

# Rapid, Automated Prediction of Abraham LSER Descriptors

Jamie Platts  
Cardiff University

20<sup>th</sup> October 1999

# Background

## LSER Approach

- Size, polarity, hydrogen bonding descriptors

# Background

## LSER Approach

- Size, polarity, hydrogen bonding descriptors
- Describe partition and passive transport in terms of these

# Background

## LSER Approach

- Size, polarity, hydrogen bonding descriptors
- Describe partition and passive transport in terms of these
- Not appropriate for active transport

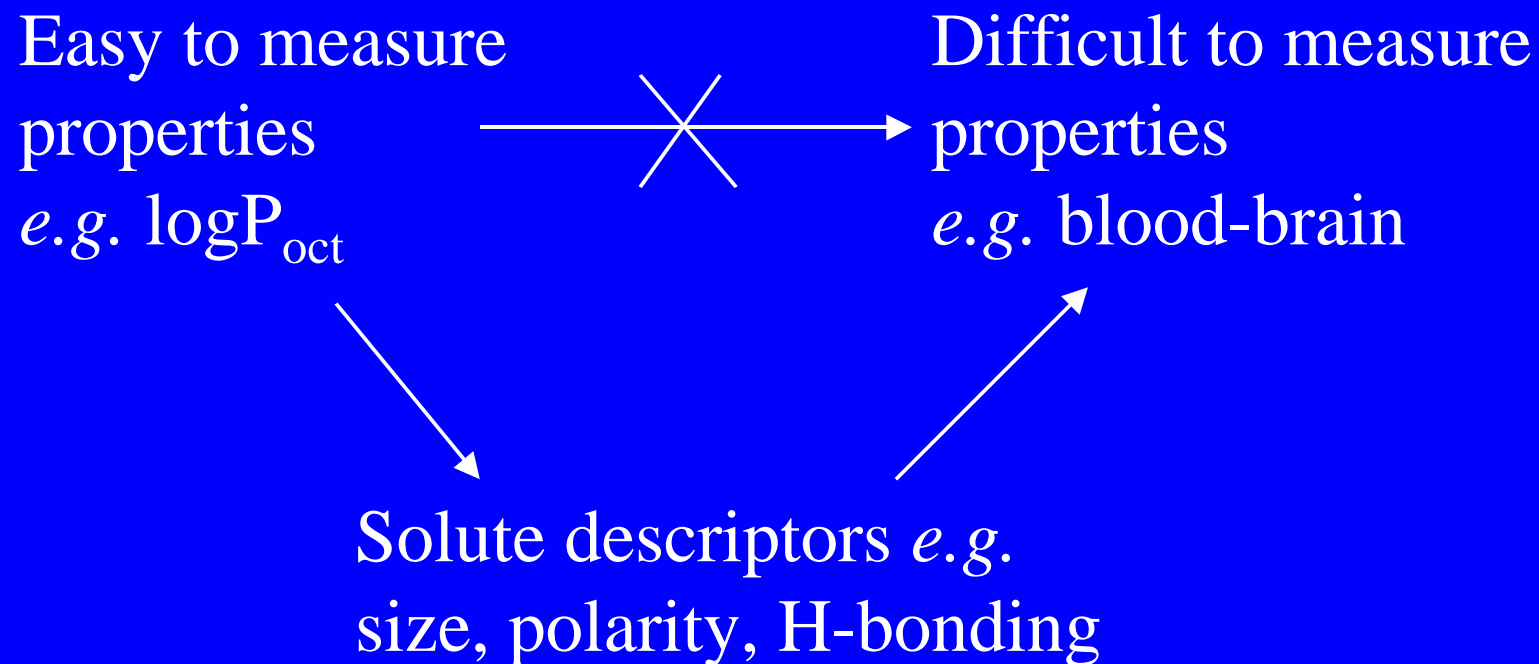
# LSER Approach

Easy to measure  
properties  
*e.g.*  $\log P_{\text{oct}}$



Difficult to measure  
properties  
*e.g.* blood-brain

# LSER Approach



# LSER Approach

## Solvation equation

- Linear combination of descriptors - general solvation equation

# LSER Approach

## Solvation equation

- Linear combination of descriptors - general solvation equation
- Descriptors all derived from free-energy related properties

## Solvation equation

$$\log SP = c + e.E + s.S + a.A + b.B + v.V_x$$

# Solvation equation

$$\log SP = c + e.E + s.S + a.A + b.B + v.V_x$$

Coefficients  $c$ ,  $e$ ,  $s$ ,  $a$ ,  
 $b$ , and  $v$  characterise  
specific solvent system

# Solvation equation

$$\log SP = c + e.E + s.S + a.A + b.B + v.Vx$$

Coefficients  $c$ ,  $e$ ,  $s$ ,  $a$ ,  
 $b$ , and  $v$  characterise  
specific solvent system

Solute descriptors  $E$ ,  $S$ ,  $A$ ,  $B$ ,  
and  $Vx$  characterise molecule

## Solvation equation

$$\log SP = c + e.E + s.S + a.A + b.B + v.V_X$$

where:

**E** = excess molar refraction (formerly  $R_2$ )

## Solvation equation

$$\log SP = c + e.E + s.S + a.A + b.B + v.Vx$$

where:

E = excess molar refraction (formerly  $R_2$ )

S = polarity/polarisability ( $\pi_2^H$ )

## Solvation equation

$$\log SP = c + e.E + s.S + a.A + b.B + v.Vx$$

where:

E = excess molar refraction (formerly  $R_2$ )

S = polarity/polarisability ( $\pi_2^H$ )

A = hydrogen bond acidity ( $\Sigma\alpha_2^H$ )

## Solvation equation

$$\log SP = c + e.E + s.S + a.A + b.B + v.V_X$$

where:

E = excess molar refraction (formerly  $R_2$ )

S = polarity/polarisability ( $\pi_2^H$ )

A = hydrogen bond acidity ( $\Sigma\alpha_2^H$ )

B = hydrogen bond basicity ( $\Sigma\beta_2^H$ )

## Solvation equation

$$\log SP = c + e.E + s.S + a.A + b.B + v.V_x$$

where:

E = excess molar refraction (formerly  $R_2$ )

S = polarity/polarisability ( $\pi_2^H$ )

A = hydrogen bond acidity ( $\Sigma\alpha_2^H$ )

B = hydrogen bond basicity ( $\Sigma\beta_2^H$ )

$V_x$  = McGowan volume

# Solvation descriptors

Some typical values of A and B

Solute	A	B
Hexane	0.00	0.00
Methanol	0.43	0.47
Phenol	0.60	0.30
Acetic acid	0.61	0.44
Urea	0.50	0.90
Triethylamine	0.00	0.79

# Solvation equation

Example: octanol/water partition

$$\log P_{\text{oct}} = 0.09 + 0.56E - 1.05S + 0.03A - 3.46B + 3.81V_x$$

$$n=613, \quad r^2=0.995, \quad \text{s.d.}=0.116$$

# Solvation equation

Example: octanol/water partition

$$\log P_{\text{oct}} = 0.09 + 0.56E - 1.05S + 0.03A - 3.46B + 3.81V_x$$

water more polar and acidic than octanol



# Solvation equation

Example: octanol/water partition

$$\log P_{\text{oct}} = 0.09 + 0.56E - 1.05S + 0.03A - 3.46B + 3.81V_x$$

Phases of equal basicity



# Solvation equation

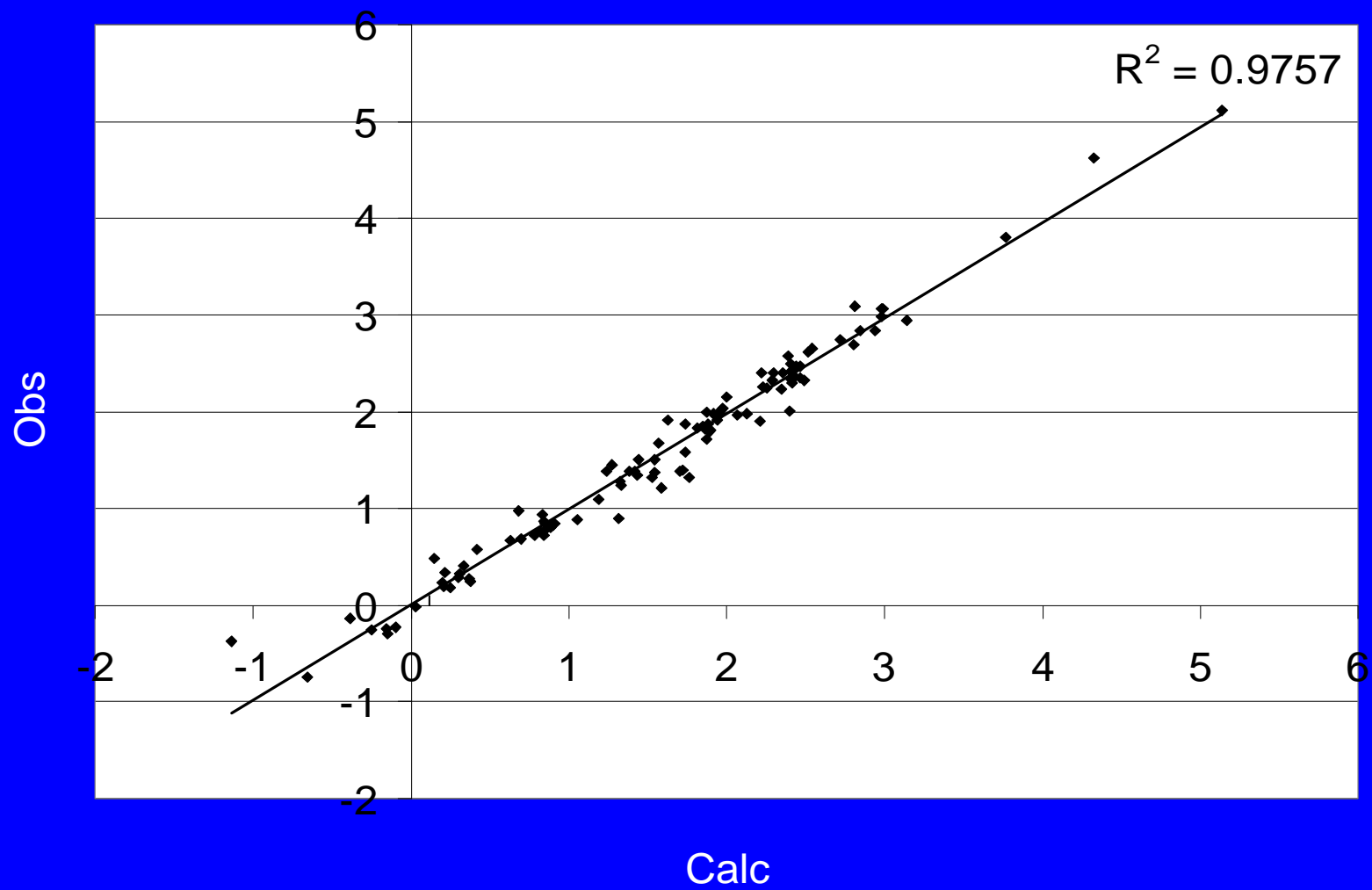
Example: octanol/water partition

$$\log P_{\text{oct}} = 0.09 + 0.56E - 1.05S + 0.03A - 3.46B + 3.81V_x$$

Cavities more easily formed in octanol



Validation for 104  $\log P_{\text{oct}}$  values:



# Solvation equation

Example: blood-brain distribution

$$\log BB = -0.04 + 0.20E - 0.69S - 0.72A - 0.70B + 0.99V_x$$

$$n=57, \quad r^2=0.907, \quad \text{s.d.}=0.197$$

# Solvation equation

Example: blood-brain distribution

$$\log BB = -0.04 + 0.20E - 0.69S - 0.72A - 0.70B + 0.99V_x$$

blood more polar, basic and acidic than brain



# Solvation equation

Example: blood-brain distribution

$$\log BB = -0.04 + 0.20E - 0.69S - 0.72A - 0.70B + 0.99V_x$$

Cavities more easily formed in brain.



# Solvation descriptors

## Derivation of Descriptors

- E and  $V_x$  simply calculated from atom/bond contributions

# Solvation descriptors

## Derivation of Descriptors

- E and  $V_x$  simply calculated from atom/bond contributions
- S, A, and B must be manually summed from fragments

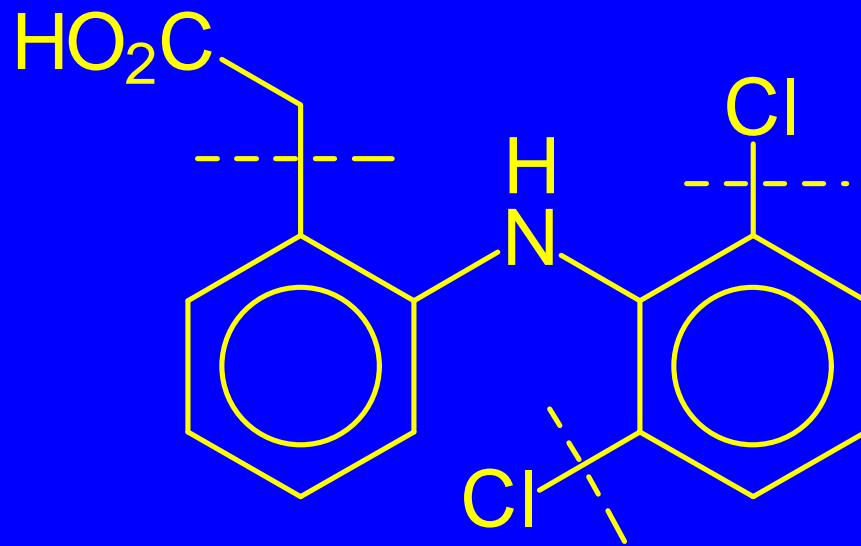
# Solvation descriptors

## Derivation of Descriptors

- E and  $V_x$  simply calculated from atom/bond contributions
- S, A, and B must be manually summed from fragments
- Comparison with measured logP's useful

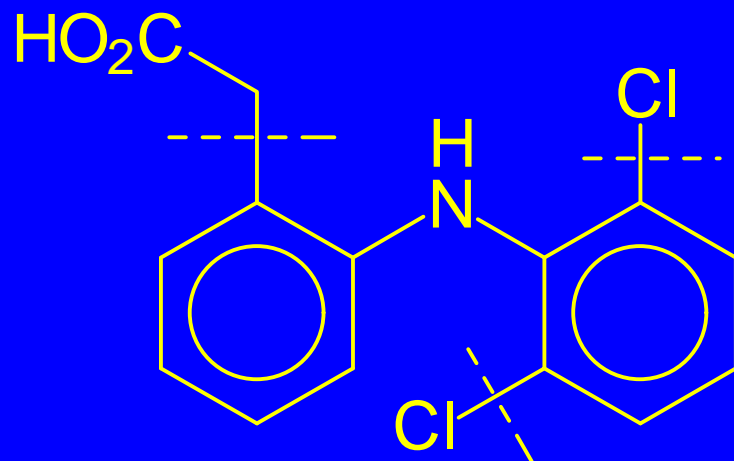
# Solvation descriptors

Example: voltarin



# Solvation descriptors

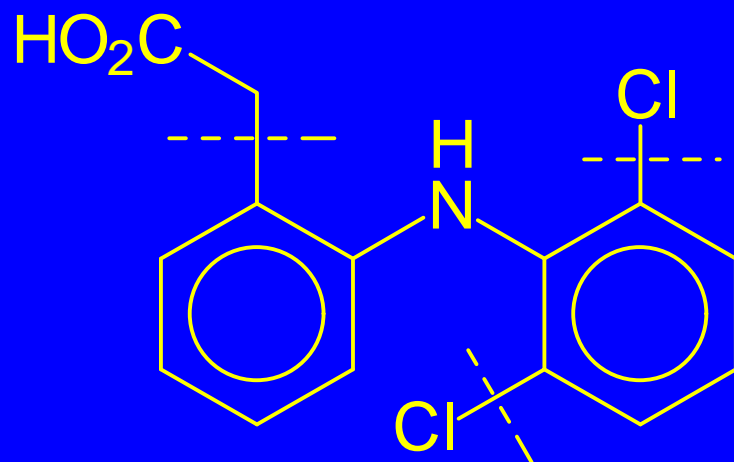
Example: voltarin



<u>Frag</u>	<u>S</u>	<u>A</u>	<u>B</u>
PhNHPh	0.88	0.10	0.57
EtCO <sub>2</sub> H	0.65	0.60	0.45
2xCl	0.26	0.00	-0.14
<u>Total</u>	<u>1.79</u>	<u>0.70</u>	<u>0.88</u>

# Solvation descriptors

Example: voltarin



Fragment	S	A	B
PhNHPh	0.88	0.10	0.57
EtCO <sub>2</sub> H	0.65	0.60	0.45
2xCl	0.26	0.00	-0.14
Total	1.79	0.70	0.88

$V_x = 2.025$  gives  $\log P_{\text{oct}}$  of 4.05

*cf* experimental value of 4.40.

# Solvation descriptors

## Derivation of descriptors

- Manual method, with experimental input, is accurate

# Solvation descriptors

## Derivation of descriptors

- Manual method, with experimental input, is accurate
- Fragmentation, look-up of fragments, comparison with logP is slow

# Solvation descriptors

## Derivation of descriptors

- Manual method, with experimental input, is accurate
- Fragmentation, look-up of fragments, comparison with logP is slow
- Typically, descriptors for 10-20 molecules per day

# Solvation descriptors

## Calculation of descriptors

- Need identified for automated procedure for derivation of descriptors

# Solvation descriptors

## Calculation of descriptors

- Need identified for automated procedure for derivation of descriptors
- Must be fast - thousands/millions per day for “high-throughput screening”

# Solvation descriptors

## Calculation of descriptors

- Need identified for automated procedure for derivation of descriptors
- Must be fast - thousands/millions per day for “high-throughput screening”
- Must be general - applicable to most organic compounds

# Automated Calculation

## Empirical observations

- Molecular descriptors known to be additive

# Automated Calculation

## Empirical observations

- Molecular descriptors known to be additive
- Functional groups values approximately constant in similar environments

# Automated Calculation

## Empirical observations

- Molecular descriptors known to be additive
- Functional groups values approximately constant in similar environments
- Intramolecular interactions important

# Automated Calculation

## Group contribution approach

- Start with 33 atom types from Klopman solubility model - designed for generality

# Automated Calculation

## Group contribution approach

- Start with 33 atom types from Klopman solubility model - designed for generality
- Define types as Daylight SMARTS

# Automated Calculation

## Group contribution approach

- Start with 33 atom types from Klopman solubility model - designed for generality
- Define types as Daylight SMARTS
- Regress counts against database descriptor values for  $\approx$  3000 molecules

# Automated Calculation

## Group contribution approach

- Improvements made by splitting atom types  
*e.g.* -NH<sub>2</sub> into amine and aniline

# Automated Calculation

## Group contribution approach

- Improvements made by splitting atom types  
*e.g.* -NH<sub>2</sub> into amine and aniline
- Added corrections for common groups *e.g.*  
SO<sub>2</sub>NH<sub>2</sub> ≠ SO<sub>2</sub> + NH<sub>2</sub>

# Automated Calculation

## Group contribution approach

- Improvements made by splitting atom types  
*e.g.* -NH<sub>2</sub> into amine and aniline
- Added corrections for common groups *e.g.*  
SO<sub>2</sub>NH<sub>2</sub> ≠ SO<sub>2</sub> + NH<sub>2</sub>
- Intramolecular interactions identified -  
hydrogen bonds and heterocycles

# Automated Calculation

## Group contribution approach

- Currently 65 fragments defined as SMARTS

# Automated Calculation

## Group contribution approach

- Currently 65 fragments defined as SMARTS
- Applies to E, S, and B

# Automated Calculation

## Group contribution approach

- Currently 65 fragments defined as SMARTS
- Applies to E, S, and B
- Klopman set not suitable for A - defined entirely new set of 45 acidic fragments

# Automated Calculation

## Group contribution approach

- Fast:  $\approx 600$  molecule per minute on standard workstation

# Automated Calculation

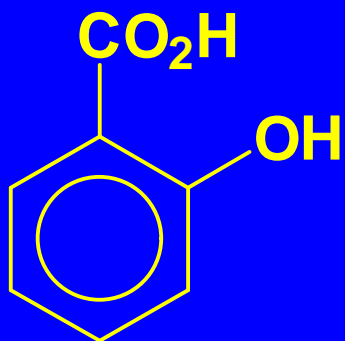
## Group contribution approach

- Fast:  $\approx 600$  molecule per minute on standard workstation
- Calculation of many solvation properties trivial once descriptors calculated

# Automated Calculation

Example

Salicylic acid



A	
Al-OH	0.35
Ar-OH	0.54
CO <sub>2</sub> H	0.24
H-bond	-0.38
<hr/>	
Total	0.75

B	
4xCH	0.04
3xC	0.00
2xOH	0.61
1x =O	0.37
1xCO <sub>2</sub> H	-0.31
1xHbond4	-0.38
<hr/>	
Total	0.39

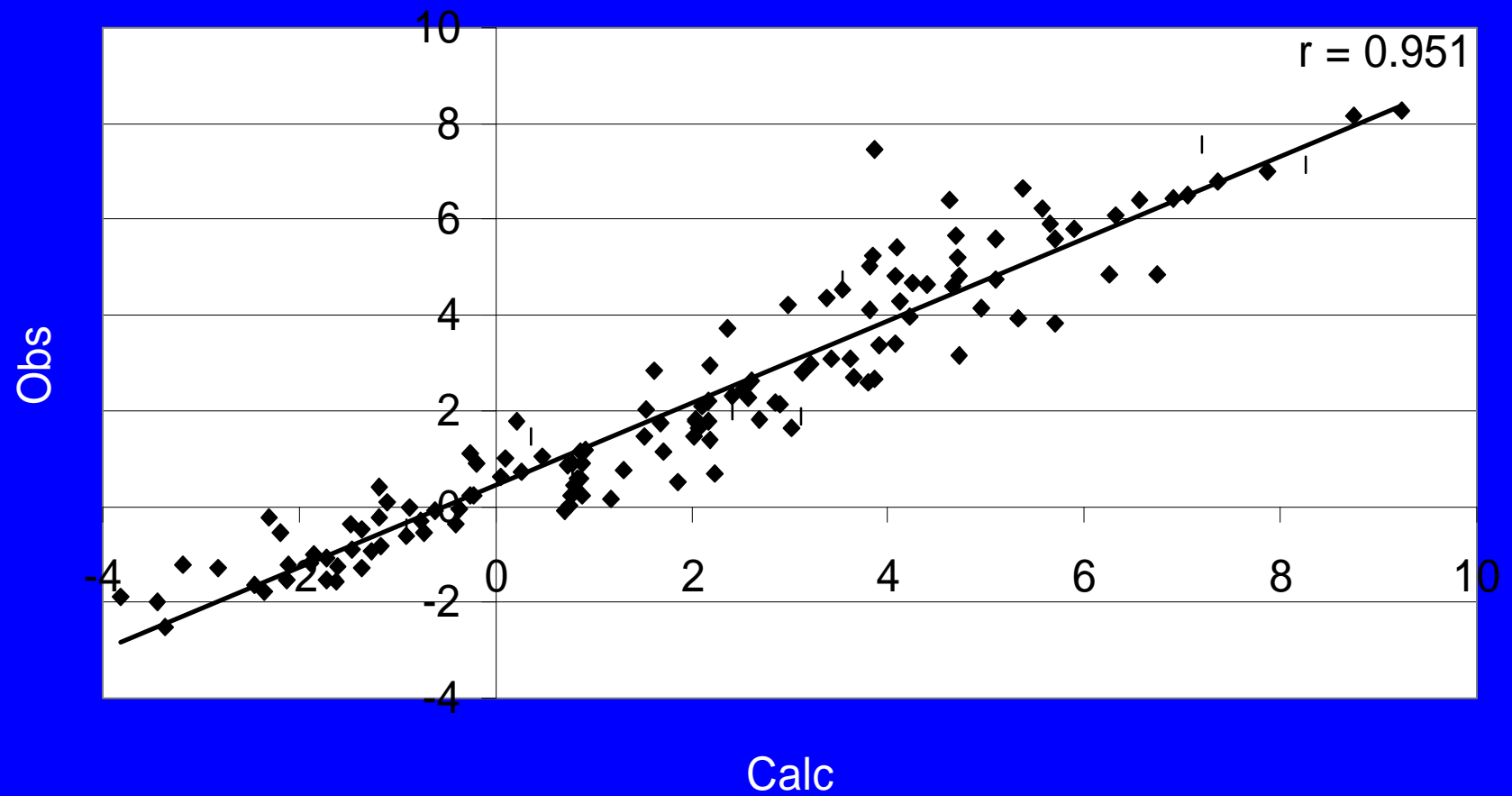
# Automated Calculation

Calculation accuracy

	$R^2$	sd
E	0.978	0.093
S	0.922	0.164
A	0.945	0.058
B	0.910	0.121

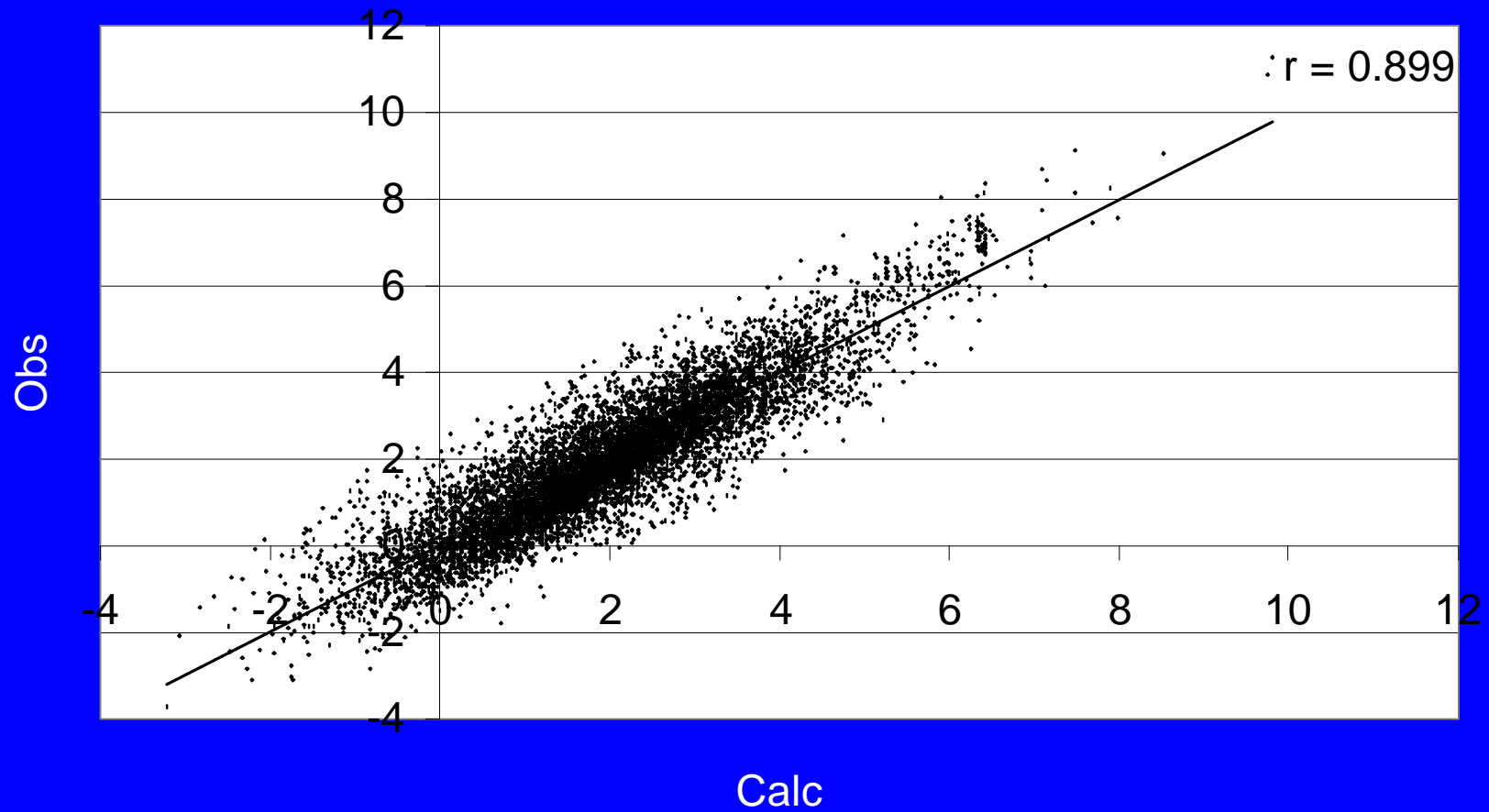
# Automated Calculation

Bodor's 137  $\log P_{\text{oct}}$  Values



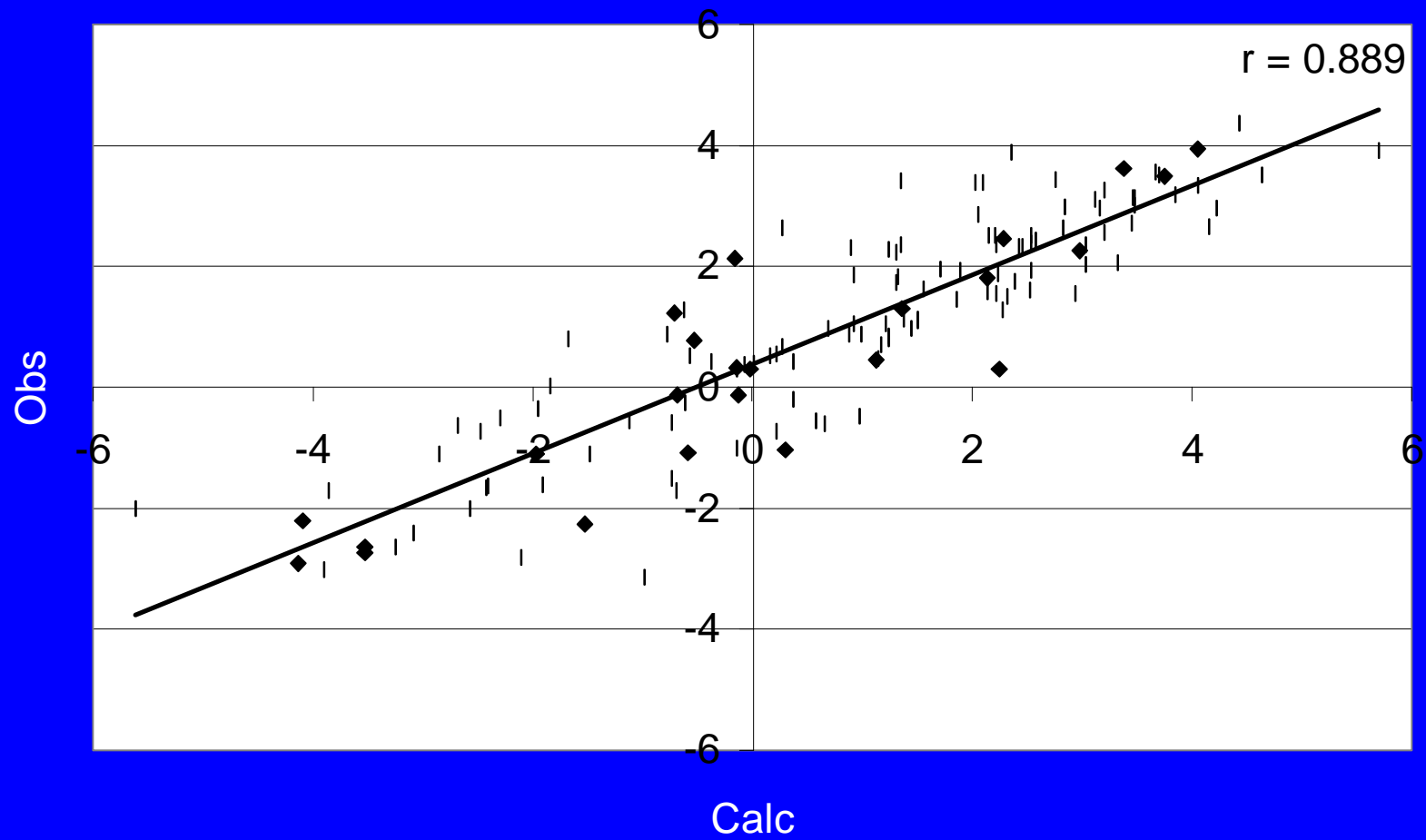
# Automated Calculation

New regression for 8943 logP\* Values



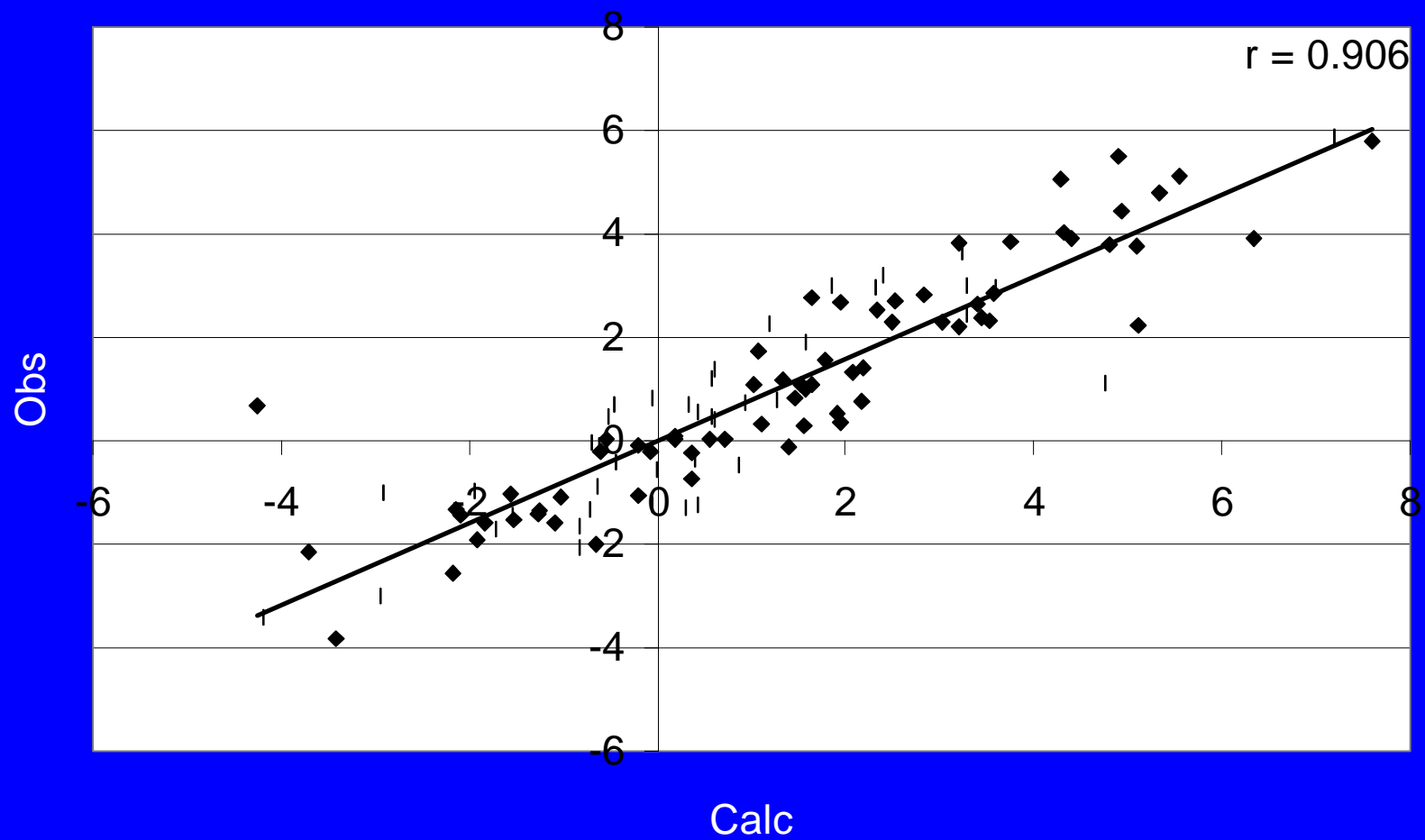
# Automated Calculation

135  $\log P_{\text{cyc}}$  values



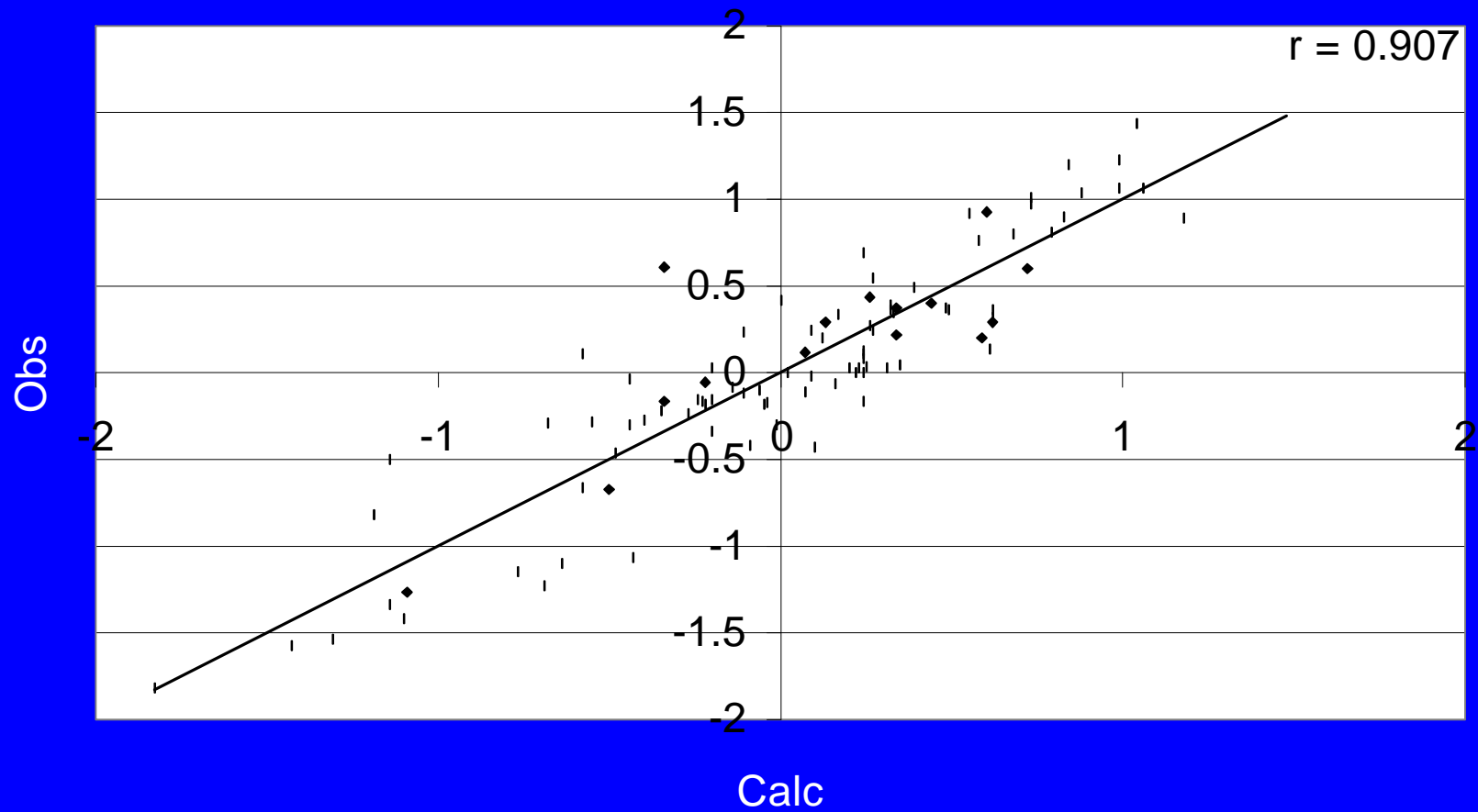
# Automated Calculation

109  $\log P_{chl}$  values



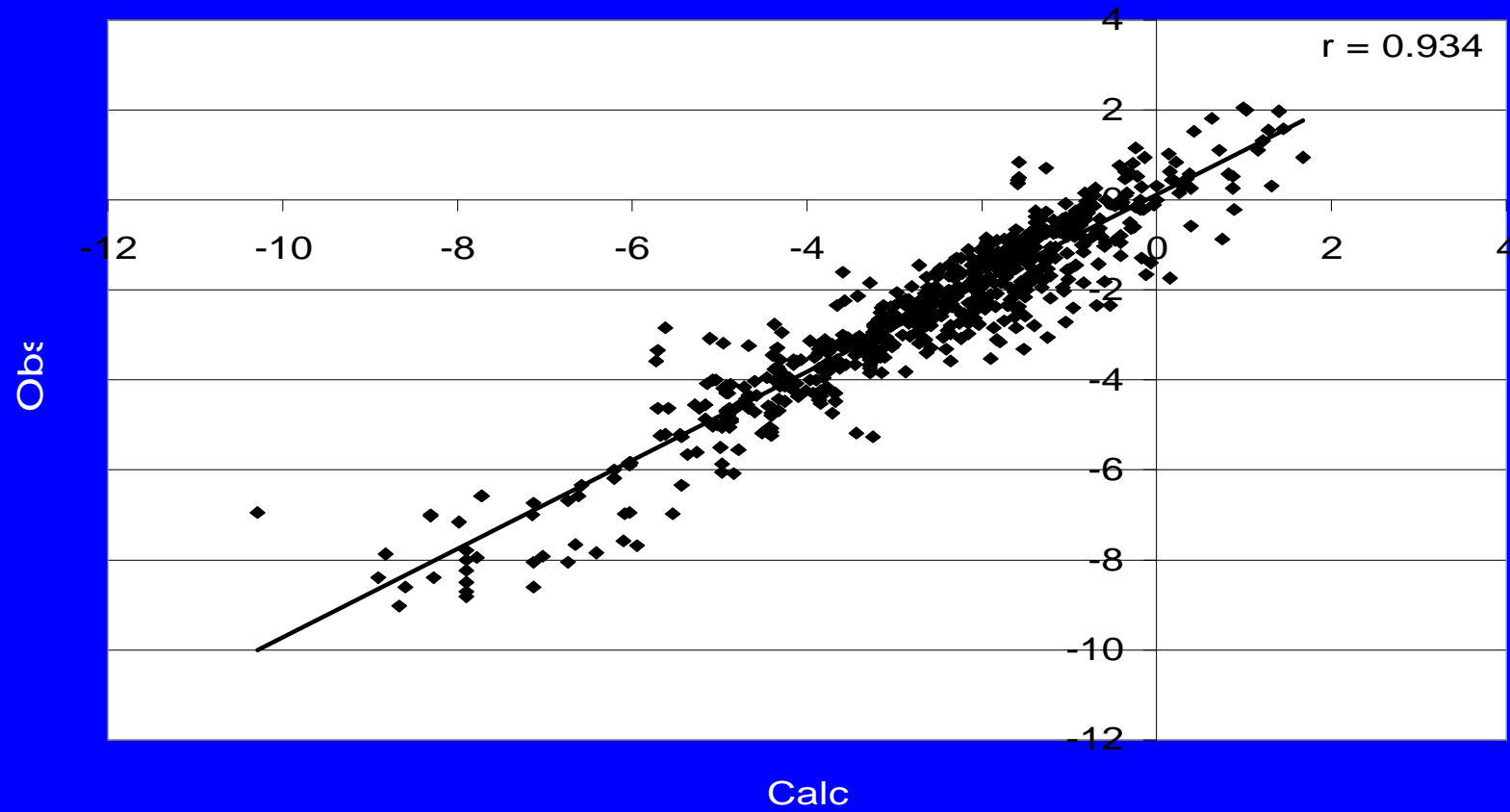
# Automated Calculation

97 logBB Values



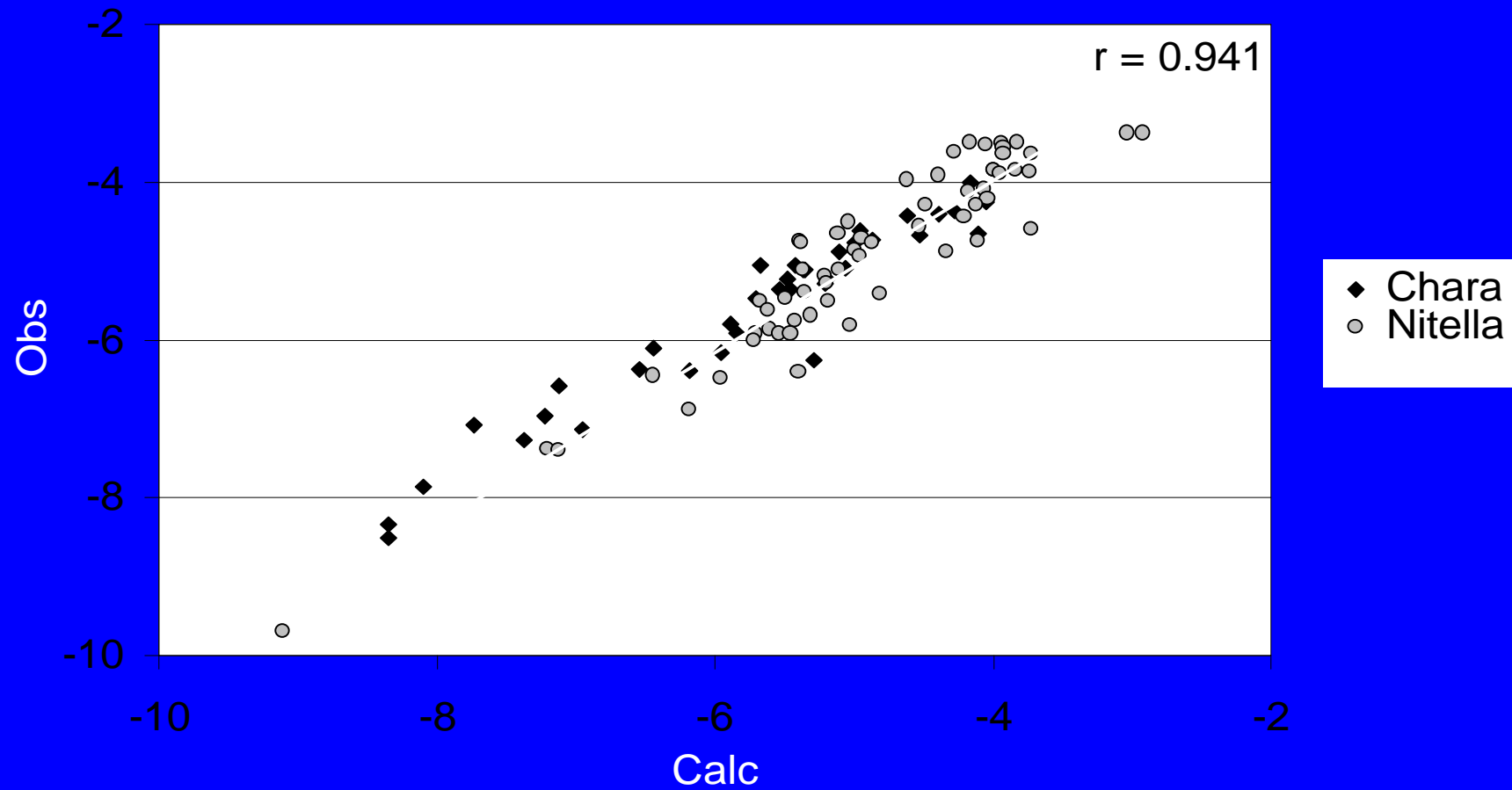
# Automated Calculation

670 logSw values



# Automated Calculation

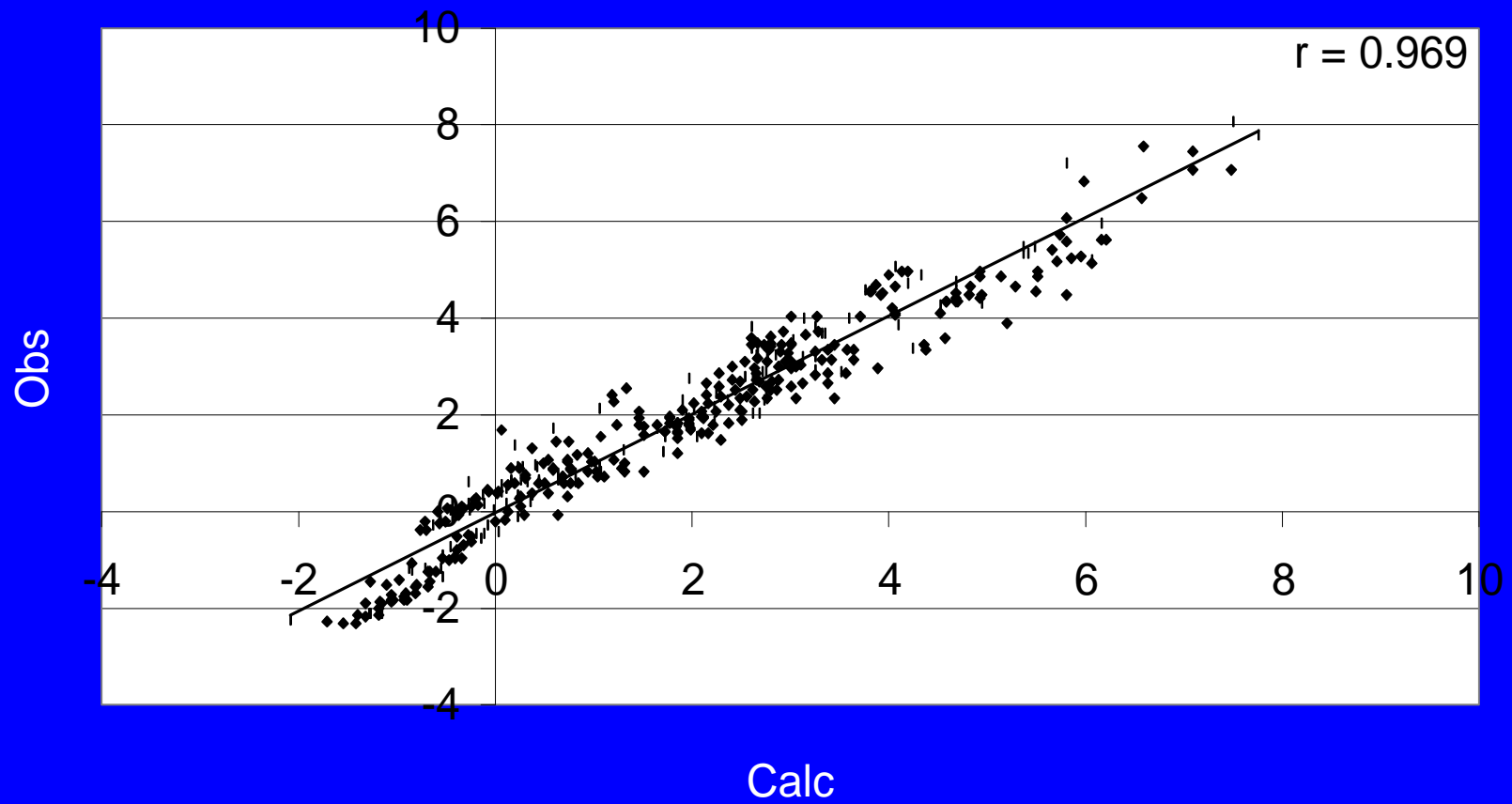
98 Plant Cell Permeation Data





# Automated Calculation

405 Air/Water logL Data



# Correlation of Descriptors

5 Descriptors too many?

- Correlation between descriptors can be a problem

# Correlation of Descriptors

## 5 Descriptors too many?

- Correlation between descriptors can be a problem
- logP\* set: E vs S  $r = 0.77$ , B vs V<sub>x</sub> 0.73

# Correlation of Descriptors

## 5 Descriptors too many?

- Correlation between descriptors can be a problem
- logP\* set: E vs S  $r = 0.77$ , B vs V<sub>x</sub> 0.73
- logBB set: E vs S  $r = 0.96$ , S vs  $\beta$  0.93

# Correlation of Descriptors

## 5 Descriptors too many?

- Correlation between descriptors can be a problem
- logP\* set: E vs S  $r = 0.77$ , B vs V<sub>x</sub> 0.73
- logBB set: E vs S 0.96, S vs  $\beta$  0.93
- Principal components to remove correlations and reduce number of descriptors

# Correlation of Descriptors

logP\* set

$$\log P(\text{oct}) = 0.33 + 0.92E - 0.83S - 0.26A - 2.49B + 2.66V_x$$

$$n = 8928, R^2 = 0.814, R^2_{\text{CV}} = 0.813, \text{sd} = 0.70, F = 7785$$

# Correlation of Descriptors

logP\* set

$$\log P(\text{oct}) = 0.33 + 0.92E - 0.83S - 0.26A - 2.49B + 2.66V_x$$

$$n = 8928, R^2 = 0.814, R^2_{CV} = 0.813, sd = 0.70, F = 7785$$

*c.f.*

$$\log P(\text{oct}) = 1.95 - 0.82PC_2 + 0.49PC_3 + 2.53PC_4 - 0.33PC_5$$

$$n = 8928, R^2 = 0.814, R^2_{CV} = 0.813, sd = 0.70, F = 9731$$

# Correlation of Descriptors

logBB set

$$\log\text{BB} = 0.06 + 0.80E - 1.45S - 0.77A - 0.16B + 0.75V_x - 0.37I_1$$

$$n = 97, R^2 = 0.822, R^2_{\text{CV}} = 0.783, \text{sd} = 0.286, F = 69$$

# Correlation of Descriptors

logBB set

$$\log\text{BB} = 0.06 + 0.80E - 1.45S - 0.77A - 0.16B + 0.75V_x - 0.37I_1$$

$$n = 97, R^2 = 0.822, R^2_{\text{CV}} = 0.783, \text{sd} = 0.286, F = 69$$

*c.f.*

$$\log\text{BB} = 0.06 - 0.15\text{PC}_1 - 0.43\text{PC}_2 + 0.80\text{PC}_3 + 1.49\text{PC}_5 - 0.37I_1$$

$$n = 97, R^2 = 0.822, R^2_{\text{CV}} = 0.790, \text{sd} = 0.286, F = 84$$

# Correlation of Descriptors

## PCA regressions

- Typically PCA reduces 5 LFER descriptors to 4

# Correlation of Descriptors

## PCA regressions

- Typically PCA reduces 5 LFER descriptors to 4
- Any advantage balanced against loss of interpretative value

# Correlation of Descriptors

## PCA regressions

- Typically PCA reduces 5 LFER descriptors to 4
- Any advantage balanced against loss of interpretative value
- 5 descriptors OK - important to design dataset well

# Automated Calculation

## Some practicalities

- Error trapping for some badly predicted molecules, *e.g.* missed atoms, zwitterions

# Automated Calculation

## Some practicalities

- Error trapping for some badly predicted molecules, *e.g.* missed atoms, zwitterions
- Input as Daylight SMILES required

# Automated Calculation

## Some practicalities

- Error trapping for some badly predicted molecules, *e.g.* missed atoms, zwitterions
- Input as Daylight SMILES required
- Web front-end and name look-up developed by Darko Butina

# Automated Calculation

## Future development

- New values for poor/missing fragments - many already identified

# Automated Calculation

## Future development

- New values for poor/missing fragments - many already identified
- Errors for individual calculations

# Automated Calculation

## Future development

- New values for poor/missing fragments - many already identified
- Errors for individual calculations
- Incorporation of pKa prediction and ionisation effects

# Automated Calculation

## Future development

- Equations for important processes - GI absorption, brain penetration

# Automated Calculation

## Future development

- Equations for important processes - GI absorption, brain penetration
- Better integration with MHA database - to be marketed as user-friendly package **Absolv** by Sirius Analytical

# Automated Calculation

## Future development

- Prediction of fragment values from structure

# Automated Calculation

## Future development

- Prediction of fragment values from structure
- MO calculations, MD & Monte Carlo simulations

# Acknowledgements

- Mike Abraham + group - UCL
- Anne Hersey, Darko Butina - GW Ware
- Lynne Trowbridge, John Comer, Tim Aitken - Sirius Analytical
- Chau Du, Klara Valko, John Bradshaw, Derek Reynolds - GW Stevenage