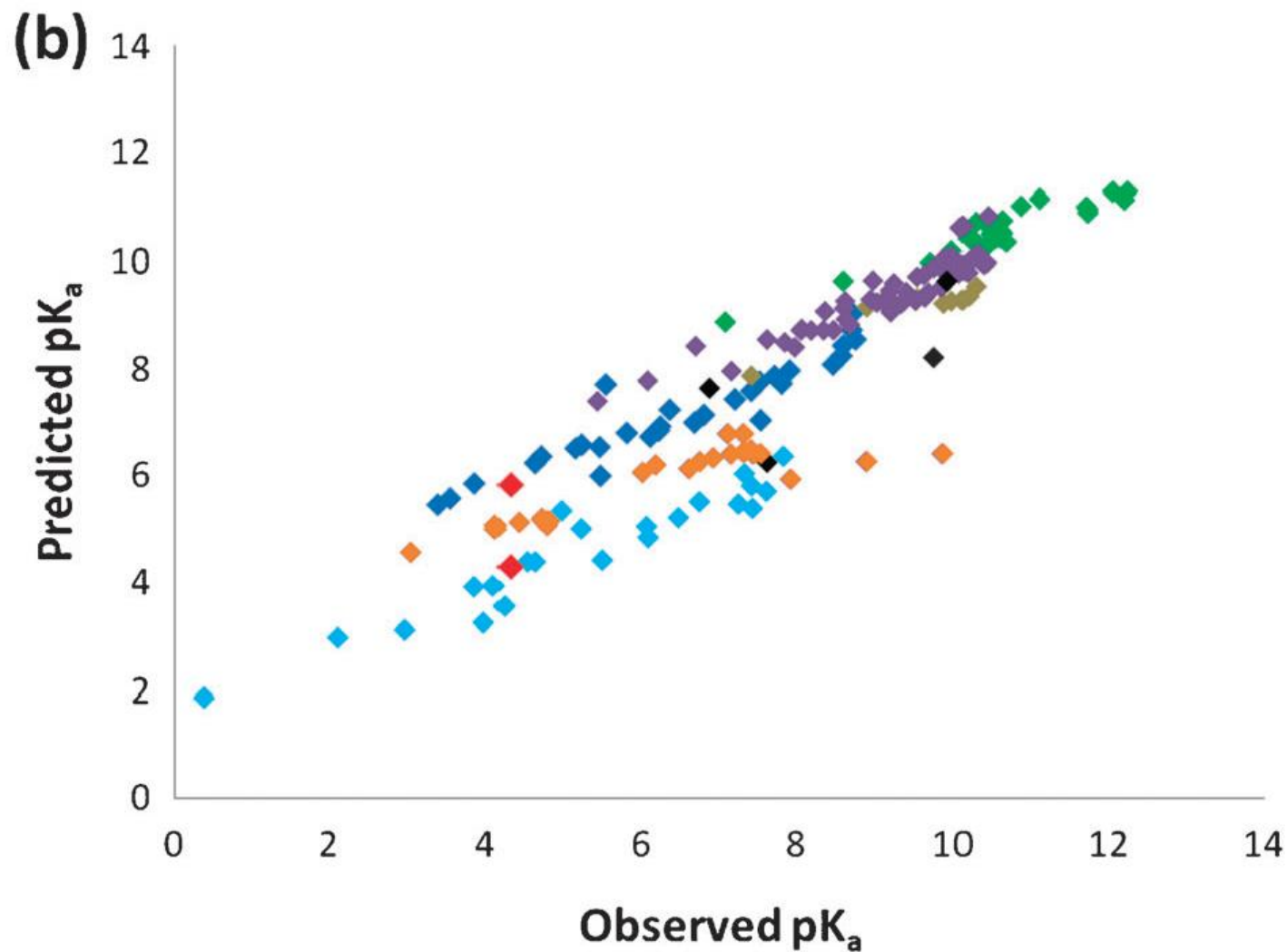
| Subsets | # LV | # Bonds | # Compounds | r^2 | q^2 | RMSEE |
|---------------------------------------|------|-----------------|-------------|-------|-------|-------|
| All | 4 | All | 171 | 0.92 | 0.88 | 0.67 |
| All | 1 | $r(\text{C-O})$ | 171 | 0.86 | 0.85 | 0.88 |
| All | 1 | $r(\text{O-H})$ | 171 | 0.52 | 0.51 | 1.62 |
| <i>Meta/Para</i> | 2 | All | 55 | 0.91 | 0.87 | 0.34 |
| <i>Meta/Para</i> | 1 | $r(\text{C-O})$ | 55 | 0.87 | 0.85 | 0.41 |
| <i>Meta/Para</i> | 1 | $r(\text{O-H})$ | 55 | 0.84 | 0.83 | 0.45 |
| <i>Ortho</i> | 4 | All | 116 | 0.92 | 0.86 | 0.72 |
| <i>Ortho</i> | 1 | $r(\text{C-O})$ | 116 | 0.85 | 0.85 | 0.99 |
| <i>Ortho</i> | 1 | $r(\text{O-H})$ | 116 | 0.47 | 0.46 | 1.84 |
| <i>Ortho</i> without <i>Ortho-IHB</i> | 5 | All | 90 | 0.94 | 0.88 | 0.65 |
| <i>Ortho</i> without <i>Ortho-IHB</i> | 1 | $r(\text{C-O})$ | 90 | 0.88 | 0.87 | 0.94 |
| <i>Ortho</i> without <i>Ortho-IHB</i> | 1 | $r(\text{O-H})$ | 90 | 0.59 | 0.58 | 1.72 |



Phenols $r(\text{C-O})$



High-correlation subsets



◆ o-halogen

◆ o-alkyl

◆ o-misc

◆ o-nitro

◆ o-methoxy/ethoxy

◆ meta/para

◆ o-IHB

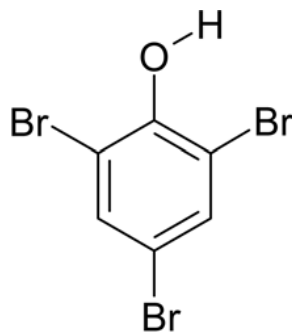
◆ 4,6-dinitro-o-cresol (compound 135)

Ortho-Halogen Phenols High-Correlation Subset

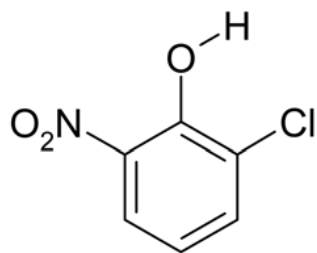
| Method | Conformation | # Bonds | # LV | # Compounds | Model statistics | | CV statistics | | | | |
|--------|--------------|----------|------|-------------|------------------|-------------|---------------|----------------|-------------|--------------|-------------|
| | | | | | r^2 | RMSEE | q^2 | RMSEP | r_{CV}^2 | $r_{CV,0}^2$ | r_m^2 |
| HF | <i>Syn</i> | All | 1 | 32 | 0.75 | 0.89 | 0.71 | 1.05 | 0.64 | 0.55 | 0.45 |
| HF | <i>Syn</i> | $r(C-O)$ | 1 | 32 | 0.88 | 0.61 | 0.87 | 0.62 | 0.87 | 0.86 | 0.79 |
| HF | <i>Syn</i> | $r(O-H)$ | 1 | 32 | 0.63 | 1.08 | 0.63 | — ^a | — | — | — |
| HF | <i>Syn</i> | All | 2 | 26 | 0.95 | 0.40 | 0.91 | 0.44 | 0.93 | 0.91 | 0.81 |
| HF | <i>Syn</i> | $r(C-O)$ | 1 | 26 | 0.97 | 0.27 | 0.97 | 0.29 | 0.97 | 0.97 | 0.94 |
| HF | <i>Syn</i> | $r(O-H)$ | 1 | 26 | 0.60 | 1.06 | 0.58 | — | — | — | — |
| HF | <i>Anti</i> | All | 2 | 32 | 0.83 | 0.74 | 0.70 | — | — | — | — |
| HF | <i>Anti</i> | $r(C-O)$ | 1 | 32 | 0.83 | 0.72 | 0.81 | — | — | — | — |
| HF | <i>Anti</i> | $r(O-H)$ | 1 | 32 | 0.61 | 1.12 | 0.53 | — | — | — | — |
| HF | <i>Anti</i> | All | 2 | 26 | 0.94 | 0.40 | 0.91 | 0.46 | 0.92 | 0.91 | 0.81 |
| HF | <i>Anti</i> | $r(C-O)$ | 1 | 26 | 0.96 | 0.35 | 0.96 | 0.35 | 0.95 | 0.95 | 0.91 |
| HF | <i>Anti</i> | $r(O-H)$ | 1 | 26 | 0.78 | 0.79 | 0.76 | — | — | — | — |
| B3LYP | <i>Syn</i> | All | 1 | 32 | 0.71 | 0.97 | 0.65 | — | — | — | — |
| B3LYP | <i>Syn</i> | $r(C-O)$ | 1 | 32 | 0.87 | 0.64 | 0.86 | — | — | — | — |
| B3LYP | <i>Syn</i> | $r(O-H)$ | 1 | 32 | 0.52 | 1.24 | 0.44 | — | — | — | — |
| B3LYP | <i>Syn</i> | All | 2 | 26 | 0.93 | 0.46 | 0.86 | 0.67 | 0.83 | 0.81 | 0.70 |
| B3LYP | <i>Syn</i> | $r(C-O)$ | 1 | 26 | 0.96 | 0.32 | 0.96 | 0.33 | 0.96 | 0.96 | 0.92 |
| B3LYP | <i>Syn</i> | $r(O-H)$ | 1 | 26 | 0.72 | 0.89 | 0.70 | — | — | — | — |

^a A dash in this Table 5–7 indicates that various cross-validation statistics were not collected as justified in the main text.

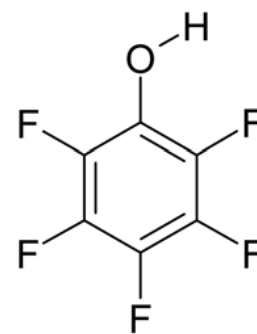
Outliers



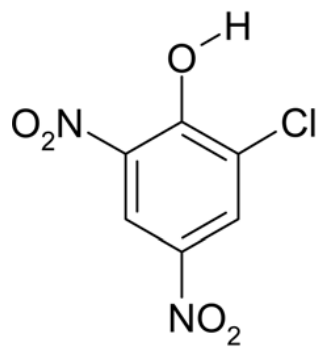
120



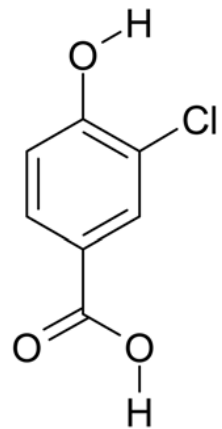
141



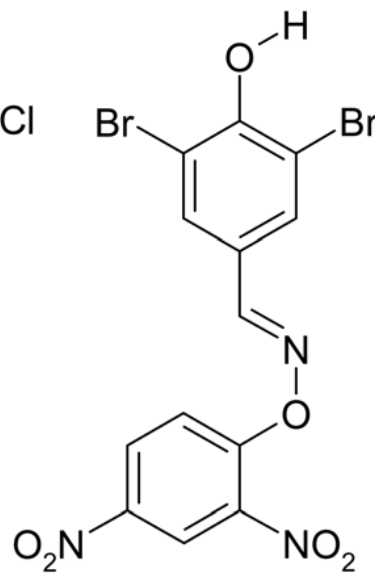
156



160



171



175

High-Correlation Subsets

- Phenols
 - ortho-phenols
 - o-nitro
 - o-halogen
 - o-alkyl
 - o-methoxy/ethoxy
 - o-IHB
 - meta/para-phenols
 - Already high quality model and no significant improvement
 - Carboxylic acids
 - Anilines
-

Equations

| Compound class | High-correlation subset | Equation | # Compounds | r^2 | RMSEE | q^2 | RMSEP | r_{CV}^2 | $r_{CV,0}^2$ | r_m^2 |
|----------------|---------------------------------|------------------------------------|-------------|-------|-------|-------|-------|------------|--------------|---------|
| Phenols | <i>o</i> -Nitro phenols | $pK_a = 137.575 r(C-O) - 337.575$ | 22 | 0.94 | 0.48 | 0.94 | 0.50 | 0.93 | 0.93 | 0.87 |
| | <i>o</i> -Halogen phenols | $pK_a = 147.411 r(C-O) - 365.540$ | 26 | 0.97 | 0.27 | 0.97 | 0.29 | 0.97 | 0.97 | 0.94 |
| | <i>o</i> -Alkyl phenols | $pK_a = 162.106 r(C-O) - 405.207$ | 25 | 0.91 | 0.34 | 0.90 | 0.37 | 0.89 | 0.87 | 0.78 |
| | <i>o</i> -Phenols-IHB | $pK_a = 160.912 r(C-O) - 397.578$ | 23 | 0.95 | 0.32 | 0.95 | 0.33 | 0.94 | 0.94 | 0.92 |
| | <i>o</i> -Methoxy/ethoxyphenols | $pK_a = 128.767 r(C-O) - 318.646$ | 24 | 0.91 | 0.29 | 0.89 | 0.53 | 0.69 | 0.57 | 0.45 |
| | <i>m-/p</i> -Phenols | $pK_a = 122.985 r(C-O) - 304.553$ | 55 | 0.87 | 0.41 | 0.85 | 0.43 | 0.85 | 0.83 | 0.72 |
| Benzoic Acids | <i>o</i> -halogen benzoic acids | $pK_a = -53.316r(C-C) + 153.475$ | 13 | 0.93 | 0.16 | 0.92 | 0.19 | 0.90 | 0.89 | 0.82 |
| | <i>o</i> -hydroxy benzoic acids | $pK_a = -975.258r(O-H) + 1758.265$ | 14 | 0.98 | 0.13 | 0.97 | 0.15 | 0.96 | 0.96 | 0.92 |
| | <i>m-/p</i> -benzoic acids | $pK_a = -770.717r(O-H) + 1390.747$ | 43 | 0.92 | 0.13 | 0.91 | 0.13 | 0.91 | 0.91 | 0.84 |
| Anilines | <i>o</i> -halogen anilines | $pK_a = 83.401r(C-N) - 216.089$ | 10 | 0.95 | 0.39 | 0.94 | 0.44 | 0.93 | 0.92 | 0.90 |
| | <i>m-/p</i> -anilines | $pK_a = 67.932r(C-N) - 175.213$ | 23 | 0.80 | 0.48 | 0.76 | 0.54 | 0.73 | 0.68 | 0.57 |
| | <i>p</i> -anilines | $pK_a = 80.644r(C-N) - 208.635$ | 11 | 0.95 | 0.27 | 0.93 | 0.33 | 0.92 | 0.92 | 0.92 |
| | <i>m</i> -aniline | $pK_a = 56.090r(C-N) - 144.080$ | 12 | 0.66 | 0.59 | 0.61 | 0.63 | 0.56 | 0.41 | 0.34 |

Conclusions

- Single-bond-length models comparable or superior to all-bond-length models
 - High-correlation subset exposes outliers
 - Unexpected bond length may deliver best model
 - RMSEE <0.5 for all models
 - Harding and Popelier, PCCP,
DOI: 10.1039/C1CP20379G (Part 2)
 - Harding and Popelier, PCCP,
DOI: 10.1039/C1CP20380K (Part 3)
-

Acknowledgements

Sponsors

EPSRC

GlaxoSmithKline

Supervisors

Prof Paul Popelier (Academic)

Dr Michael Kranz (Industrial)

pK_a Work

Dr David Wedge (UoM)

All colleagues at UoM and GSK

Backup Slides

Outlier Detection

