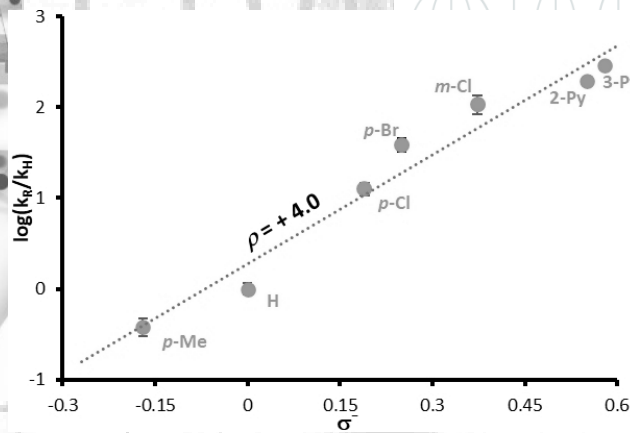
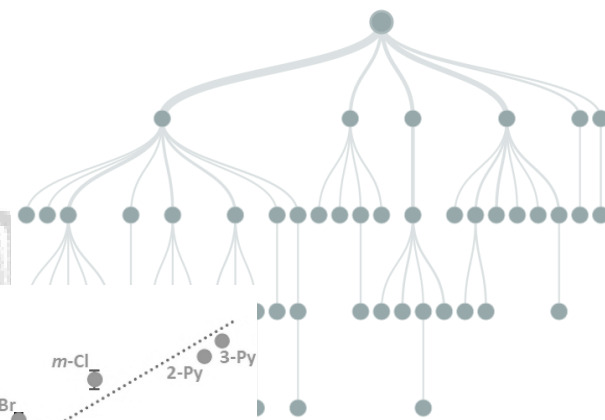
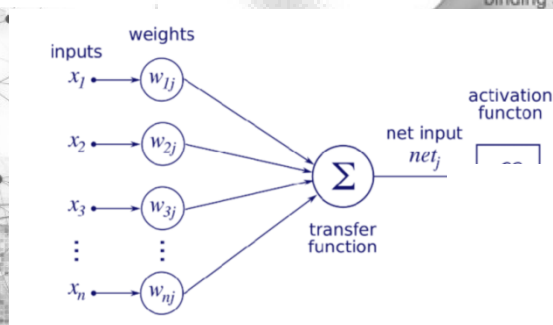
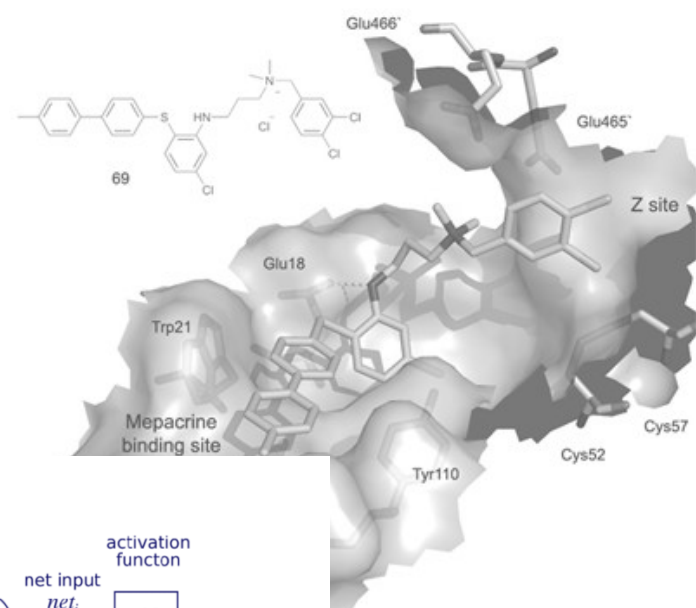


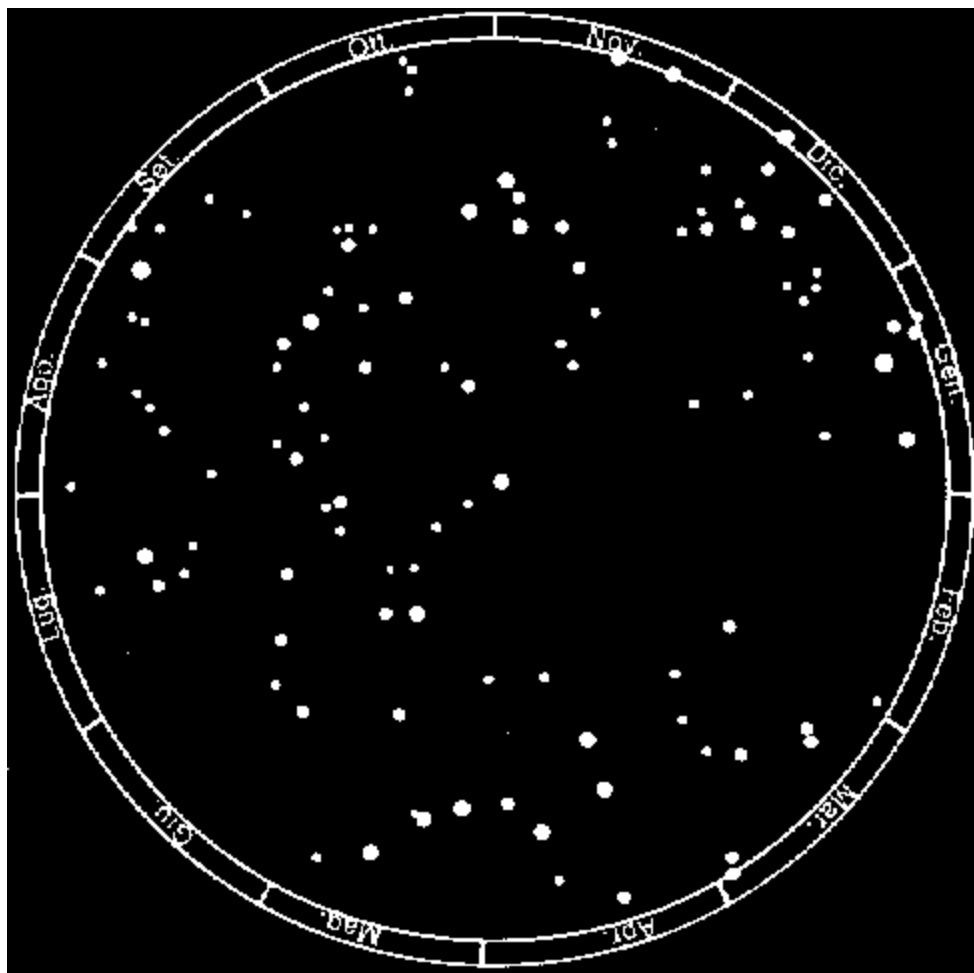
Introduzione ai metodi di Intelligenza Artificiale...



Stefano

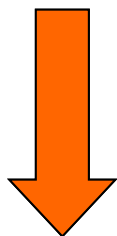
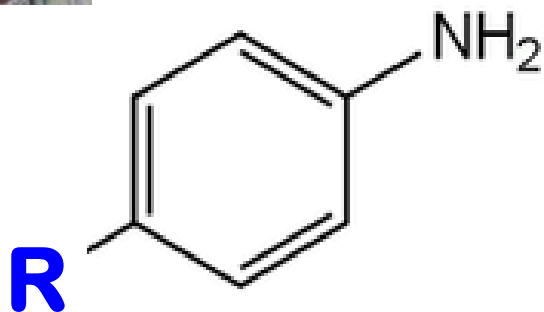


(Q)SAR and surroundings...





The easiest way to describe the concept of (Q)SAR...



Chemicals

$R(1) = \text{NH}_2$

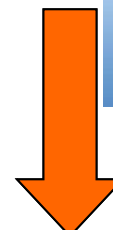
$R(2) = \text{NHCH}_3$

$R(3) = \text{N}(\text{CH}_3)_2$

...

$R(n) = \text{NHC}_6\text{H}_5$

f(x)



Activities (LC_{50} , μM)

A_1

A_2

A_3

...

A_n



or alternatively...

Activity Space

Chemical Space

	E.P. (A)	E.P. (B)	E. P. (C)	E.P. (D)	E.P. (E)
Comp.1	LC ₁ (A)				
Comp.2	LC ₂ (A)				
Comp.3	LC ₃ (A)				
Comp.4	LC ₄ (A)				
Comp.n	LC _n (A)				

Screening for the specific endpoint



We will discuss later about the dimension of this table ($n \times m$)!!!

Activity Space

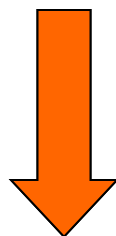
Chemical Space

	E.P. (A)	E.P. (B)	E. P. (C)	E.P. (D)	E.P. (E)
Comp.1	LC ₁ (A)				
Comp.2	LC ₂ (A)				
Comp.3	LC ₃ (A)				
Comp.4	LC ₄ (A)				
Comp.n	LC _n (A)				

Screening for the specific endpoint



(Q)SAR it is not necessary to remember this definition:



Activities (IC_{50} , μM)

Any in vivo or in vitro data is affected by both PHARMACODYNAMIC and PHARMACOKINETICS properties of the specific assay.

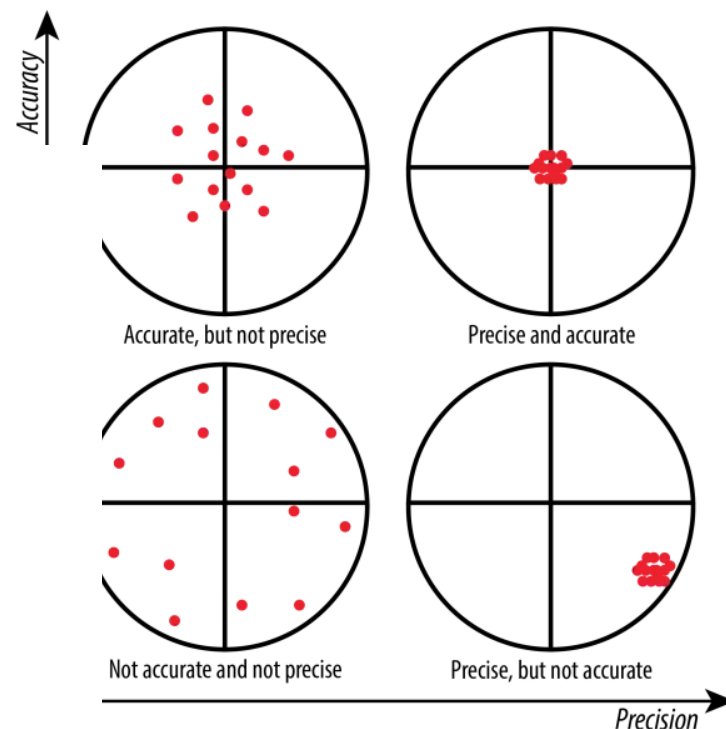
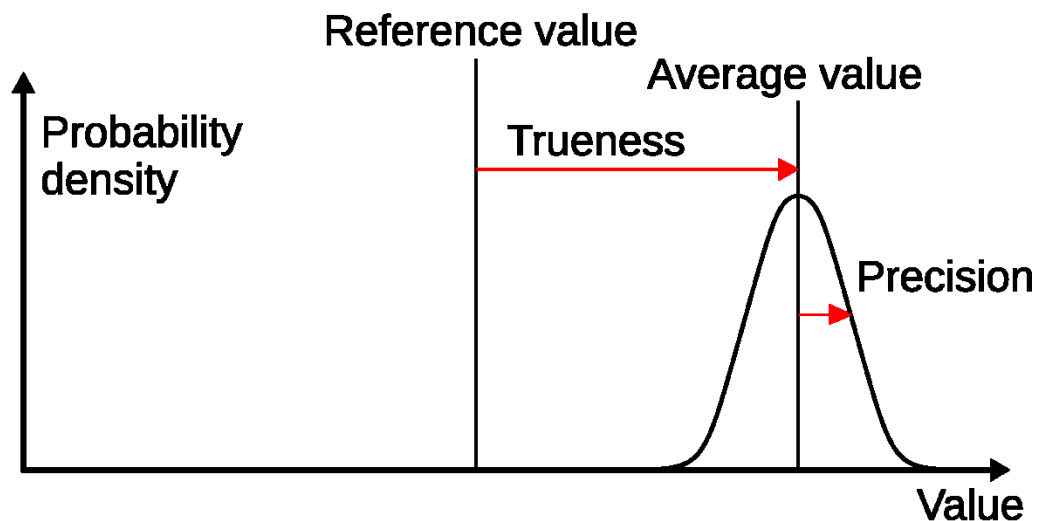


Please: don't start any (quantitative) structure-activity relationship if you're not confident of *quality* of your data activity!!!

$LC_{50} =$	1.0	$\mu M \neq$	
	1.0 ± 1.0	$\mu M \neq$	
	1.0 ± 0.5	$\mu M \neq$	
	1.0 ± 0.1	$\mu M \neq$	
	1.0 ± 0.05	μM	



Please: don't start any (quantitative) structure-activity relationship if you're not confident of *quality* of your data activity!!!

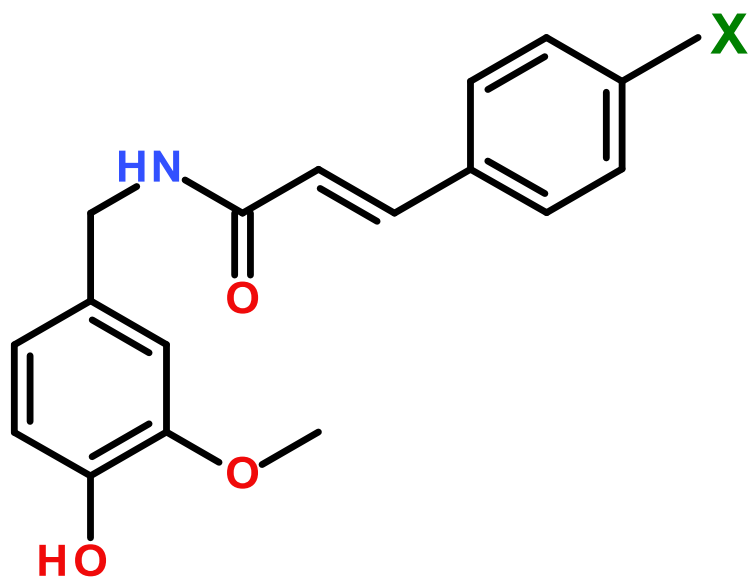


Credits: https://en.wikipedia.org/wiki/Accuracy_and_precision



We can start we are real example:

The work reported from The Sandoz Institute for Medical Research on the development of novel analgesic agents (1) can be used as an example of a simple QSAR. In this study, vanillylamides and vanillylthioureas related to capsaicin were prepared and their activity was tested in an in vitro assay which measured $^{45}\text{Ca}^{2+}$ influx into dorsal root ganglia neurons. The data, which was reported as the $\text{EC}_{50}(\mu\text{M})$, is shown in Table 1 (**note that compound 6f is the most active of the series**).



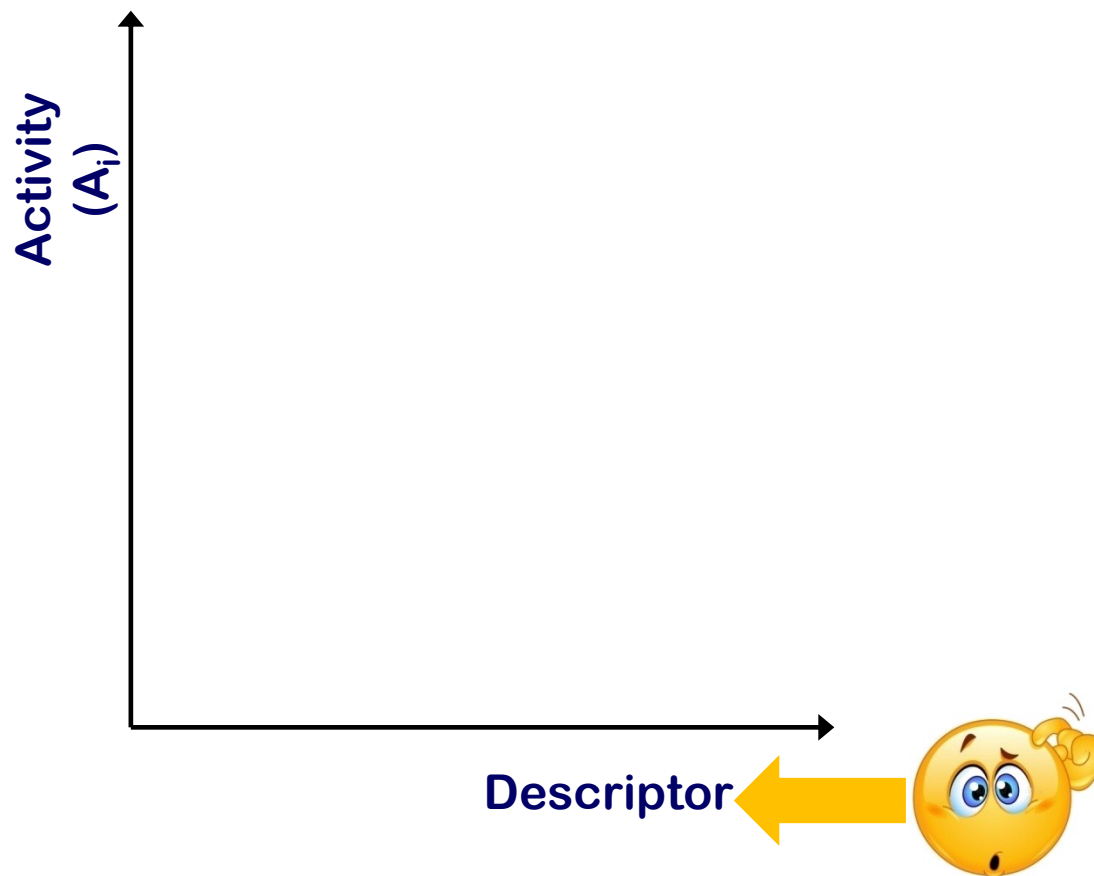
Cmpd number	Cmpd name	X	EC_{50} (μM)
1	6a	H	11.80 ± 1.90
2	6b	Cl	1.24 ± 0.11
3	6d	NO_2	4.58 ± 0.29
4	6e	CN	26.50 ± 5.87
5	6f	C_6H_5	0.24 ± 0.30
6	6g	$\text{N}(\text{CH}_3)_2$	4.39 ± 0.67
7	6h	I	0.35 ± 0.05
8	6i	NHCHO	?



1. Christopher S.J. Walpole, Roger Wigglerworth, Stuart Bevan, Elizabeth A. Campbell, Andy Dray, Iain F. James, Kay J. Masdin, Martin N. Perkins and Janet Winter, J. Med. Chem., 36, 2381 (1993).



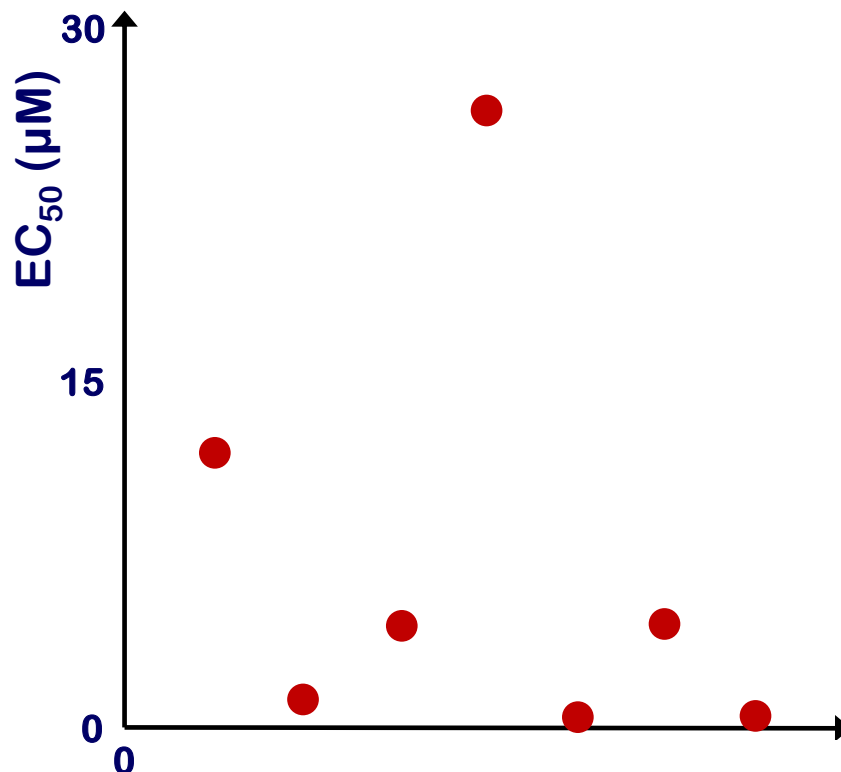
(Q)SAR: follow me in this wonderful experience



**Scatterplot... an interesting place where
scouting for **patterns!!!****



Starting from our table, what is the first descriptor that we can use?

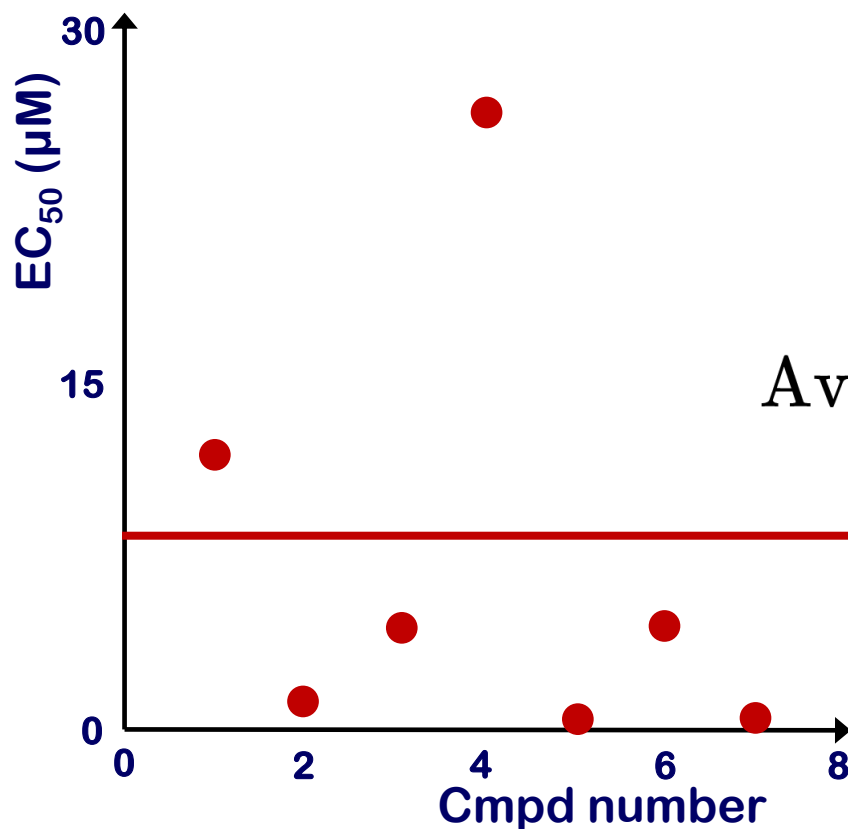


What is your feeling?





Starting from our table, what is the first descriptor that we can use?

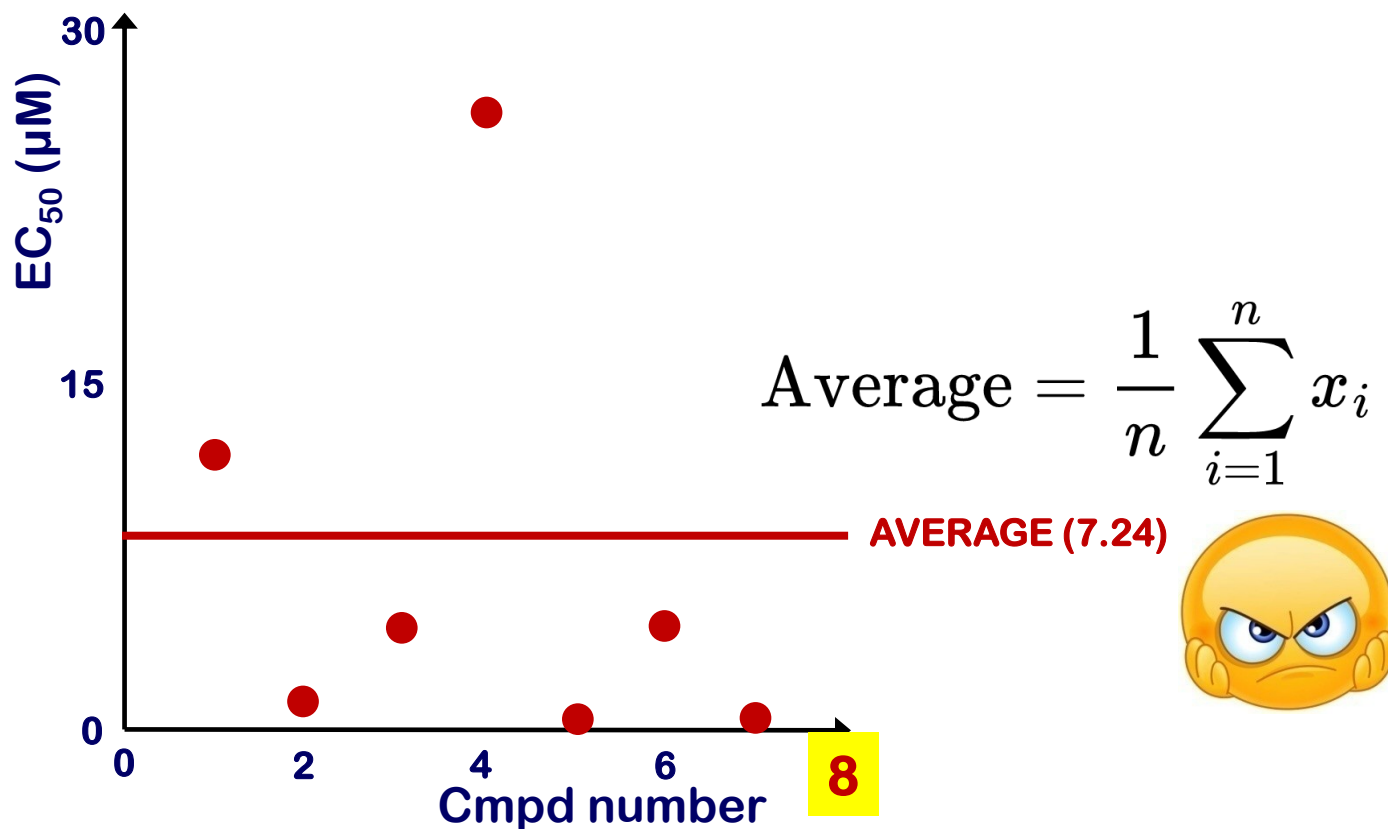


$$\text{Average} = \frac{1}{n} \sum_{i=1}^n x_i$$

Well, what we can do now... we can calculate the **average**, of course!



How useful could be the average value to accurately predict the EC₅₀ of cmpd 8?

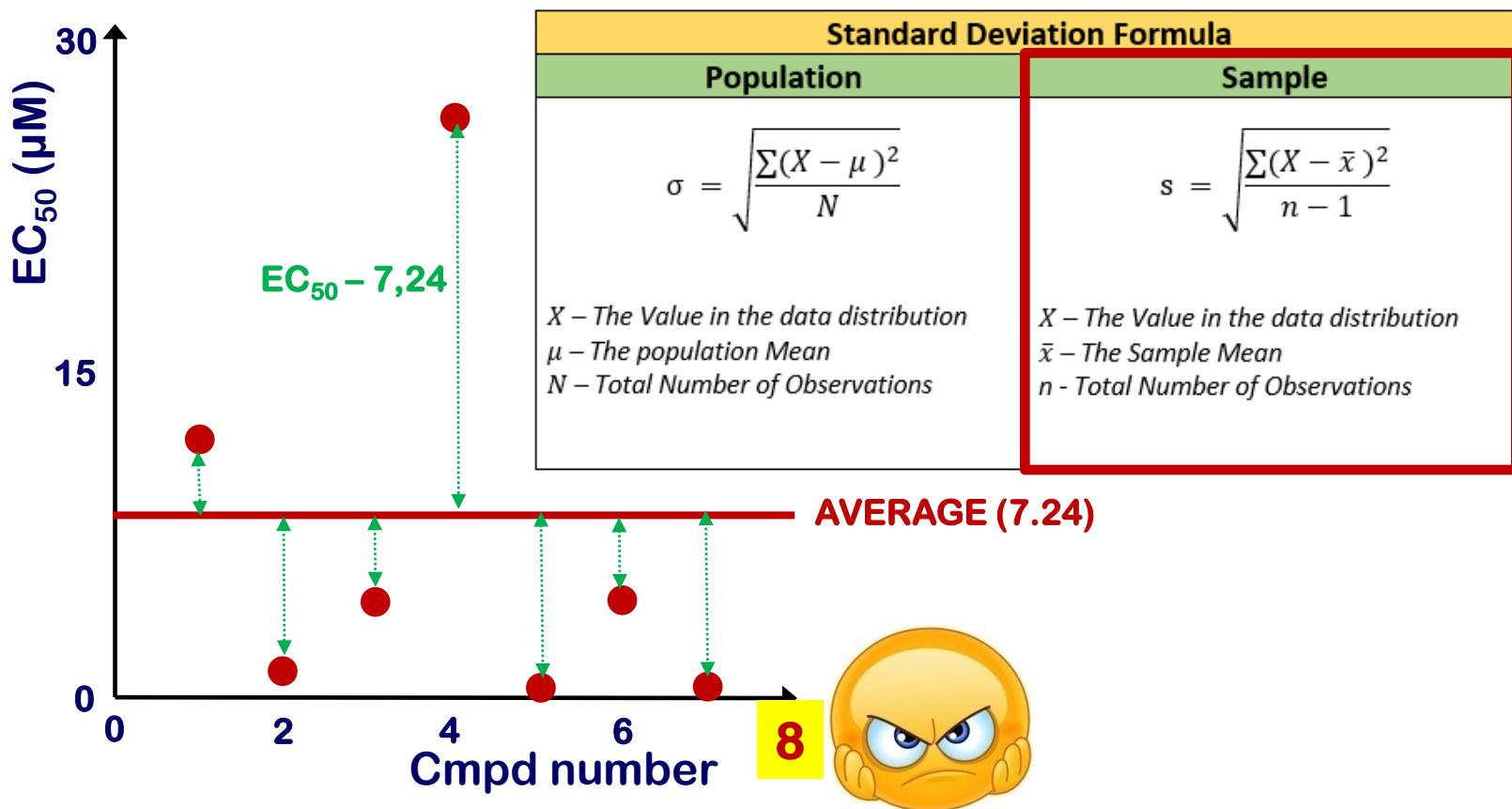


The average, 7.24, provides a guess for the value of compound 8 but, how good is this guess?



Do you remember the **standard deviation** concept? Just a very quick refresh!!!

Credits: https://en.wikipedia.org/wiki/Standard_deviation



The standard deviation of the data, *s*, shows how far the activity values are spread about their average. This value provides an indication of the quality of the guess by showing the amount of variability inherent in the data. **In our case 9.48!!!**



Do you remember the standard deviation concept? Just a very quick refresh!!!

Credits: https://en.wikipedia.org/wiki/Standard_deviation

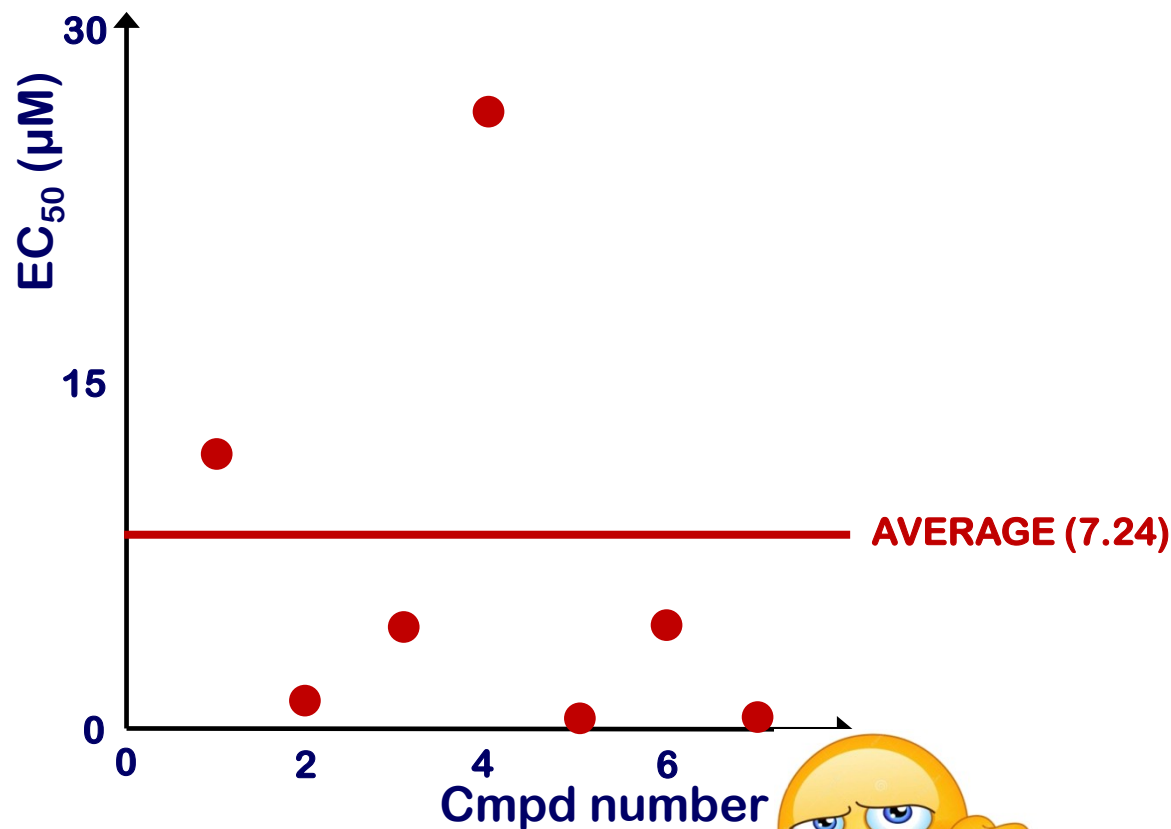
Standard Deviation Formula	
Population	Sample
$\sigma = \sqrt{\frac{\sum (X - \mu)^2}{N}}$ <i>X – The Value in the data distribution</i> <i>μ – The population Mean</i> <i>N – Total Number of Observations</i>	$s = \sqrt{\frac{\sum (X - \bar{x})^2}{n - 1}}$ <i>X – The Value in the data distribution</i> <i>$\bar{x}$ – The Sample Mean</i> <i>n – Total Number of Observations</i>



Why n – 1? The n-1 (*called Bessel's correction*) in the sample variance formula is a correction factor to adjust the result because the sample mean tends to underestimate the population mean.



We must change the nature of the descriptor but how?



Back to 1964...



Spencer M. Free and James W. Wilson: A Mathematical Contribution to Structure Activity Studies

Journal of Medicinal Chemistry

© Copyright 1964 by the American Chemical Society

VOLUME 7, NUMBER 4

JULY 6, 1964

A Mathematical Contribution to Structure-Activity Studies

SPENCER M. FREE, JR., AND JAMES W. WILSON

Research and Development Division, Smith Kline and French Laboratories, Philadelphia, Pennsylvania

Received February 4, 1964

A mathematical technique is suggested as a means of describing structure-activity relationships of a series of chemical analogs. The data requirements included specific side chain arrangements and performance characteristics of all analogs tested. Two examples illustrate the use of the additive mathematical model where the performance characteristics are measures of biological activity. The results rank the structural changes per position by estimating the amount of biological response attributed to each change. The estimates are both positive and negative. Several uses for the mathematical solution are suggested.

Credits: <https://pubs.acs.org/doi/10.1021/jm00334a001>



Spencer M. **Free** and James W. **Wilson**: A Mathematical Contribution to Structure Activity Studies.

Organic chemists who study analog series relate differences in structure to performance characteristics of each compound. The introduction of electronic computers offers opportunities to enhance this effort. Information retrieval techniques provide chemists with selected lists of analogs associated with biological response data. When such data have not been developed through first hand experience, and/or a large number of compounds are available, the determination of structure-activity relationships is more difficult. When the available data meet a limited number of restrictions, mathematical techniques can supplement the organic chemist's intuition.

Credits: <https://pubs.acs.org/doi/10.1021/jm00334a001>



Spencer M. Free and James W. Wilson: A Mathematical Contribution to Structure Activity Studies.

The example is presented to illustrate the additive property of the analogs. The mathematical models described in this paper will be based upon the assumption that there is some such “additivity” in series of analogs.

Continuing with the example to illustrate the mathematical model one writes a formula as follows

response = average + effect of R_1 substituent + effect of R_2 substituent

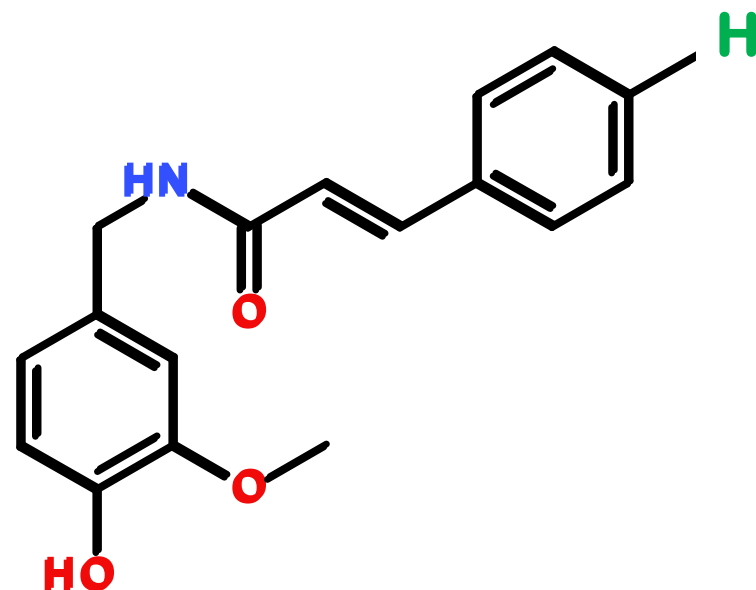
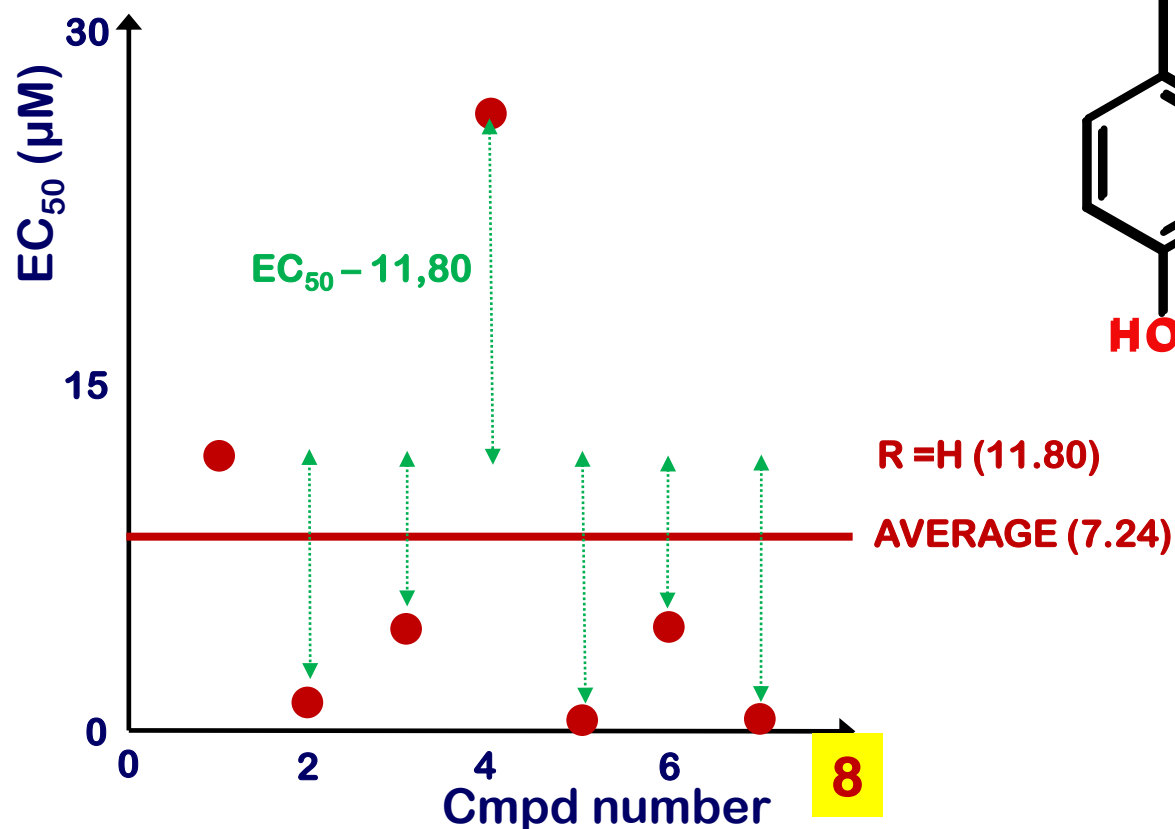


Credits: <https://pubs.acs.org/doi/10.1021/jm00334a001>



Spencer M. **Free** and James W. **Wilson**: A Mathematical Contribution to Structure Activity Studies.

How it Works:





Spencer M. **Free** and James W. **Wilson**: A Mathematical Contribution to Structure Activity Studies.

How it Works: $\Delta EC_{50} : H \Rightarrow R$

Cmpd number	Cmpd name	X	EC ₅₀ (μM)	ΔEC ₅₀ (μM)
1	6a	H	11.80 ± 1.90	0
2	6b	Cl	1.24 ± 0.11	- 10.56
3	6d	NO ₂	4.58 ± 0.29	- 7.22
4	6e	CN	26.50 ± 5.87	14.7
5	6f	C ₆ H ₅	0.24 ± 0.30	- 11.56
6	6g	N(CH ₃) ₂	4.39 ± 0.67	- 7.41
7	6h	I	0.35 ± 0.05	- 11.45
8	6i	NHCHO	???	???



Spencer M. **Free** and James W. **Wilson**: A Mathematical Contribution to Structure Activity Studies.

How we can use this ΔEC_{50}

a) Unfortunately, we cannot estimate the activity of compound **8**!!!



b) We assume that the biological activity of a molecule can be represented as follow:

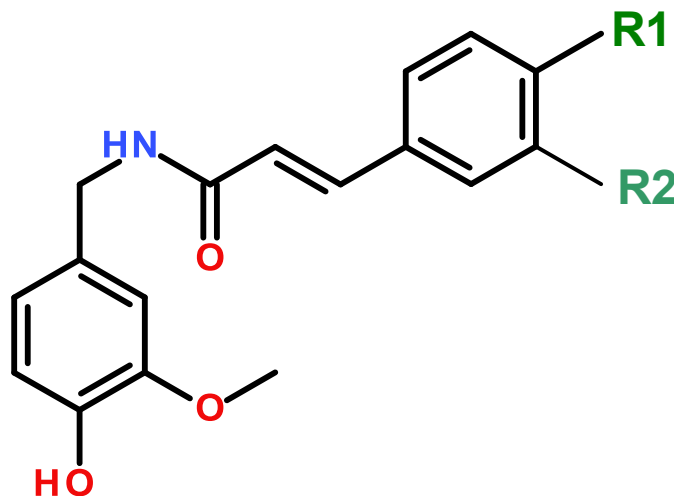
$$EC_{50} (R) = EC_{50} (H) + \Delta EC_{50} (R)$$



Spencer M. **Free** and James W. **Wilson**: A Mathematical Contribution to Structure Activity Studies.

How we can use this ΔEC_{50}

- a) Following Free-Wilson approach in the case of at least two (or more) substituent, such as in this example:





Spencer M. **Free** and James W. **Wilson**: A Mathematical Contribution to Structure Activity Studies.

How we can use this ΔEC_{50}

- a) In this case, it can assume that the biological activity of a molecule can be represented as follow:

$$EC_{50}(R) = EC_{50}(H) + \Delta EC_{50}(R1) + \Delta EC_{50}(R2)$$

- b) If it can be assumed that the contribution of the substituent is independent from where is located, it is very easy to anticipate the biological activity of new disubstituted analogs:



Spencer M. **Free** and James W. **Wilson**: A Mathematical Contribution to Structure Activity Studies.

or generalizing:

$$\text{Activity} = \sum a_i x_i + C$$

a_i are the regression coefficients for each substituent;

x_i are binary variables (**1 if the substituent is present, 0 if absent**);

C is a constant (**could you read the meaning of the constant?**)



Spencer M. **Free** and James W. **Wilson**: A Mathematical Contribution to Structure Activity Studies.

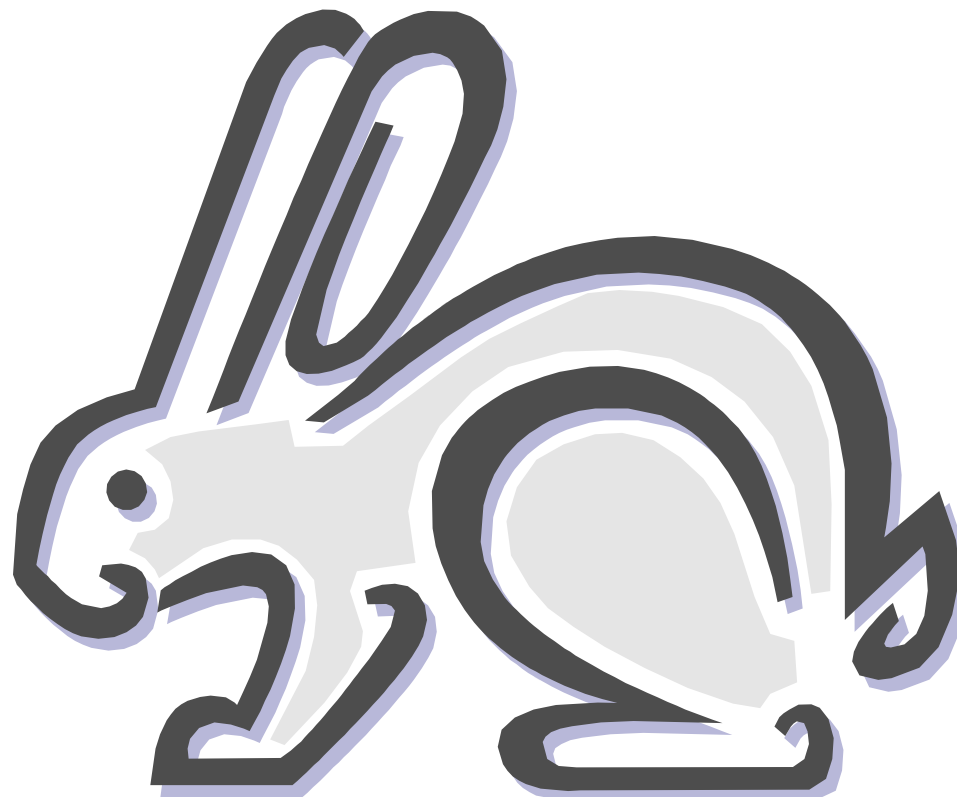
$$\text{Activity} = \sum a_i x_i + C$$

Despite the descriptive simplicity of the additive (linear) model proposed by Free and Wilson, there are **rare** examples of quantitative structure-activity relationships that have been shown to be describable with acceptable predictive accuracy!

Why?



Is the world made of straight lines?



... at that of the Free-Wilson time it seemed so!!!



I like to introduce Louis Hammett... and his sigma (σ)



Louis Plack Hammett (April 7, 1894 — February 9, 1987)

1940
Physical Organic Chemistry.
New York; McGraw-Hill.



From the birth of organic physical chemistry... to an

96

LOUIS P. HAMMETT

Vol. 59

[CONTRIBUTION FROM THE DEPARTMENT OF CHEMISTRY, COLUMBIA UNIVERSITY]

The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives

BY LOUIS P. HAMMETT

The effect of a substituent in the meta or para position of the benzene ring upon the rate or upon the equilibrium of a reaction in which the reacting group is in a side chain attached to the ring may be represented by a simple formula which is valid within a reasonable precision in a surprising variety of cases. The formula is

$$-RT \ln K + RT \ln K^0 = \Delta F = A/d^2 \left(\frac{B_1}{D} + B_2 \right)$$

K is a rate constant or an equilibrium constant for a substituted reactant, K^0 is the corresponding quantity for the unsubstituted reactant, ΔF is a free energy change or its kinetic analog, d is the distance from the substituent to the reacting group, D is the dielectric constant of the medium in which the reaction occurs, and the quantities A , B_1 and B_2 are constants independent of temperature and solvent. Of these A depends only upon the substituent and its position in the ring relative to the reacting group (with one exception, the two values necessary for the para nitro group), while B_1 and B_2 depend only upon the reaction.

The most important practical feature of equation (1) is the separation of the effect of a substituent into two constants, one of which depends on

σ is a substituent constant, dependent upon the substituent; ρ is a reaction constant, dependent upon the reaction, the medium and the temperature. Since the only data available consist of values of the $\sigma\rho$ product, it is necessary to assign an arbitrary value to some one σ or ρ . The choice of a value of unity for the ρ constant in the ionization equilibrium of substituted benzoic acids in water solution at 25° was determined by the large amount of accurate data available from the recent work of Dippy and co-workers.³ On this basis the difference between the logarithm of the ionization constant of a substituted benzoic acid and the logarithm of the ionization constant of benzoic acid gives the value of the σ constant for that substituent. With the nucleus of σ values thus provided, ρ values have been derived by least squares methods for other reactions, and from these in turn σ values have been obtained for substituents whose effects upon the ionization constant of benzoic acid are unknown or inaccurately known. After any new σ value was obtained it was used for the calculation of subsequent ρ values, so that the order of the calculations, which is that of the key numbers in the

L.P. Hammett, *The Effect of Structure upon the Reactions of Organic Compounds. Benzene Derivatives*. J. Am. Chem. Soc. 1937, 59, 96-103.



From the electronic effects of the substituent to the definition of the constant σ :

pKa for some substituted benzoic acids (H₂O, 25°C)
and corresponding s values of σ the substituents.

$$\sigma_x = pKa(C_6H_5COOH) - pKa(XC_6H_4COOH)$$

$$\sigma_x = \log \frac{Ka(XC_6H_4COOH)}{Ka(C_6H_5COOH)}$$

Substituent X	pKa (XC ₆ H ₄ COOH)	σ_x
H	4.20	0
<i>m</i> -OCH ₃	4.09	0.11
<i>m</i> -F	3.86	0.34
<i>m</i> -NO ₂	3.49	0.71
<i>p</i> -NO ₂	3.42	0.78
<i>p</i> -CH ₃	4.37	-0.17
<i>p</i> -OCH ₃	4.48	-0.28

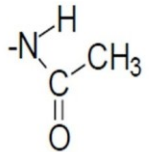
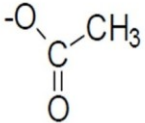
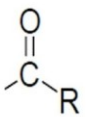
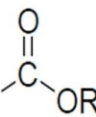
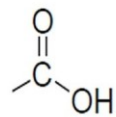
Sostituente	σ_m	σ_p
NO ₂	0.71	0.78
CN	0.61	0.70
CF ₃	0.43	0.54
CH ₃ COO	0.39	0.31
Br	0.39	0.23
CH ₃ CO	0.38	0.48
CO ₂ C ₂ H ₅	0.37	0.45
Cl	0.37	0.22
CHO	0.36	0.44
CO ₂ H	0.35	0.44
I	0.35	0.28
F	0.34	0.06
C≡CH	0.20	0.23
SCH ₃	0.15	≈ 0
OH	0.13	-0.38
OCH ₃	0.11	-0.28
C ₆ H ₅	0.05	≈ 0
H	0	0
CH ₃	-0.06	-0.17
C ₂ H ₅	-0.07	-0.15
CH(CH ₃) ₂	-0.07	-0.15
C(CH ₃) ₃	-0.10	-0.20
N(CH ₃) ₂	-0.15	-0.63
NH ₂	-0.16	-0.57

Substituent Effect Theory

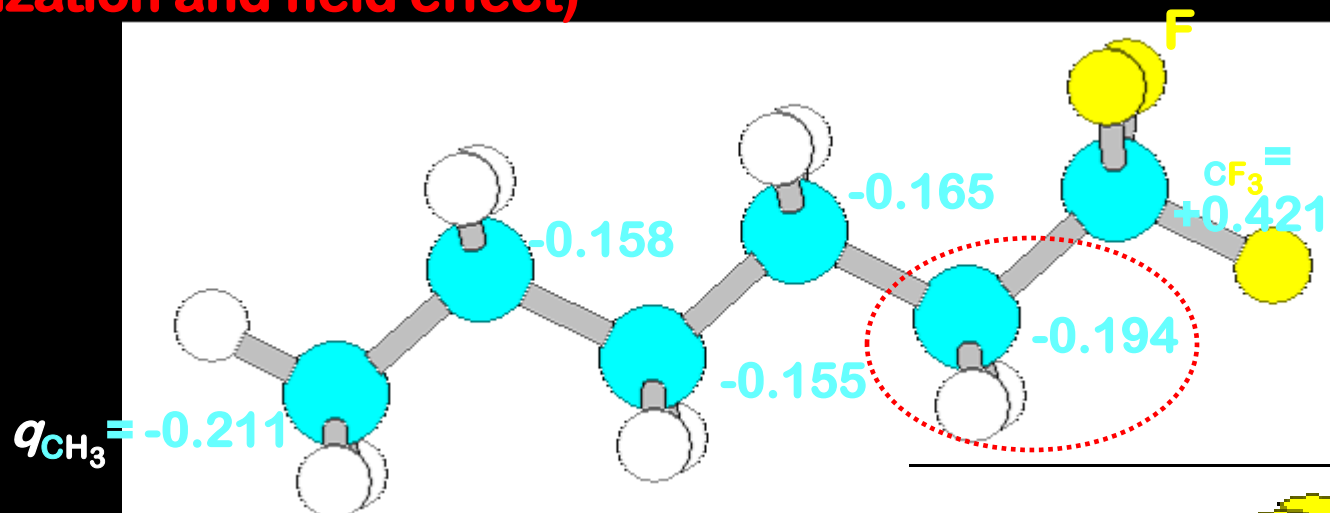
To summarize:

+R π -donating; -R π -withdrawing
 +I σ -donating -I σ -withdrawing

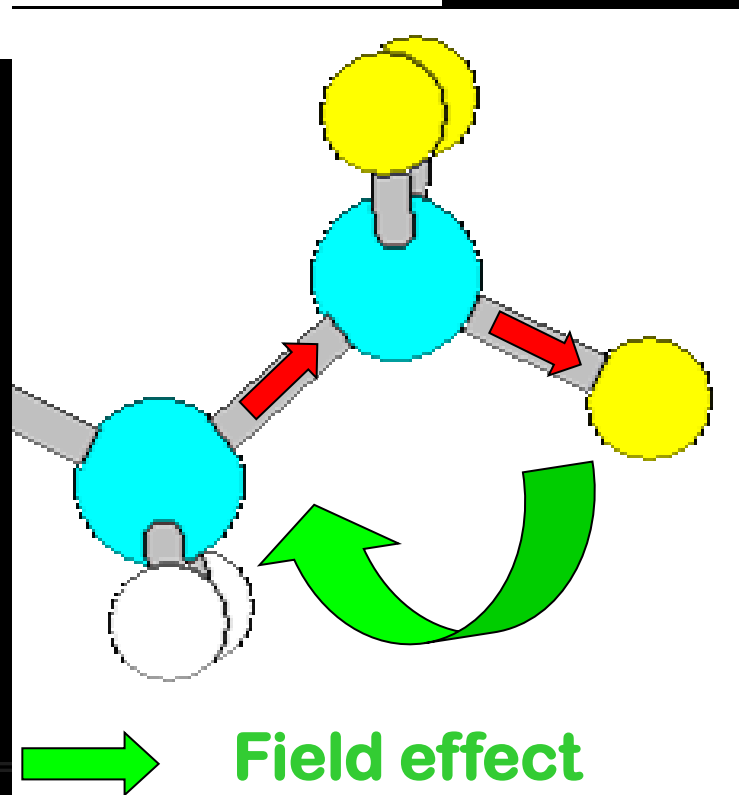
- Often the types of substituents are

+R, +I	+R, -I	-R, -I
alkyl, CH ₃ -, CH ₃ CH ₂ -, trialkylsilyl-,	-NH ₂ , -OH, -X, -SH -NR ₂ , -OR, -SR  	   -C≡N -SO ₃ H -NO ₂ -SO ₂ R

Inductive effect (polarization and field effect)

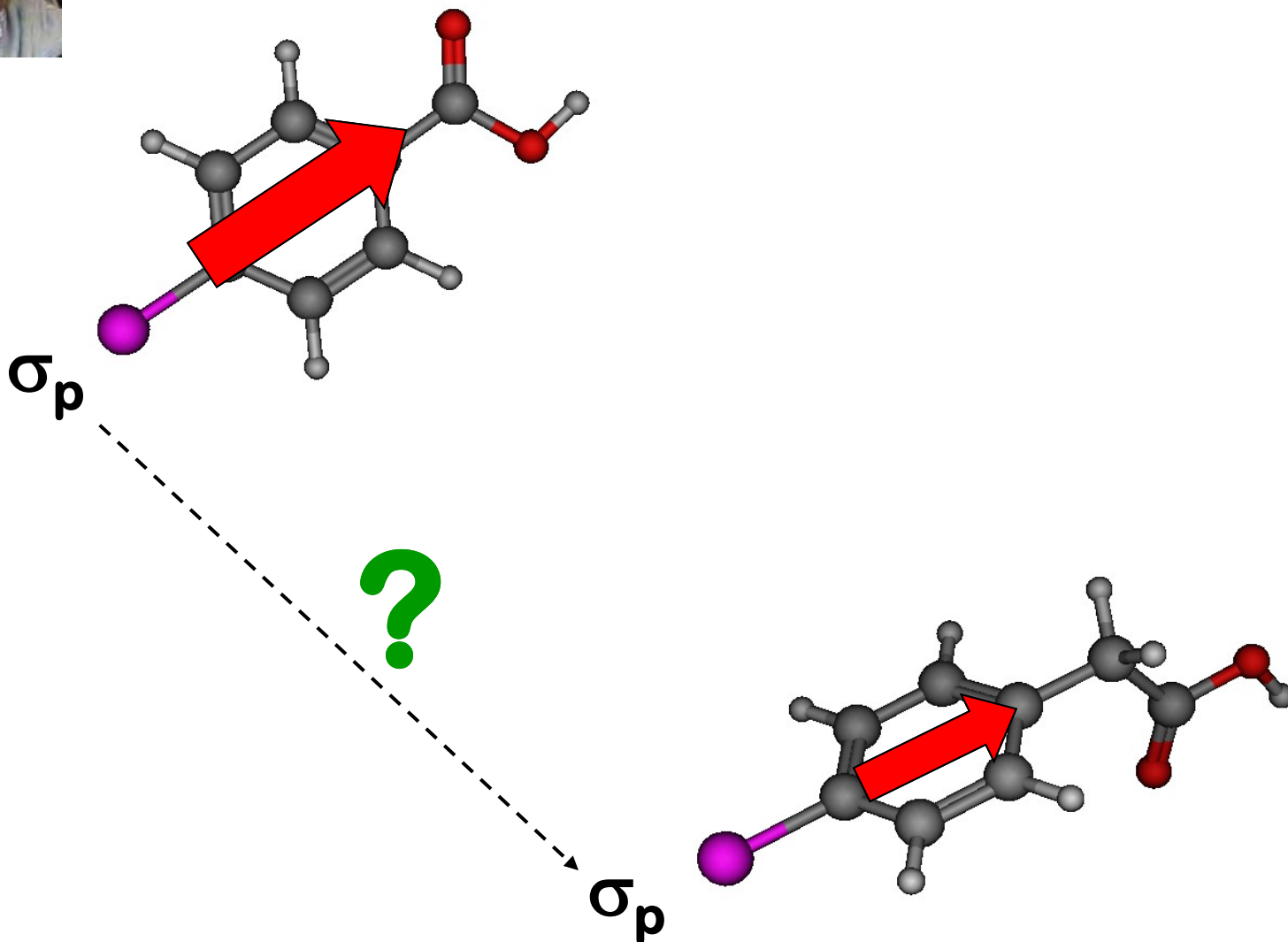


➔ Polarization effect





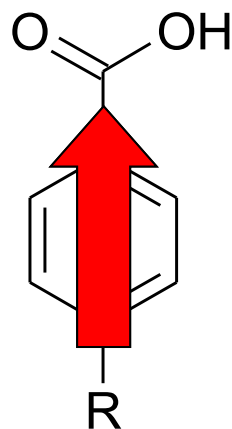
... but is sigma (σ) an invariant property of the substituent?



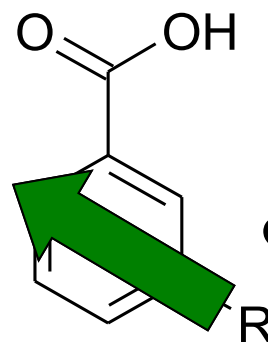
Electronic effect of the substituent and constant σ :

But what electronic effect was Hammett discussing as he defined his constant σ ?

$$\sigma_x = \log \frac{K_a(XC_6H_4COOH)}{K_a(C_6H_5COOH)}$$

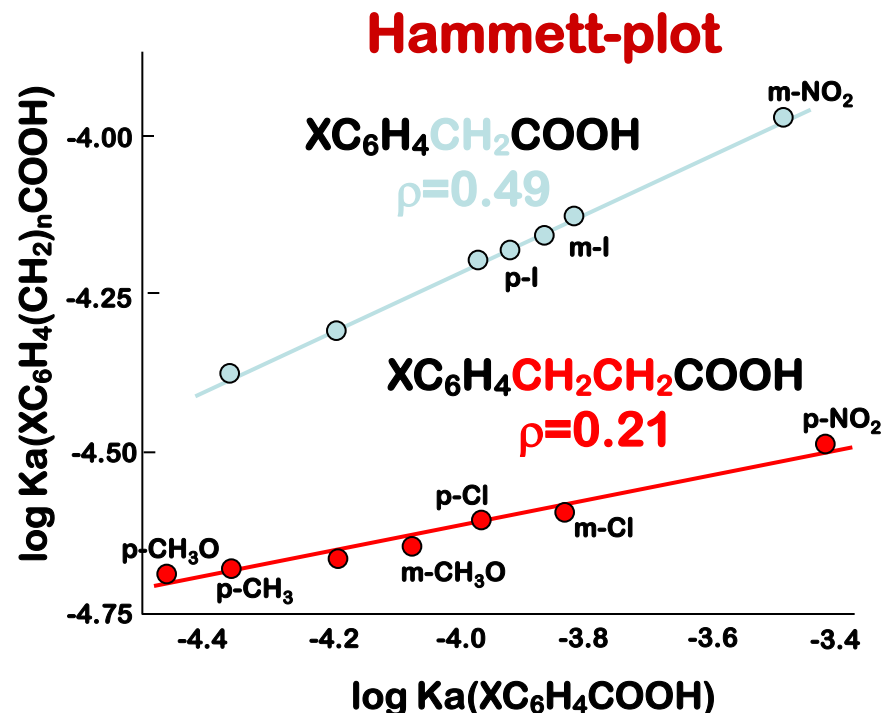
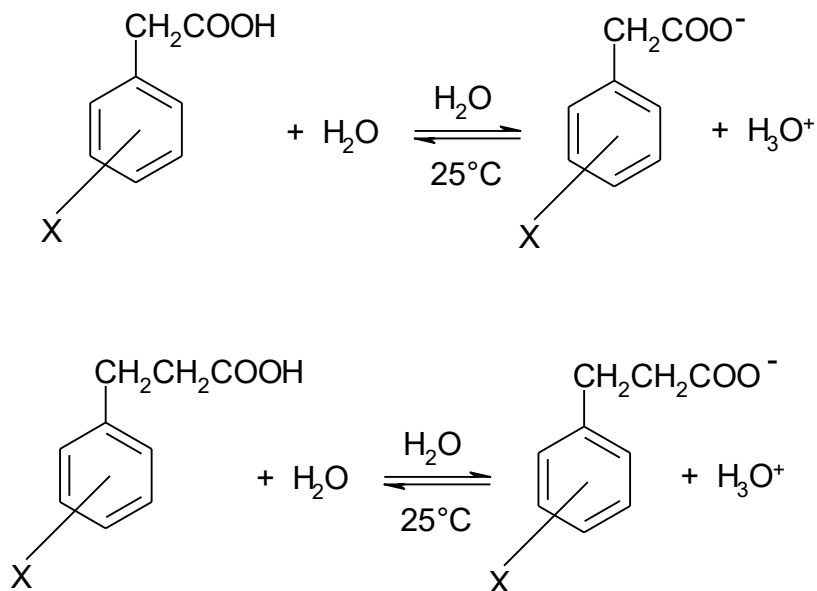


$$\sigma_p = \sigma_I + \sigma_R$$



$$\sigma_m = \sigma_I + \alpha\sigma_R$$

Application of the constant σ : dissociation of *homo*-benzoic acids



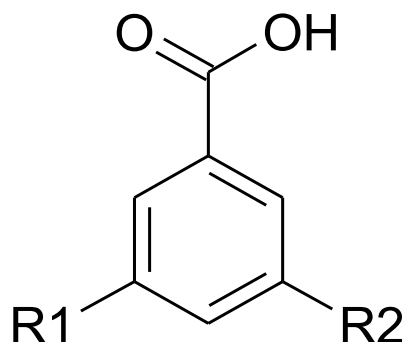
$$\log \frac{K_a(\text{XC}_6\text{H}_4\text{CH}_2\text{COOH})}{K_a(\text{C}_6\text{H}_5\text{CH}_2\text{COOH})} = \rho \log \frac{K_a(\text{XC}_6\text{H}_4\text{COOH})}{K_a(\text{C}_6\text{H}_5\text{COOH})}$$

$$\log \frac{K_a(\text{XC}_6\text{H}_4\text{CH}_2\text{COOH})}{K_a(\text{C}_6\text{H}_5\text{CH}_2\text{COOH})} = \rho \sigma_X$$



The first example of "additivity" in organic chemistry:

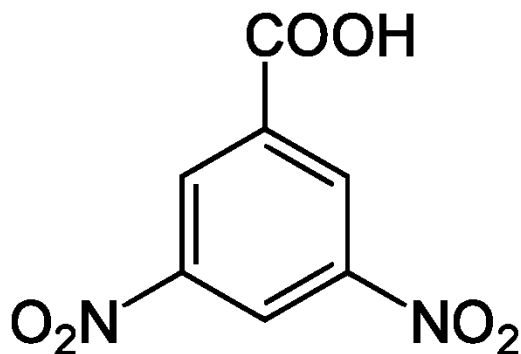
For benzoic acids serie is 1



$$\log K_{a(R1,R2)} = \log K_{a_{rif}} + \rho \sum \sigma_{Ri}$$

$$pK_{a(R1,R2)} = pK_{a_{rif}} - \rho \sum \sigma_{Ri}$$

a nice and convincing example:



$$pK_{a(R1,R2)} = 4.2 - (0.71 \times 2) = 2.78$$

$$pK_{a(R1,R2)} = 2.82 \text{ (exp)}$$



... and after Hammett

The fragments method of Hansch-Fujita.

(*J. Am. Chem. Soc.* 1964, 86, 5175)

Dec. 5, 1964

SUBSTITUENT CONSTANT, π , FROM PARTITION COEFFICIENTS

5175

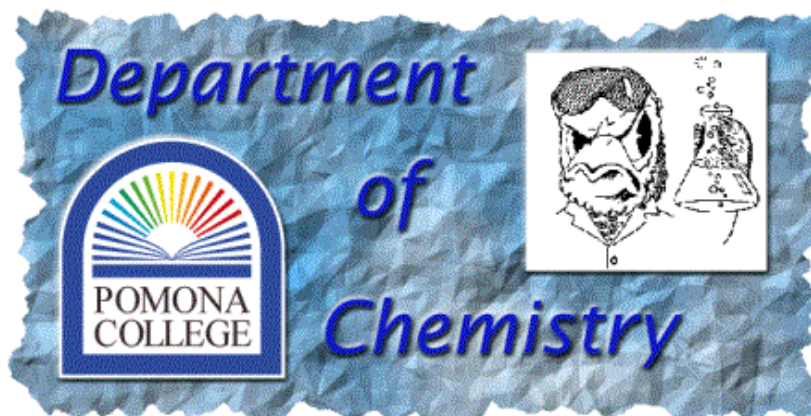
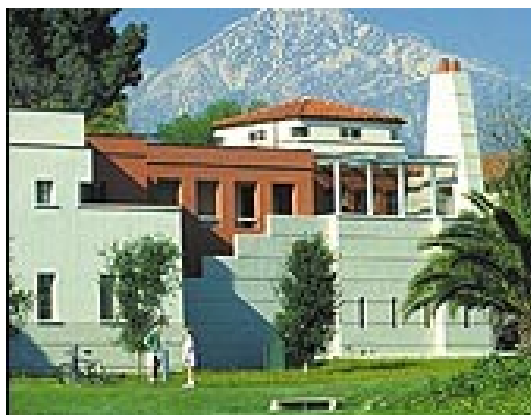
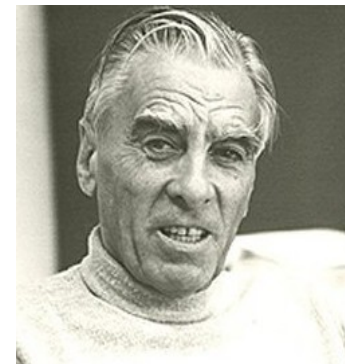
[CONTRIBUTION FROM THE DEPARTMENT OF CHEMISTRY, POMONA COLLEGE, CLAREMONT, CALIF.]

A New Substituent Constant, π , Derived from Partition Coefficients

BY TOSHIO FUJITA,^{1a} JUNKICHI IWASA,^{1b} AND CORWIN HANSCH

RECEIVED FEBRUARY 19, 1964

The partition coefficients between 1-octanol and water have been determined for 203 mono- and disubstituted benzenes. From these values a substituent constant, π , has been calculated for 67 functional groups. The constant π is defined as: $\pi = \log P_X - \log P_H$ where P_X is the partition coefficient of a derivative and P_H is that of the parent compound. π has been derived for many of the functions from eight different systems: benzene, nitrobenzene, aniline, phenol, benzyl alcohol, benzoic acid, phenylacetic acid, and phenoxyacetic acid. It is found that, although π varies continuously for a given function depending on its electronic environment, the range over which it varies is not great. In certain of the systems, π -values are related by a simple linear expression.

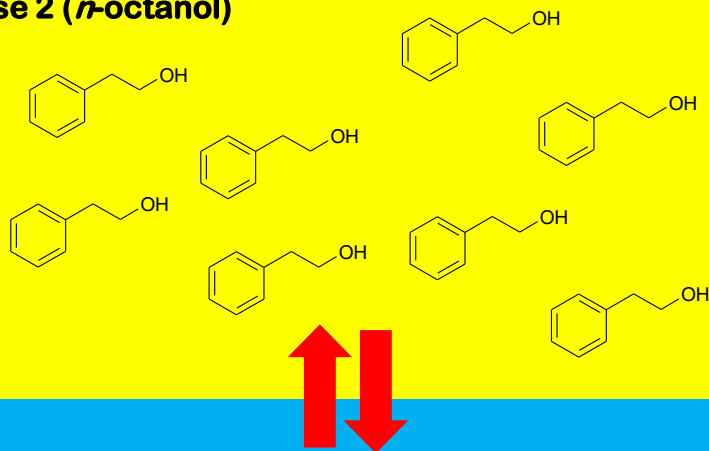


POMONA COLLEGE (Claremont, California)

Hydrophobicity and partition coefficient

$$\mu_{(phase2)} = \mu_{(phase2)}^0 - RT \ln[C_{(phase2)}]$$

Phase 2 (*n*-octanol)



At the equilibrium:

$$\mu_{(phase1)} = \mu_{(phase2)}$$

$$\mu_{(phase1)}^0 - RT \ln[C_{(phase1)}] = \mu_{(phase2)}^0 - RT \ln[C_{(phase2)}]$$

$$\mu_{(phase1)}^0 - \mu_{(phase2)}^0 = RT \ln[C_{(phase1)}] - RT \ln[C_{(phase2)}]$$

$$\mu_{(phase1)}^0 - \mu_{(phase2)}^0 = RT \ln \left[\frac{C_{(phase1)}}{C_{(phase2)}} \right]$$

$$\left[\frac{C_{(phase1)}}{C_{(phase2)}} \right] = P \quad \text{Partition coefficient}$$

$$\mu_{(phase1)} = \mu_{(phase1)}^0 - RT \ln[C_{(phase1)}]$$



Hydrophobicity and partition coefficient

$$\left[\frac{C_{(phase1)}}{C_{(H_2O)}} \right] = P \quad \text{Partition coefficient}$$

We can define: “*hydrophobic*” a compound with $P > 1$;
 “*hydrophilic*” a compound with $P < 1$.

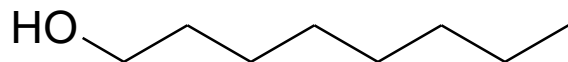
$$\log \left[\frac{C_{(phase1)}}{C_{(H_2O)}} \right] = \log P$$

We can define: “*hydrophobic*” a compound with $\log P > 0$;
 “*hydrophilic*” a compound with $\log P < 0$.



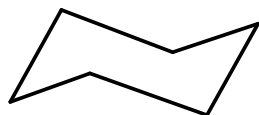
Hydrophobicity and partition coefficient

How we can choose the second phase:



***n*-octanol**

- immiscible in water even if 27% of water dissolves in it... so the first region of hydration of the solute is preserved;
- UV transparent;
- Low vapor pressure.



cyclohexane

Immiscible in water and very low amount of water dissolves in it... so also the first region of hydration of the solute is lost.
The differences between the logP values in *n*-octanol and cyclohexane is a measure of the de-hydration energy of a solute.



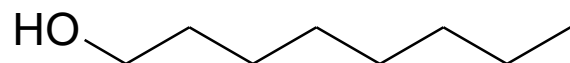


Do you see any similarity?

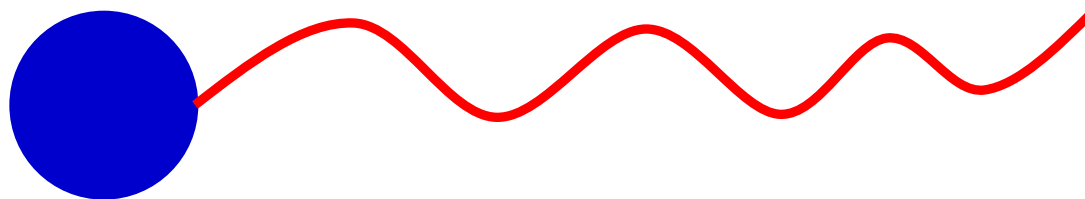
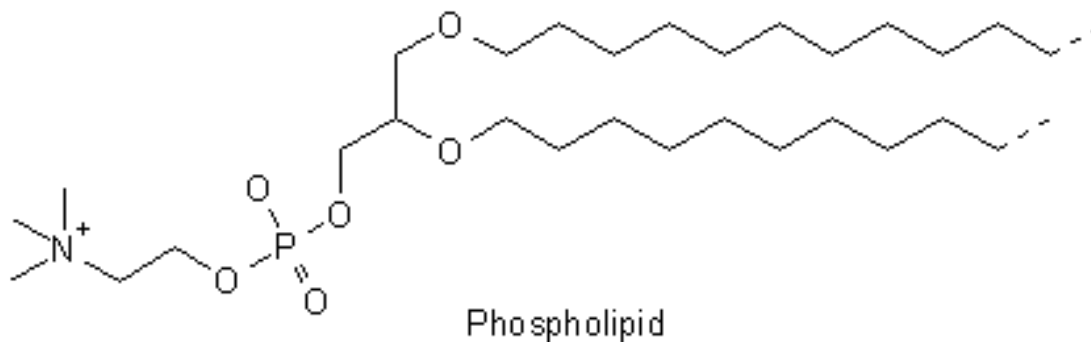
...in addition to this?



How we can choose the second phase:

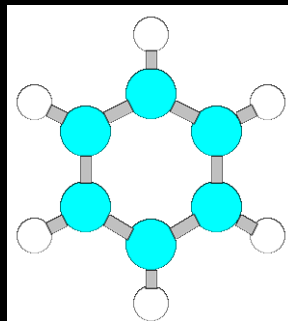


n-octanol • “Similarity” with biological membrane;

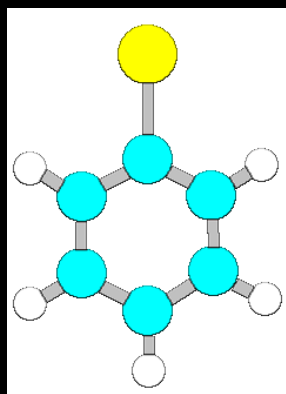
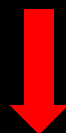




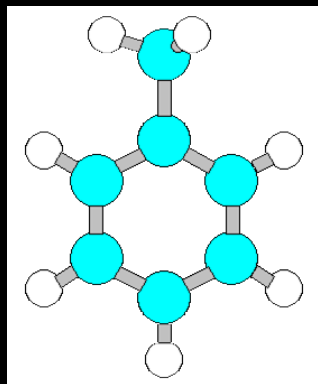
The hydrophobic constant of substituent, π



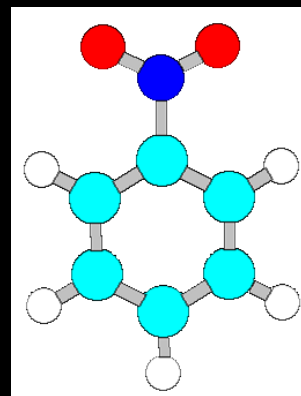
$\log P = 2.13$



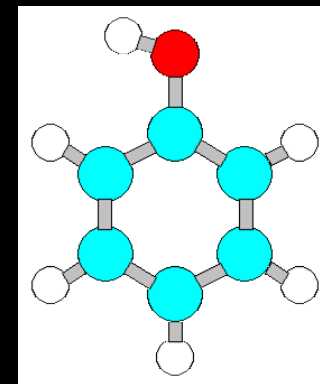
$\log P = 2.99$



$\log P = 2.69$



$\log P = 1.85$



$\log P = 1.48$

$$\pi_X = \log P_{C_6H_5X} - \log P_{C_6H_6}$$

$$\pi_{Br} = 0.86$$

$$\pi_{CH_3} = 0.56$$

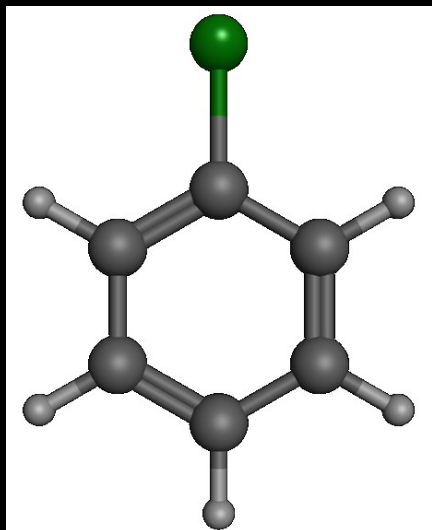
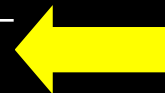
$$\pi_{NO_2} = -0.28$$

$$\pi_{OH} = -0.67$$



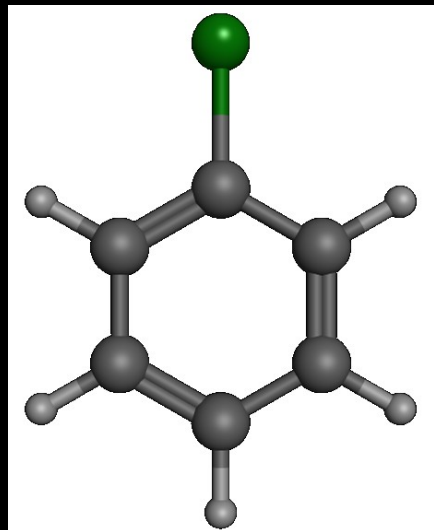
The hydrophobic constant of substituent, π

Substituent	π aromatic
F	0.14
Cl	0.71
Br	0.86
I	1.12
OH	-0.67
OCH ₃	-0.02
SCH ₃	0.61
CN	-0.57
COOH	-0.28
COOCH ₃	-0.01
COCH ₃	-0.55
NH ₂	-1.23
N(CH ₃) ₂	-0.28
NO ₂	-0.28
CH ₃	0.56



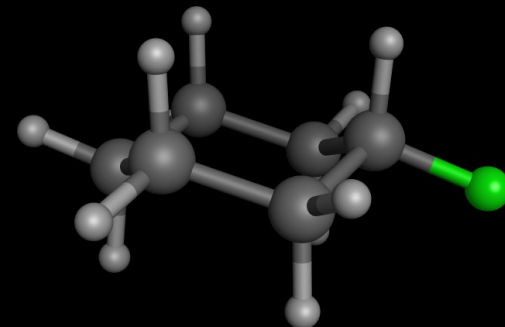


The hydrophobic constant of substituent, π



Ref. benzene
 $\log P = 2.13$

Substituent	π aromatic
F	0.14
Cl	0.71
Br	0.86
I	1.12
OH	-0.67
OCH ₃	-0.02
SCH ₃	0.61
CN	-0.57
COOH	-0.28
COOCH ₃	-0.01
COCH ₃	-0.55
NH ₂	-1.23
N(CH ₃) ₂	-0.28
NO ₂	-0.28
CH ₃	0.56

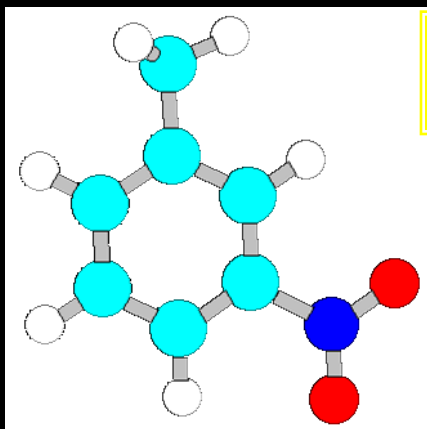


Ref. cyclohexane
 $\log P = 3.44$



The additive rule:

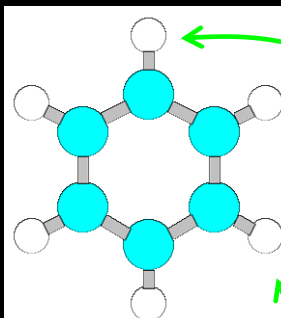
It works? Just check together:



$\log P: 2.45$

IMPRESSIVE!!!!

$\log P_{\text{calculated}} (\text{ClogP}):$



$$\log P_{\text{C}_6\text{H}_6} = 2.13$$

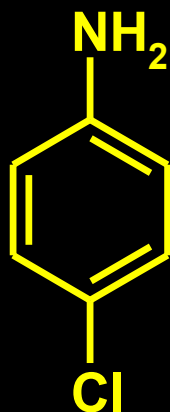
$$\pi_{\text{CH}_3} = 0.56$$

$$\pi_{\text{NO}_2} = -0.28$$

$$2.41$$

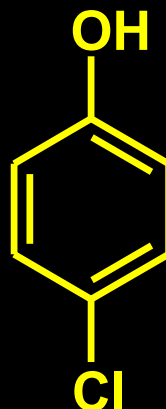


The hydrophobic constant of substituent, π



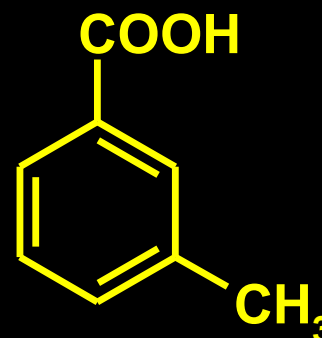
$\log P$ 1.80

$\text{Clog } P$ 1.61



$\log P$ 2.30

$\text{Clog } P$ 2.14



...

$\log P$ 2.40

$\text{Clog } P$ 2.38

...not too bad!!!

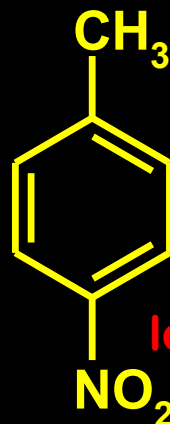


The hydrophobic constant of substituent, π

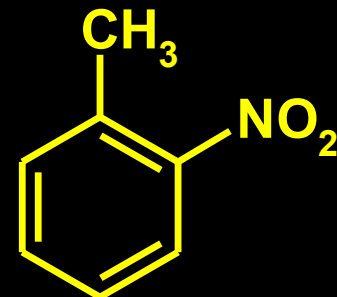
But not so good...



$\log P$ 2.45

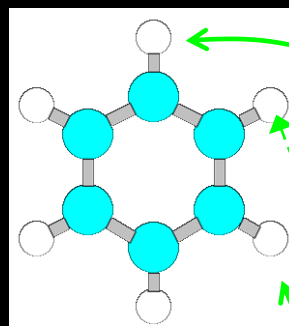


$\log P$ 2.42



$\log P$ 2.30

$\log P_{\text{calculated}}$ (ClogP):



$$\log P_{\text{C}_6\text{H}_6} = 2.13$$

$$\pi_{\text{CH}_3} = 0.56$$

$$\pi_{\text{NO}_2} = -0.28$$

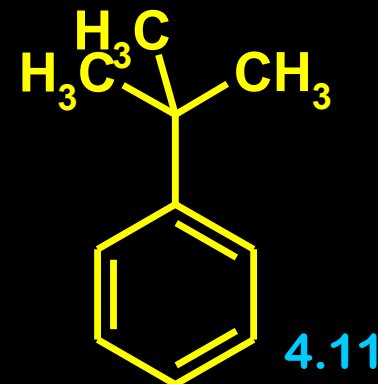
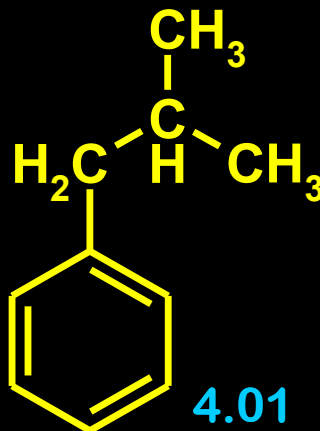
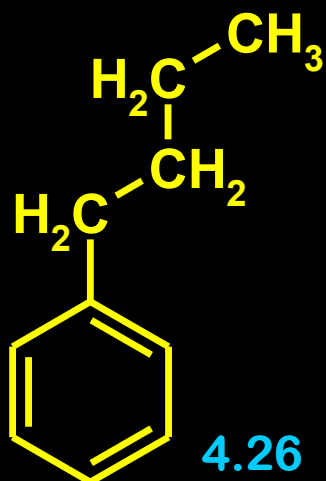
$$\hline 2.41$$

?

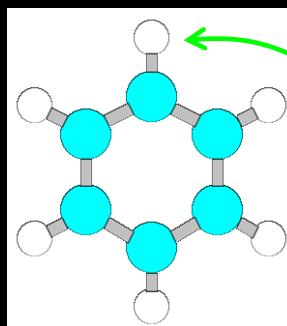


The hydrophobic constant of substituent, π

again...



$\log P_{\text{calculated}}$ (ClogP):



$$\log P_{\text{C}_6\text{H}_6} = 2.13$$

$$4 \times \pi_{\text{CH}_3} = 2.24$$

$$4.37$$

?



Corwin... we have a problem!

$$\pi_H = 0$$


Is this true?

But how we can demonstrate this is not true?



Corwin... we have a problem!

Here is how he fixed the problem... as an engineer usually done introduction the magic *correction factors*!!!

Remember : a *correction factor* is any mathematical adjustment made to a calculation to account for deviations in either the sample or the method of measurement.

The hydrophobic correction factors, $\Delta\pi$



- Branched carbon chain ($\Delta\pi = -0.20$);
- Double bond ($\Delta\pi = -0.30$);
- Intra-molecular H-bond ($\Delta\pi = 0.65$);
- Ring condensation ($\Delta\pi = -0.20$)
- ...

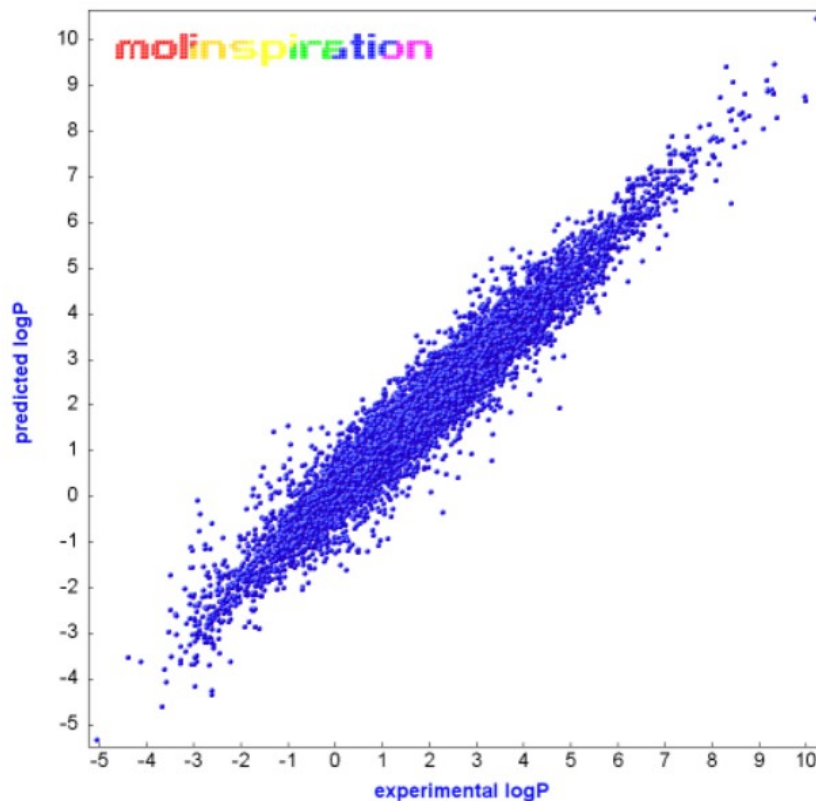
Finally the $\log P$ calculated by Hansch-Fujita:

$$\text{Clog } P = \log P_{\text{ref}} + \sum \pi_{\chi_i} + \sum \Delta\pi$$



Nowadays, are these methods reliable?

... you decide!!!



$n = 12'202$, $r^2 = 0.944$, $r = 0.972$, $\text{stdev} = 0.428$, $\text{mae} = 0.328$

credits: <https://www.molinspiration.com/cgi-bin/properties>



ClogP... a wonderfully precious molecular descriptor!



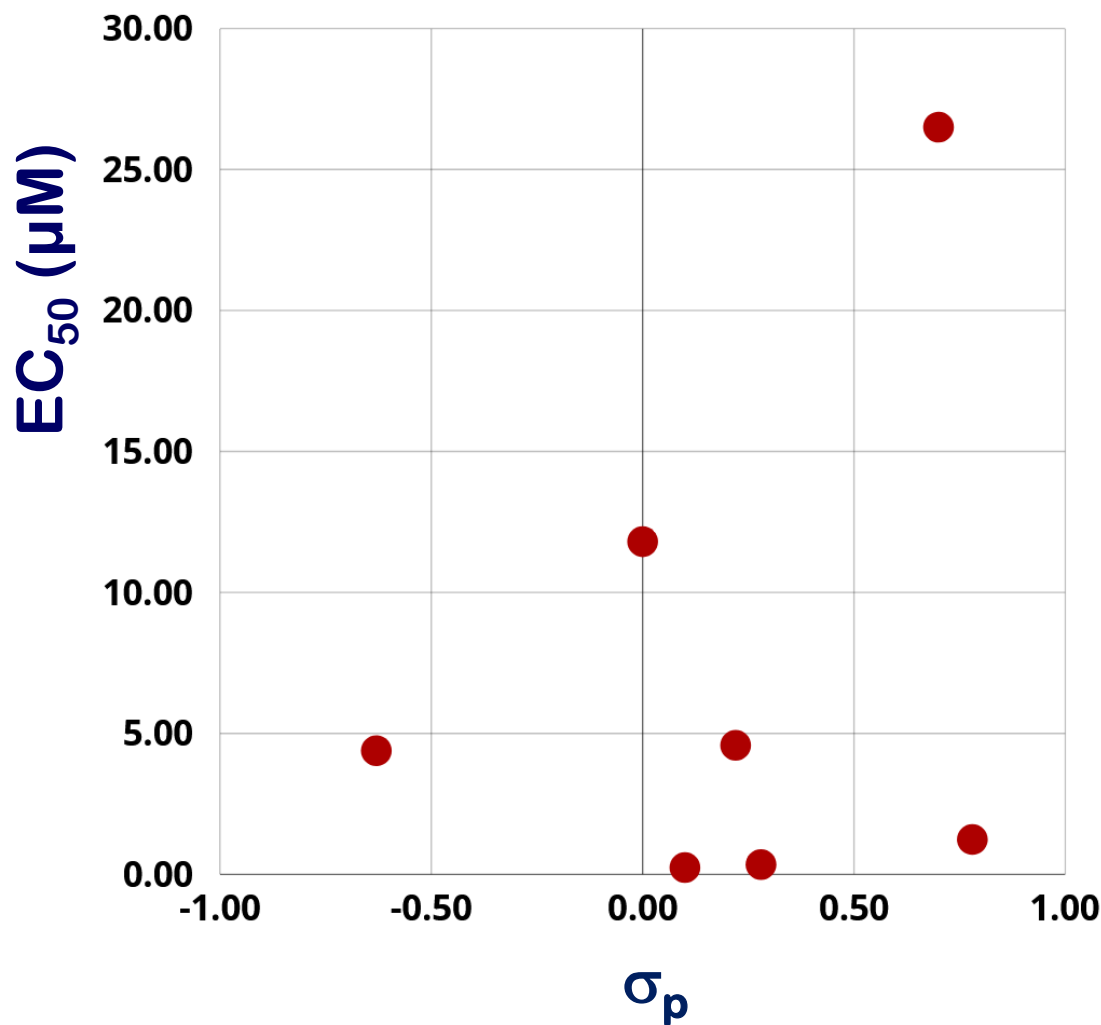
Can we replace in our scatterplot the “cmpd number” with “ σ_p ”?



Cmpd number	Cmpd name	X	EC ₅₀ (μM)	σ_p
1	6a	H	11.80 ± 1.90	0
2	6b	Cl	1.24 ± 0.11	0.78
3	6d	NO ₂	4.58 ± 0.29	0.22
4	6e	CN	26.50 ± 5.87	0.70
5	6f	C ₆ H ₅	0.24 ± 0.30	0.10
6	6g	N(CH ₃) ₂	4.39 ± 0.67	-0.63
7	6h	I	0.35 ± 0.05	0.28
8	6i	NHCHO	???	0.12



Can we replace in our scatterplot the “cmpd number” with “ σ_p ”?





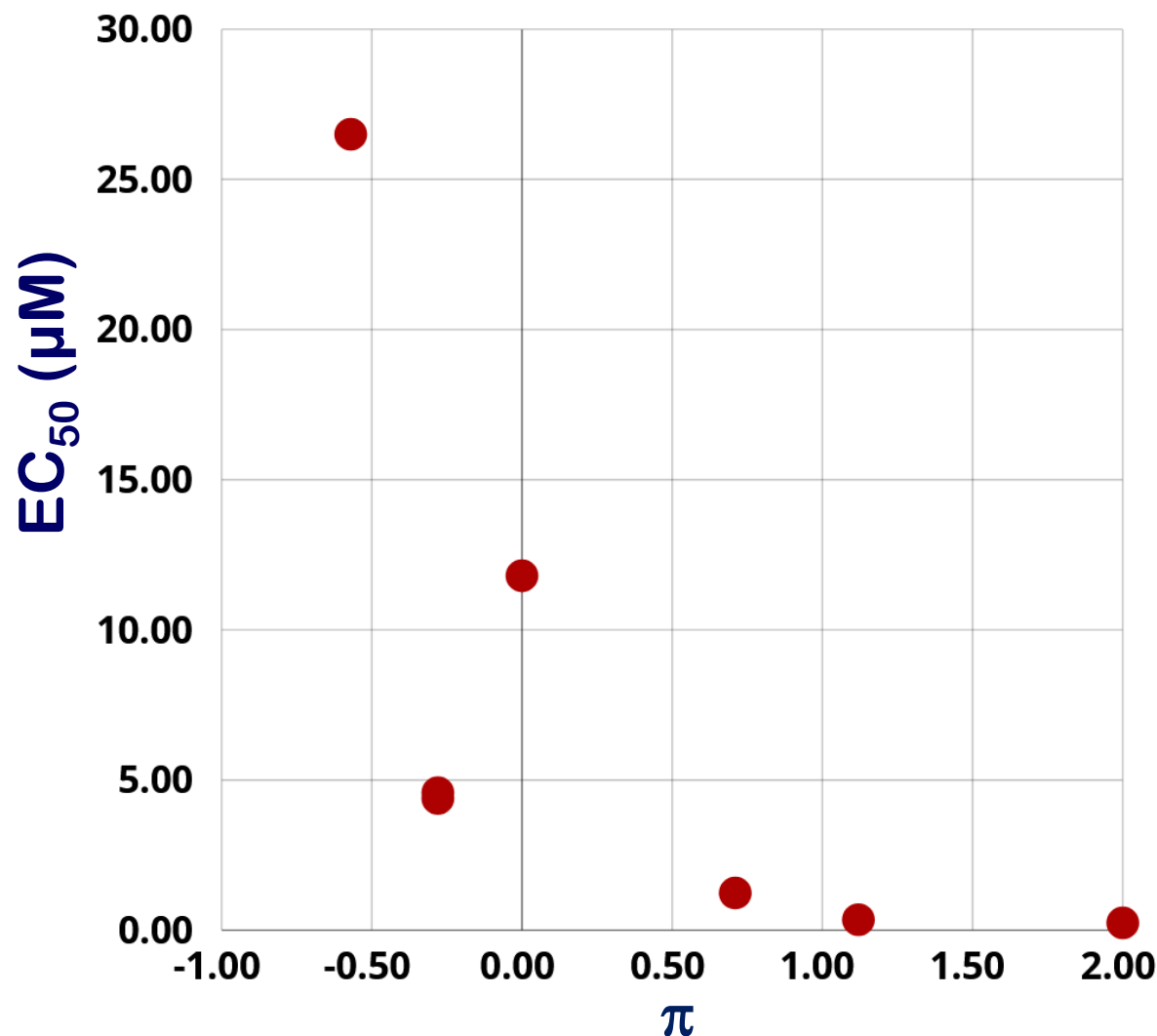
... or can we replace in our scatterplot the “ σ_p ” with “ π ”?



Cmpd number	Cmpd name	X	EC ₅₀ (μM)	σ_p	π
1	6a	H	11.80 ± 1.90	0	0
2	6b	Cl	1.24 ± 0.11	0.78	0.71
3	6d	NO ₂	4.58 ± 0.29	0.22	- 0.28
4	6e	CN	26.50 ± 5.87	0.70	- 0.57
5	6f	C ₆ H ₅	0.24 ± 0.30	0.10	2.00
6	6g	N(CH ₃) ₂	4.39 ± 0.67	-0.63	- 0.28
7	6h	I	0.35 ± 0.05	0.28	1.12
8	6i	NHCHO	???	0.12	- 0.70



... or can we replace in our scatterplot the “ σ_p ” with “ π ”?

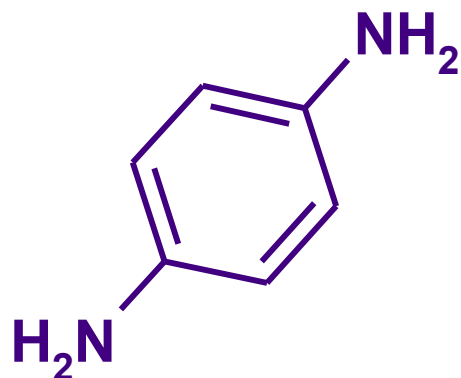




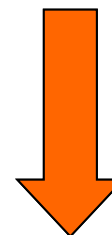
... and now?



(Q)**S**AR: we are ready to this...



Structure



Properties

PM = 108,14

pKa = 6.2

Volume = 93,9

MP = 142

PSA = 52

nC = 6

logP = -0.3

...



(Q)**S**AR: we are ready to this...

Real world

Virtual world

Chemical Compound (CC)



Chemical Structure (CS)



Chemical Properties (CP)

**Numerical
representations of CS**



Molecular Descriptors (MD)

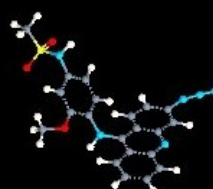


(Q)**S**AR: we are ready to this...

DRAGON 7.0 is able to calculate **5270** molecular descriptors.

File Fingerprints View Analysis Settings ?

Viewer Source Info Weights Matrices



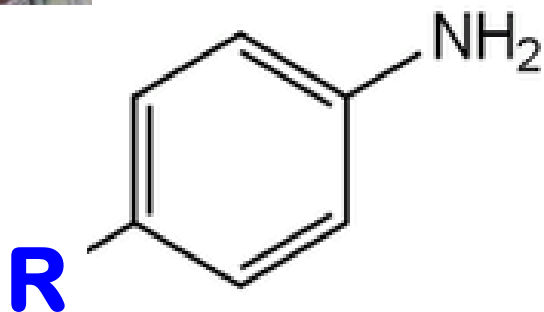
No.	NAME	CATS2D_00_DD	CATS2D_01_DD	CATS2D_02_DD	CATS2D_03_DD	CATS2D_04_DD	CATS2D_05_DD	CATS2D_06_DD	CATS2D_07_DD	CATS2D_08_DD	CATS2D_09_DD	CATS2D_00_DA
7	[N-]=[N+]=ClC=C(C=CCl	1	0	0	0	0	0	0	0	0	0	1
8	[N-]=[N+]=CC(=O)NCC(=O	1	0	0	0	0	0	0	0	0	0	0
9	[N-]=[N+]=CC(=O)OCC(N)	2	0	0	1	0	0	0	0	0	0	1
10	[N-]=[N+]=NC1CC(OC1(C	2	0	0	0	0	0	0	1	0	0	1
11	[N-]=[N+]=NC1ccc(cc1)Nc	1	0	0	0	0	0	0	0	0	0	0
12	[N-]=[N+]=NC4ccc3c(nc1c	2	0	0	0	0	1	0	0	0	0	0
13	[N-]=[N+]=NC4ccc3c(nc1c	2	0	0	0	0	1	0	0	0	0	0
14	[N-]=[N+]=NCC(N)C(=O)C	2	0	0	1	0	0	0	0	0	0	1
15	[N-]=[N+]=NCC(O)CN=[N	1	0	0	0	0	0	0	0	0	0	1
16	[N-]=[N+]=NC1cccc1	0	0	0	0	0	0	0	0	0	0	0
17	[N-]=[N+]=NCC1cccc1	0	0	0	0	0	0	0	0	0	0	0
18	[O-][N+](O)=ClCCC1	0	0	0	0	0	0	0	0	0	0	0
19	[O-][N+]=NC1ccc(c1)Cl	0	0	0	0	0	0	0	0	0	0	0
20	[O-][N+](=NC1ccc(OCC)cc	0	0	0	0	0	0	0	0	0	0	0
21	[O-][N+](=NC1cccc1)c2cc	0	0	0	0	0	0	0	0	0	0	0
22	[O-][N+](=NCC)CC	0	0	0	0	0	0	0	0	0	0	0
23	[O-][N+](=O)N=C1/NCCN	1	0	0	0	0	0	0	0	0	0	0
24	[O-][N+](=O)C(Br)=C/Cl=	0	0	0	0	0	0	0	0	0	0	0



R. Todeschini



The easiest way to describe the concept of (Q)SAR...



f(x)



Chemicals

$R(1) = \text{NH}_2$

$R(2) = \text{NHCH}_3$

$R(3) = \text{N}(\text{CH}_3)_2$

...

$R(n) = \text{NHC}_6\text{H}_5$

Activities ($\text{LC}_{50}, \mu\text{M}$)

A_1

A_2

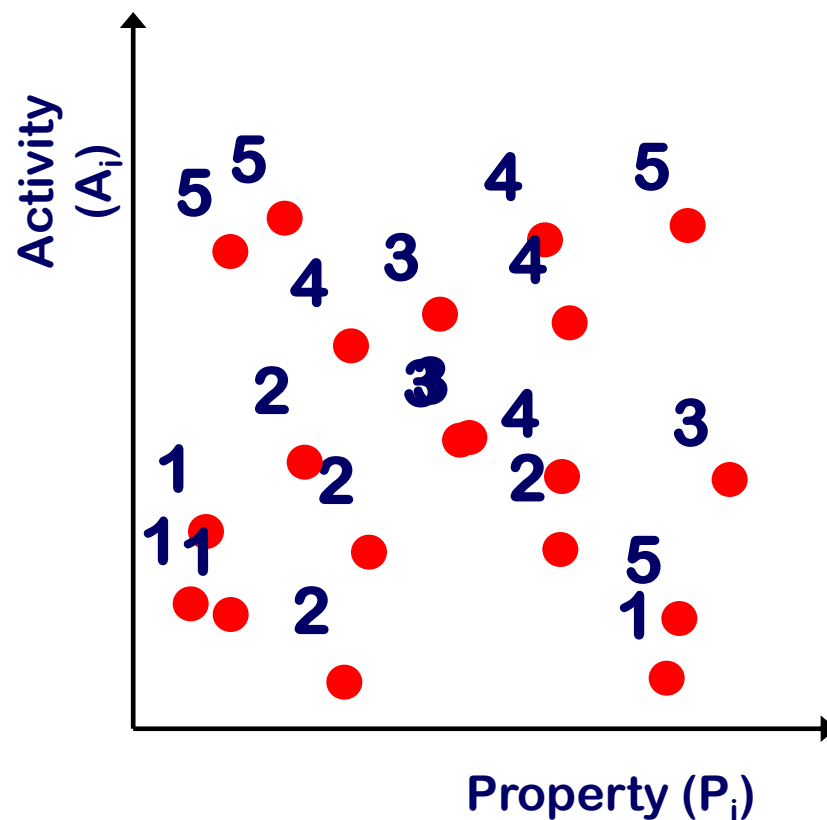
A_3

...

A_n



(Q)SAR: follow me in this wonderful experience



**Scatterplot... an interesting place where
scouting for patterns!!!**

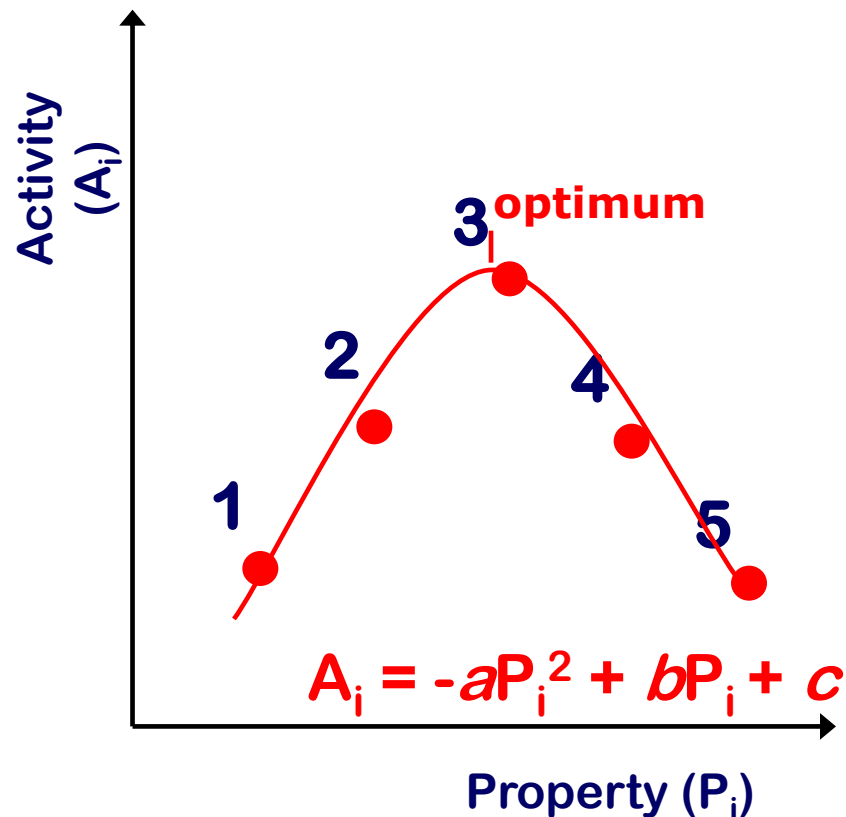
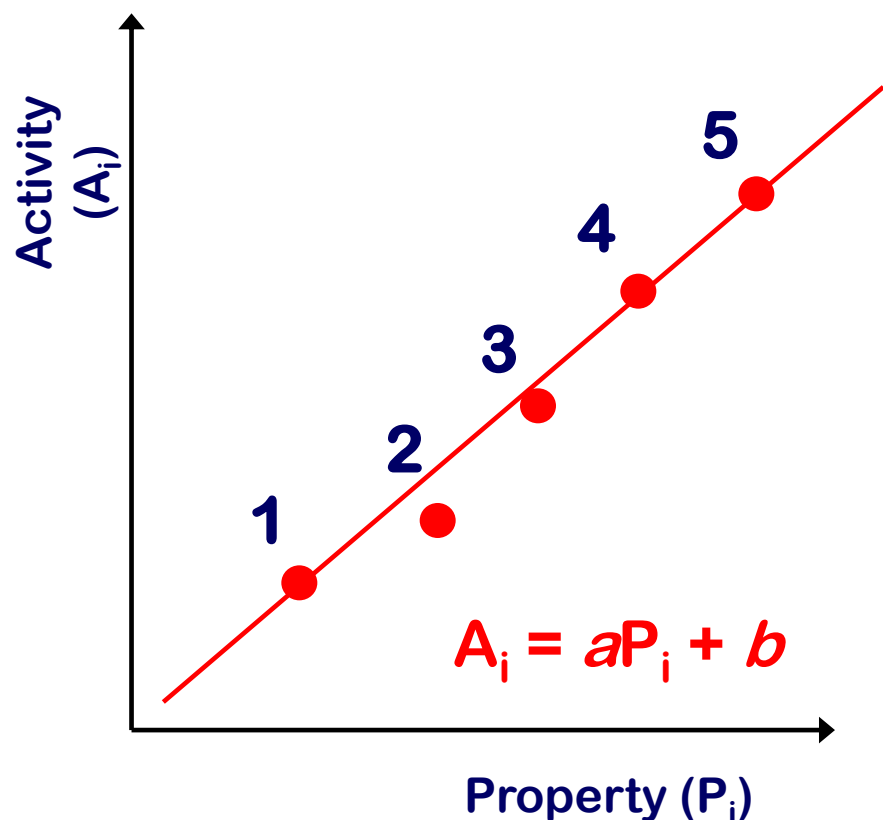


(Q)SAR: how can we select the good
“molecular descriptor(s)”?

f(x)



(Q)SAR: this is the base of a quantitative structure-activity relationship (QSAR): find *patterns*!

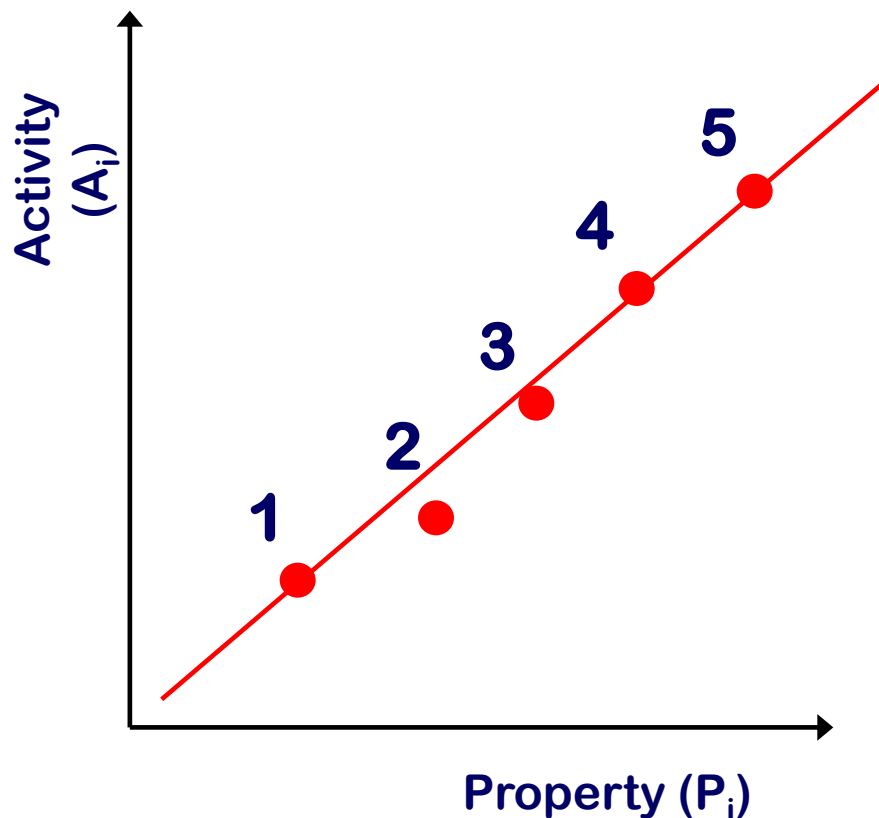




**Yes, we can look for “*regularity*”
(**pattern**) between the variability
of molecular descriptors and the
corresponding variability of
experimental activities.**



The beauty of mathematics:



$$A_i = a P_i + b$$

Discrete (few x - y correspondences)
Continuum (∞ x - y correspondences)

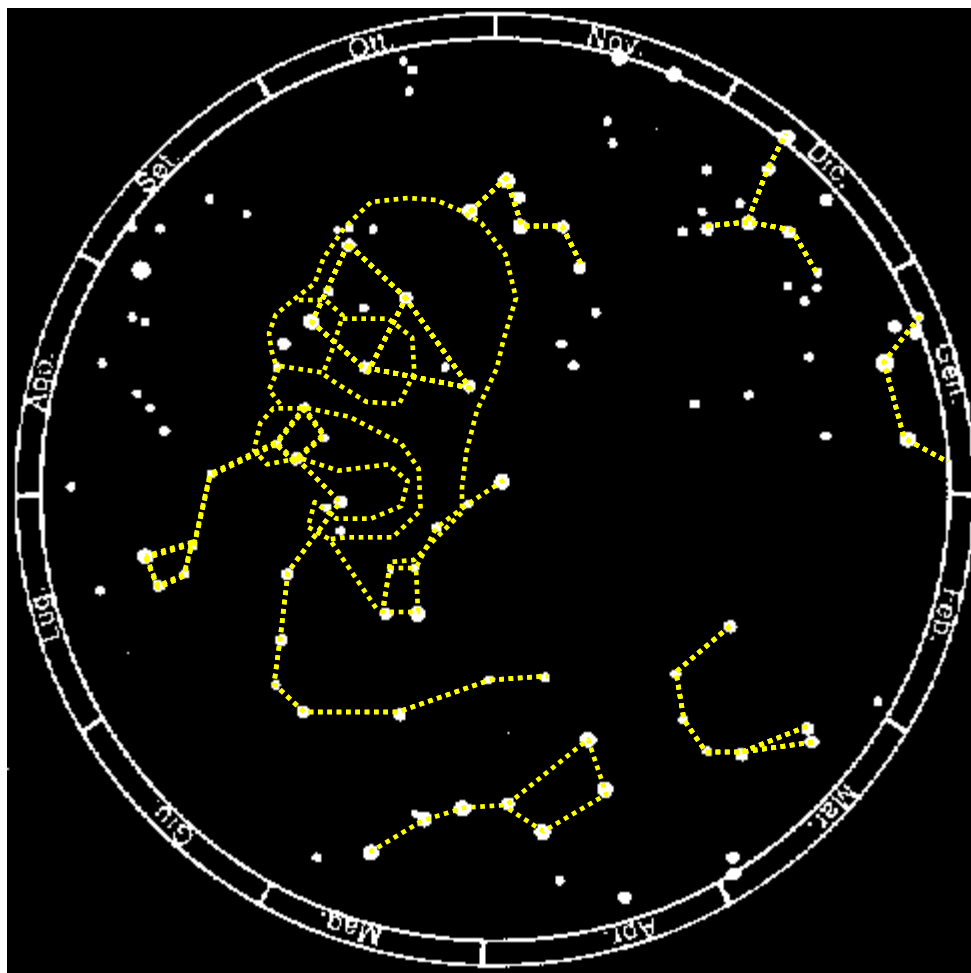


Patterns are gorgeous:

- Patterns can be mathematically condensed in equations;
- Pattern can be used to describe relationships among variables;
- Patterns can be used to predict new data;
- Patterns can be used to verify exiting data;



sometimes too gorgeous...



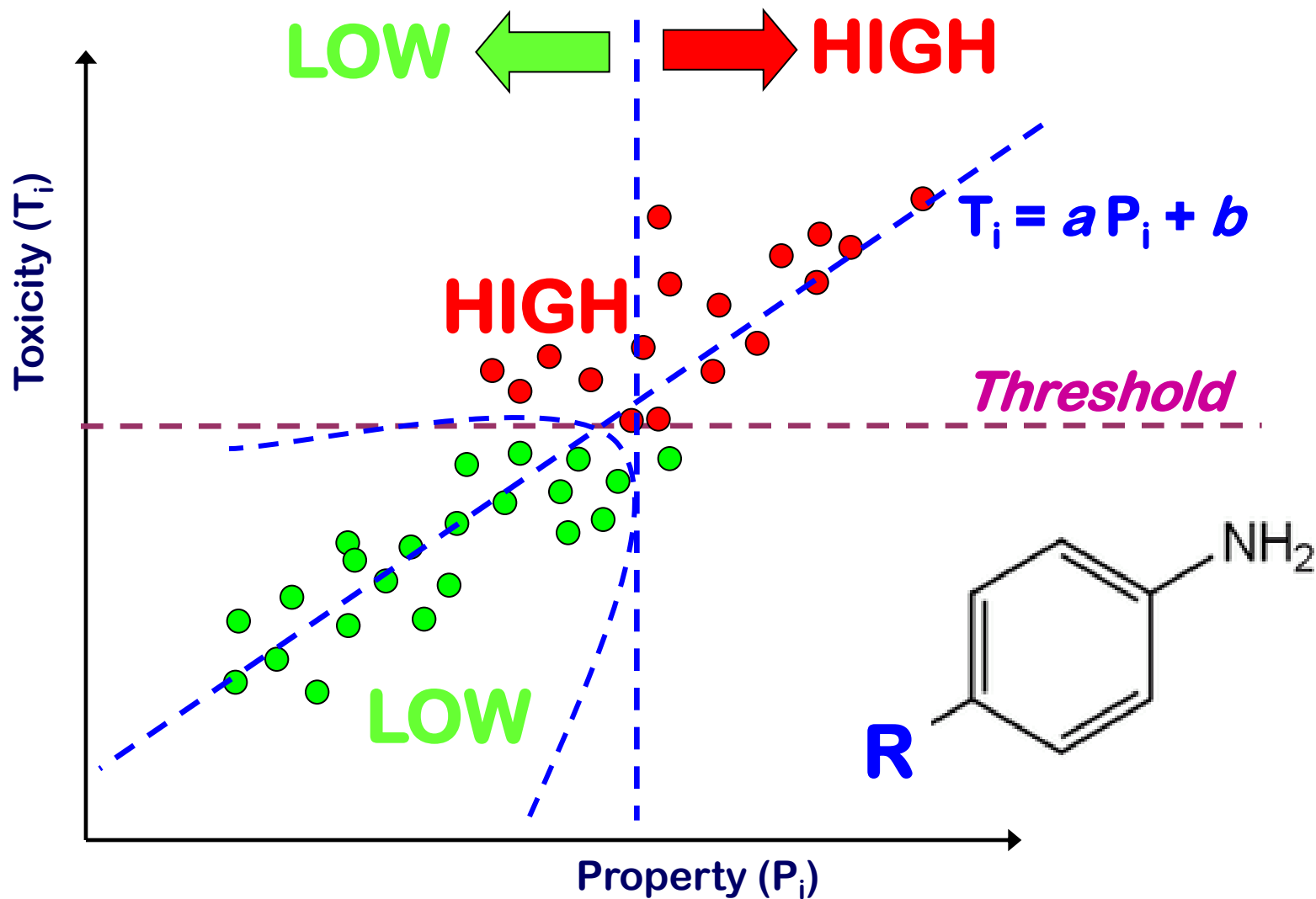


sometimes too gorgeous...

- ⊙ classification models (qualitative response)
- ⊙ regression models (quantitative response)
- ⊙ ranking models (ordered response)

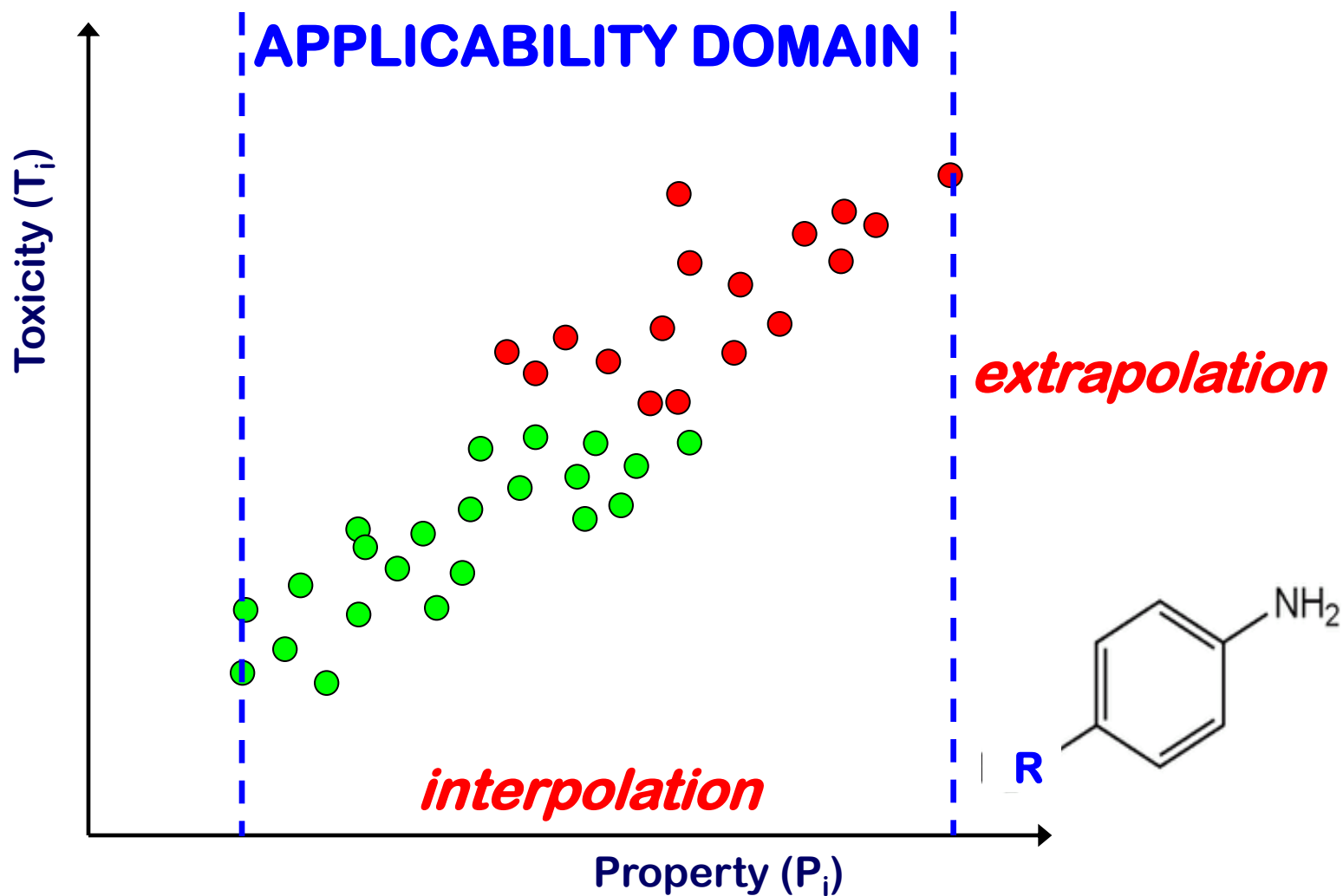


The scatter plot: the best place where explore (Q)SAR.





The scatter plot: the best place where explore (Q)SAR.



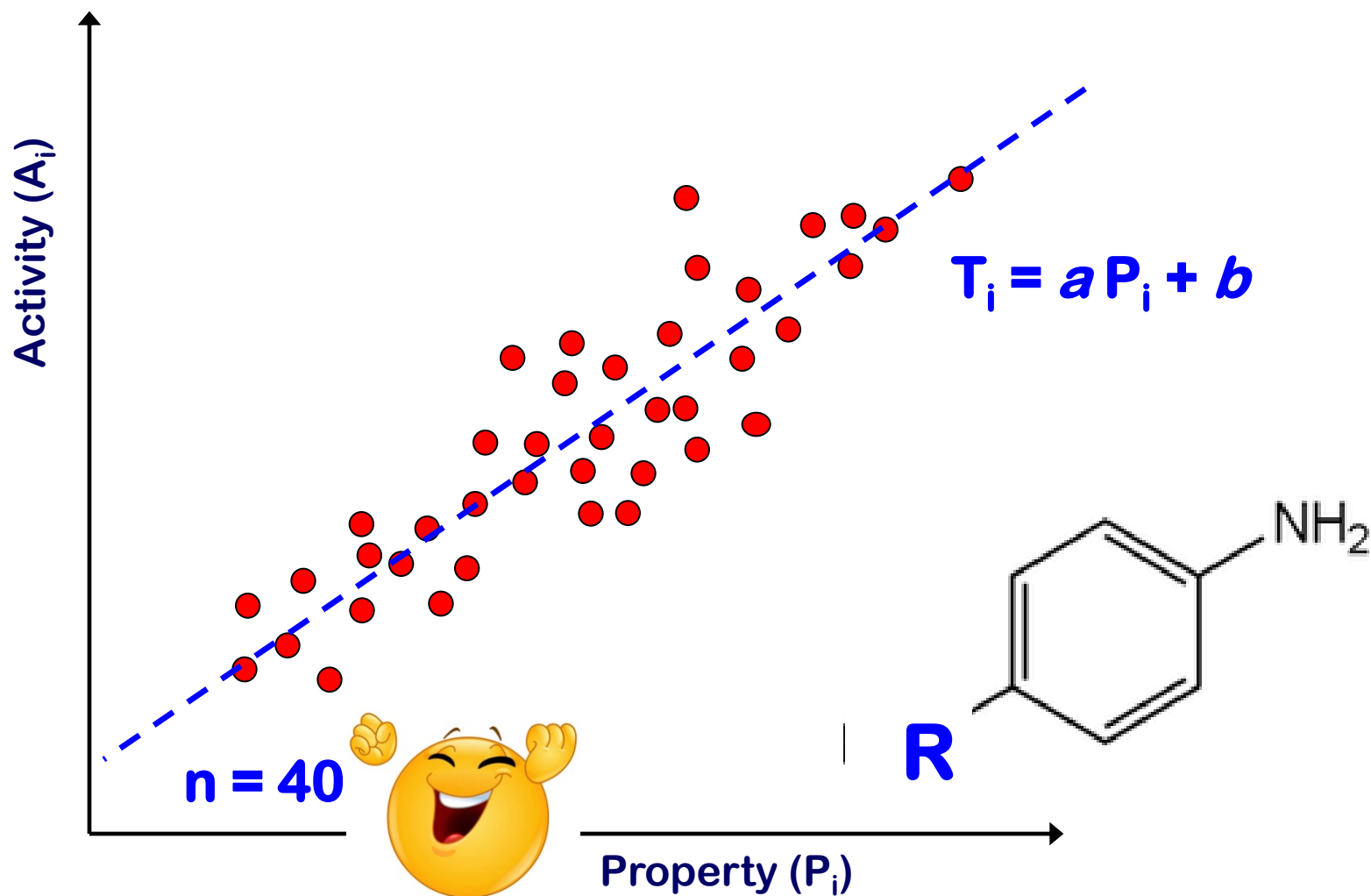


sometimes too gorgeous... the first powerfull example of AI:

- ⊙ classification models (qualitative response)
- ⊙ regression models (quantitative response)
- ⊙ ranking models (ordered response)



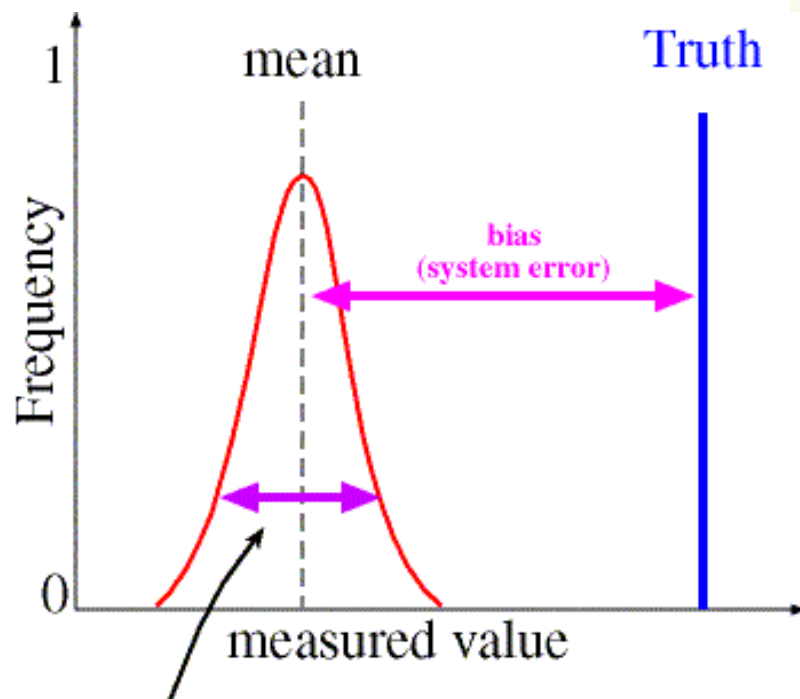
The scatter plot: the best place where explore (Q)SAR: look at this experiment





The two statistical **gold** rules do build up linear models:

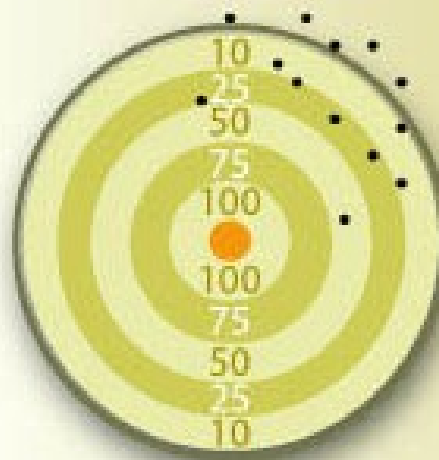
- For each independent variable (*molecular descriptor*) you need at least five (5) dependent variable values (*activities*).
- The dependent variable values (*activities*) must be accurate and precise.



Large random error



Large systematic error



Large random and
systematic error



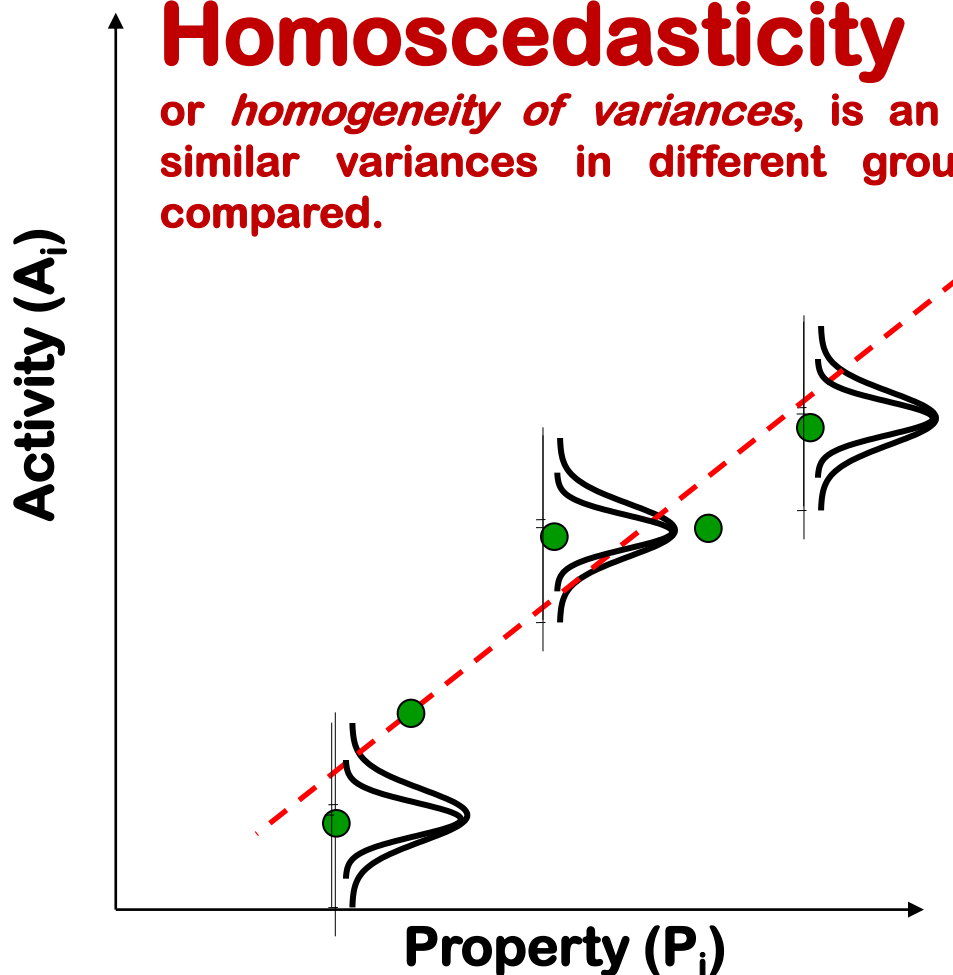
Small random and
systematic error



Now, how can we select the “good” linear model:

Homoscedasticity

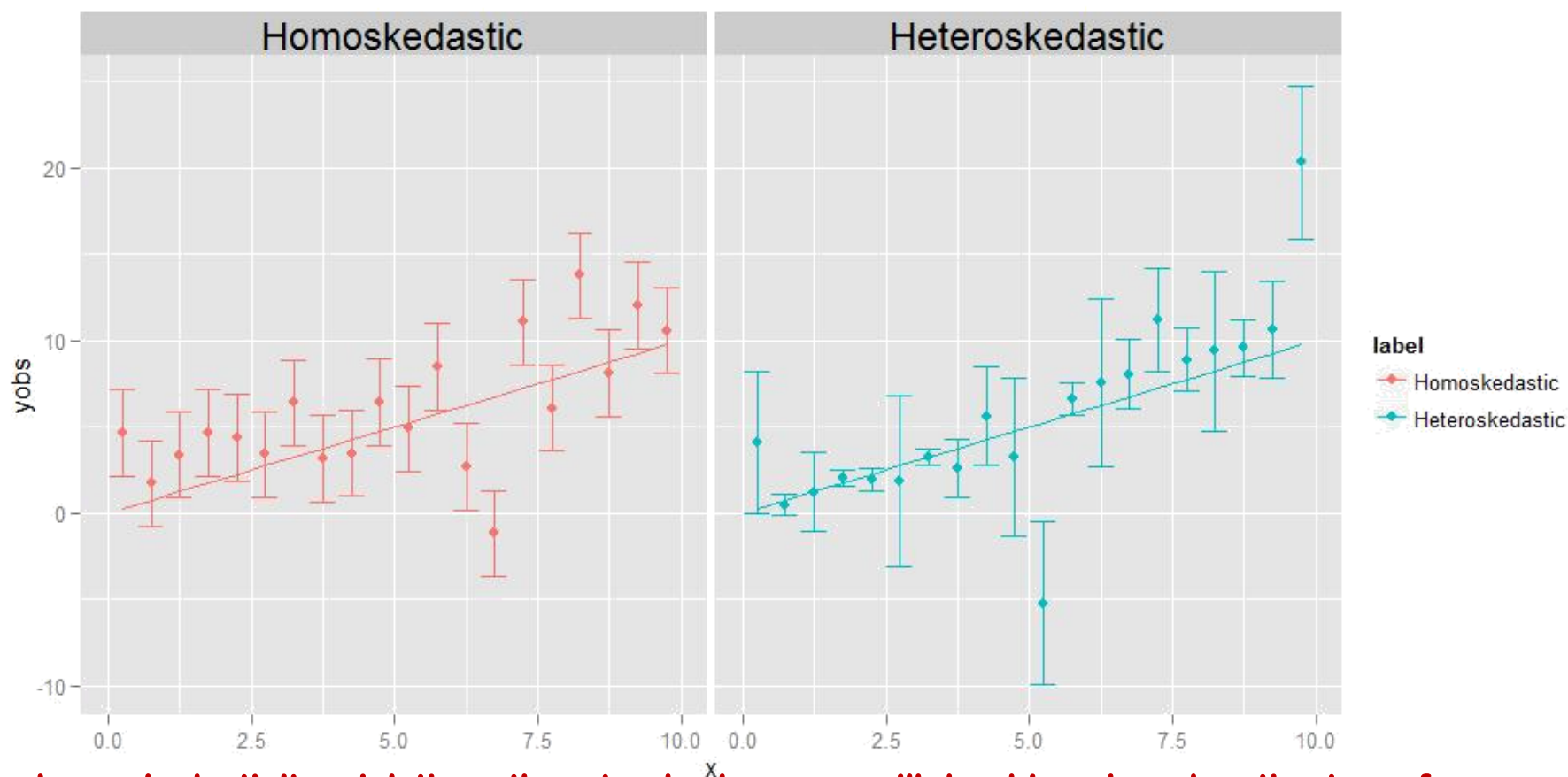
or *homogeneity of variances*, is an assumption of equal or similar variances in different groups (experiments) being compared.



credits: https://en.wikipedia.org/wiki/Homoscedasticity_and_heteroscedasticity_homogeneity_of_variance.



Now, how can we select the “*good*” linear model:

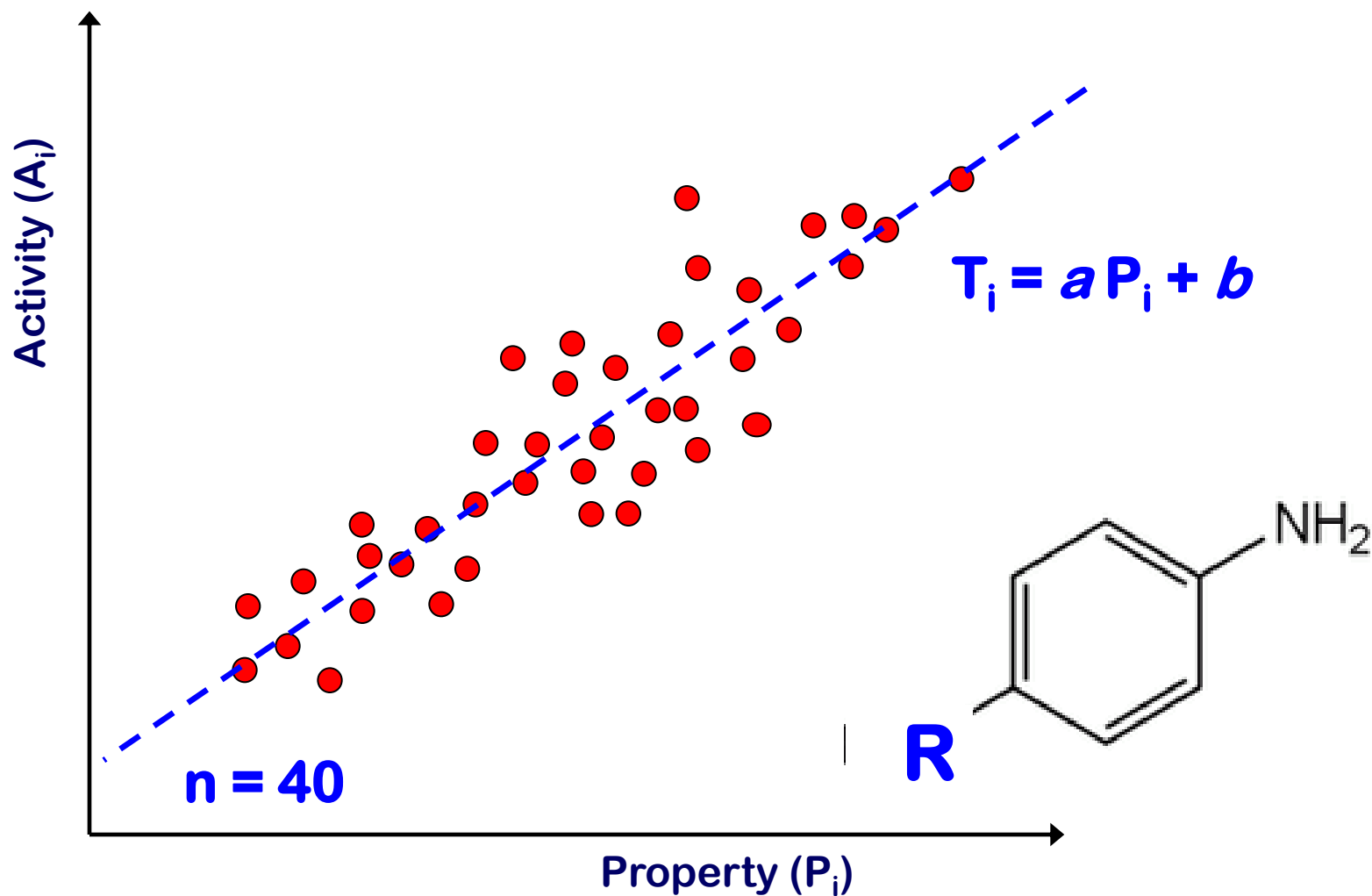


For homoskedasticity violation, the standard errors will be biased and estimates of regression coefficients will be less efficient. In practice, this usually mean overestimating the precision of your model

credits: https://en.wikipedia.org/wiki/Homoscedasticity_and_heteroscedasticity_homogeneity_of_variance.



The scatter plot: the best place where explore (Q)SAR: look at this experiment





do you remember... Least Squares Analysis?

Credits: https://en.wikipedia.org/wiki/Least_squares

1. What is a Least Squares Regression Line?

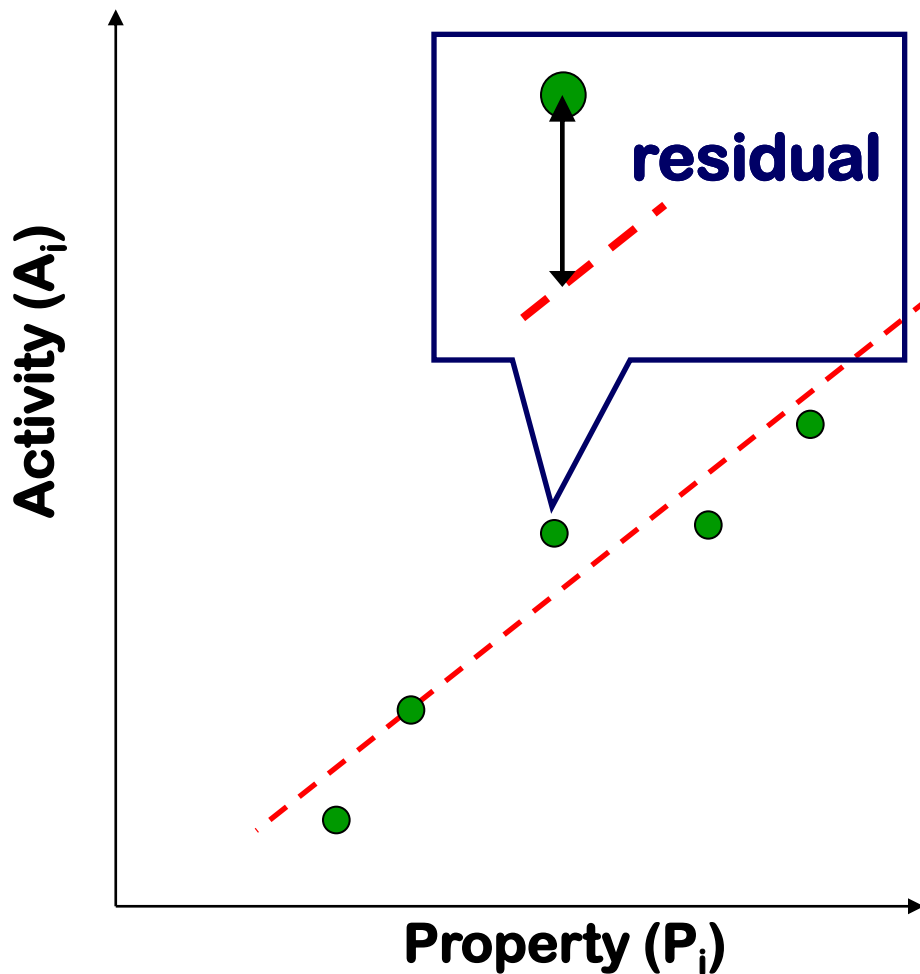
If your data shows a **linear relationship** between the X and Y variables, you will want to find the line that best fits that relationship. That line is called a **Regression Line** and has the equation $\hat{y} = a + b x$. The Least Squares Regression Line is the line that makes the vertical distance from the data points to the regression line as small as possible. It's called a "least squares" because the best line of fit is one that minimizes the variance (the sum of squares of the errors).

.



do you remember... Least Squares Analysis?

Credits: https://en.wikipedia.org/wiki/Least_squares



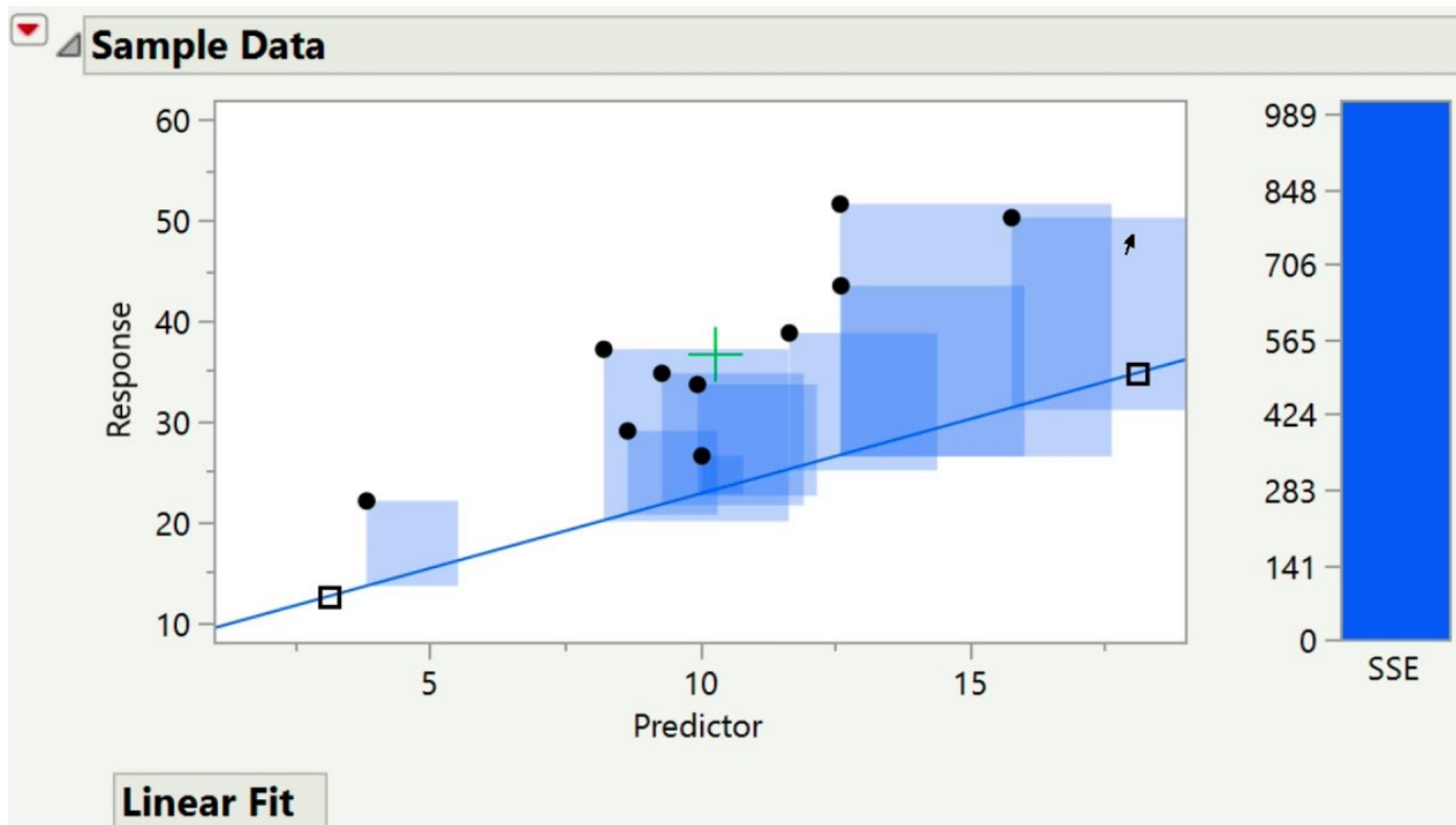
LSA is a method for linear regression that determines the values of unknown quantities in a statistical model by minimizing the sum of the **residuals**, the difference between the predicted ($\hat{y}$) and observed values (y) squared.

$$e = \hat{y} - y$$



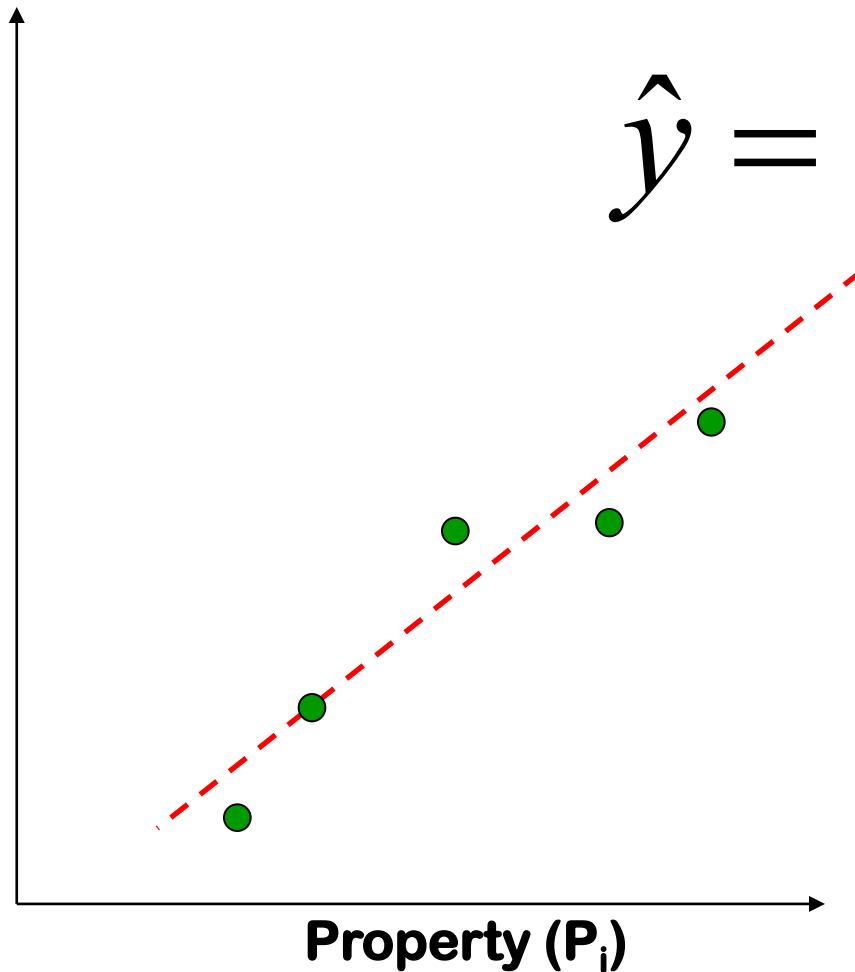
do you remember... Least Squares Analysis?

This can be a bit hard to visualize but the main point is you are aiming to find the equation that fits the points as closely as possible.



Please look this video: <https://youtu.be/bhKXKTaQ96M>

Activity (A_i)



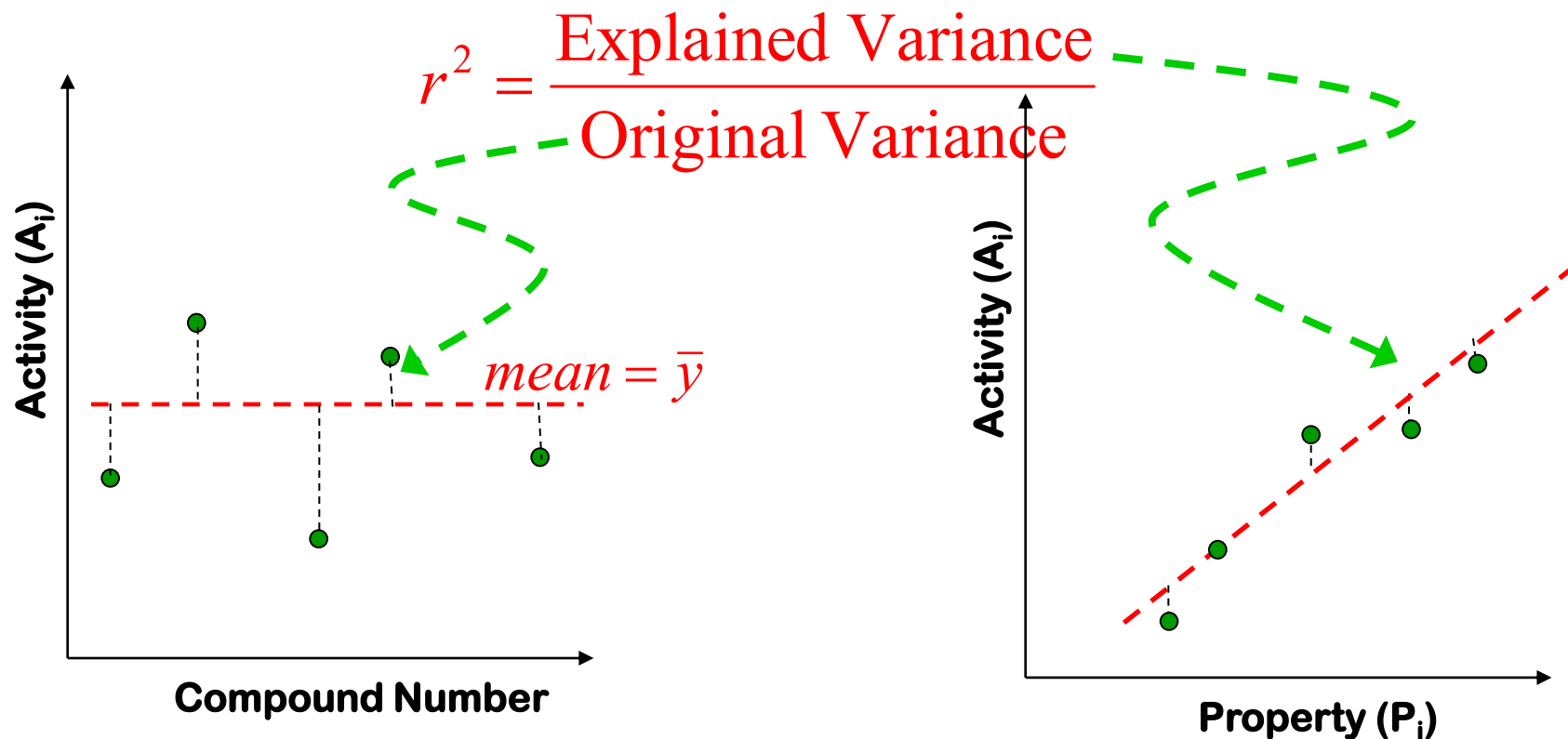
$$\hat{y} = a + bx$$

$$b = \frac{n \sum xy - \sum x \sum y}{n \sum x^2 - (\sum x)^2}$$

$$a = \frac{\sum y}{n} - b \frac{\sum x}{n}$$

Credits: https://en.wikipedia.org/wiki/Least_squares

Goodness of fit: variation in the data is quantified by the *coefficient of determination (r^2)* which measures how closely the observed data tracks the fitted regression line. Errors in either the model or in the data will lead to a bad fit. This indicator of fit to the regression line is calculated as:



Original variance = Explained variance (*i.e.*, variance explained by the equation) + Unexplained variance (*i.e.*, residual variance around regression line)

Calculating r^2

- **Original variance:**

$$TSS = \sum_{i=1}^N (y_i - \bar{y})^2$$

- **Explained variance:**

$$ESS = \sum_{i=1}^N (y_{i,calc} - \bar{y})^2$$

- **Variance around regression line:**

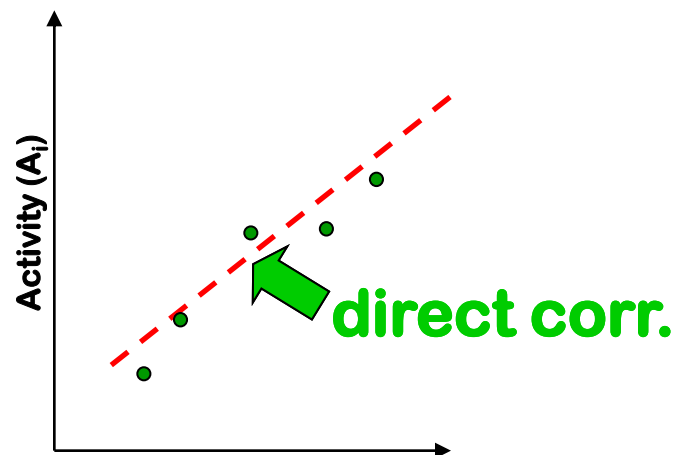
$$RSS = \sum_{i=1}^N (y_i - y_{calc,i})^2$$

$$r^2 = \frac{ESS}{TSS} \equiv \frac{TSS - RSS}{TSS} \equiv 1 - \frac{RSS}{TSS}$$

$$0 < r^2 < 1$$

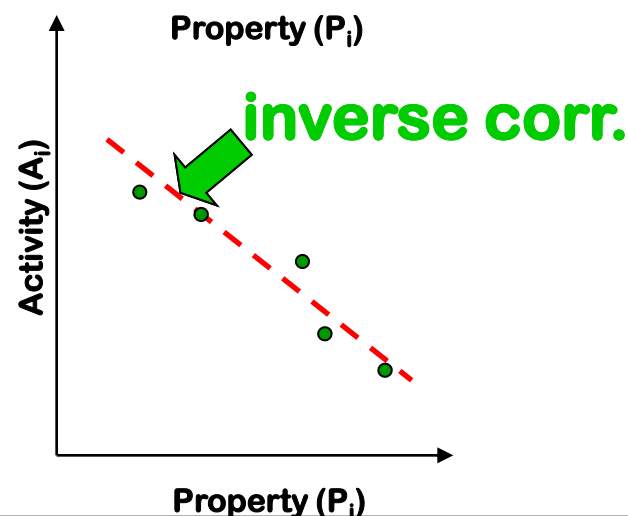
Possible values reported for r^2 fall between 0 and 1. For example: with r^2 of 0.83, you can say that 83% of the variability in activity can be explained by the different value of the selected molecular property. ***The remaining 17% of variability is due to other unexplained factors.***

Goodness of fit: the *Pearson correlation coefficient* (*r*) is the square root of r^2 expressed as a decimal. Its *size* is always between 0 and 1. The *sign* of the correlation coefficient depends on the slope of the regression line:



$$r^2 = \frac{ESS}{TSS} \quad r = \sqrt{\frac{ESS}{TSS}}$$

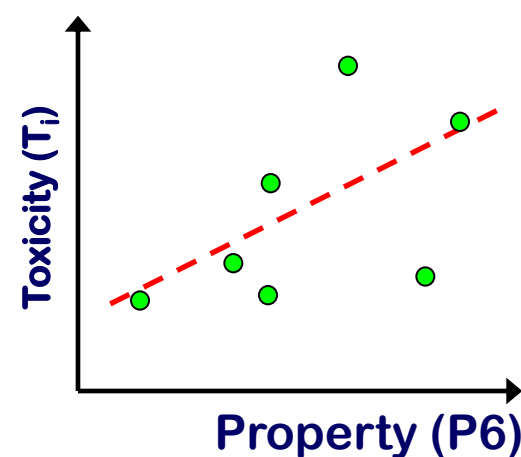
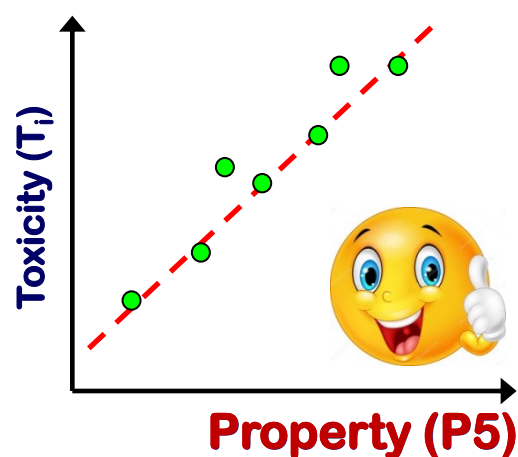
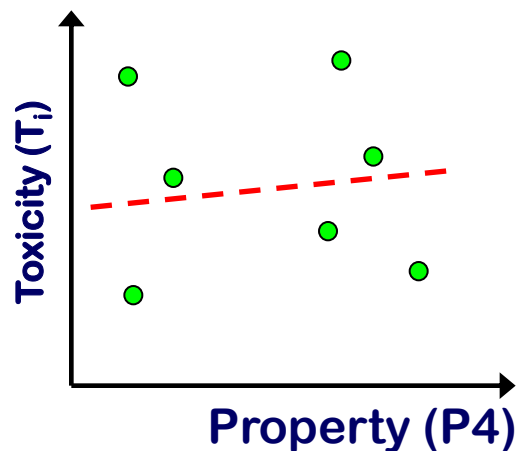
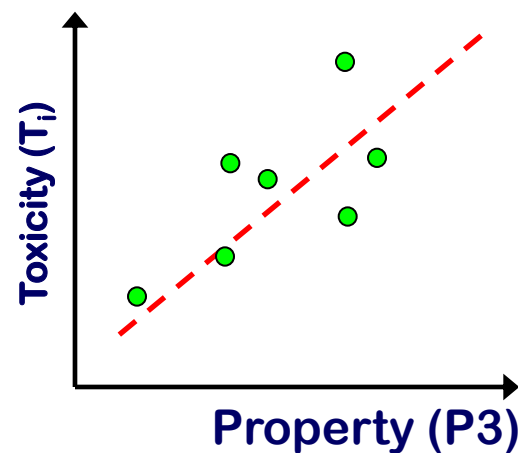
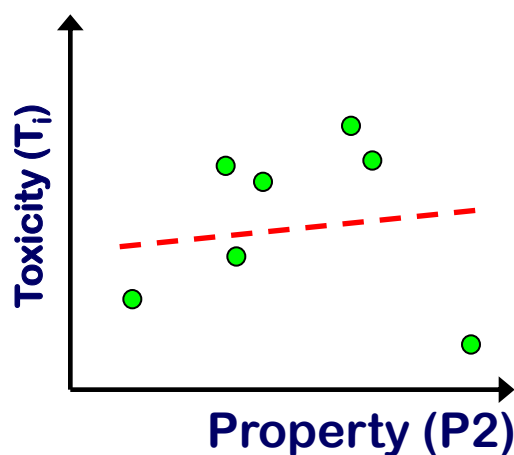
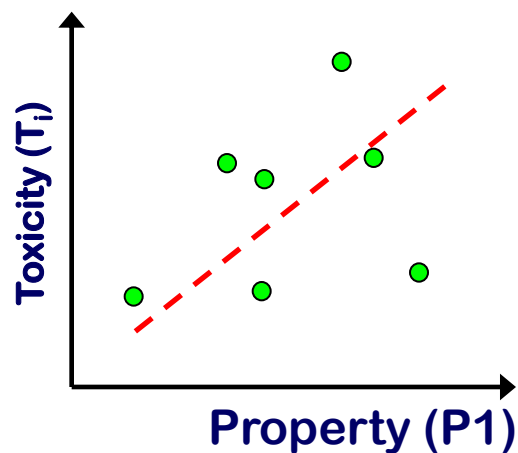
$$0 < r < 1$$



A perfect correlation of ± 1 occurs only when the data points all lie exactly on a straight line. A correlation greater than **0.8** would be described as strong, whereas a correlation less than **0.5** would be described as weak.



How we can select the *good* descriptors? *Compare r!*



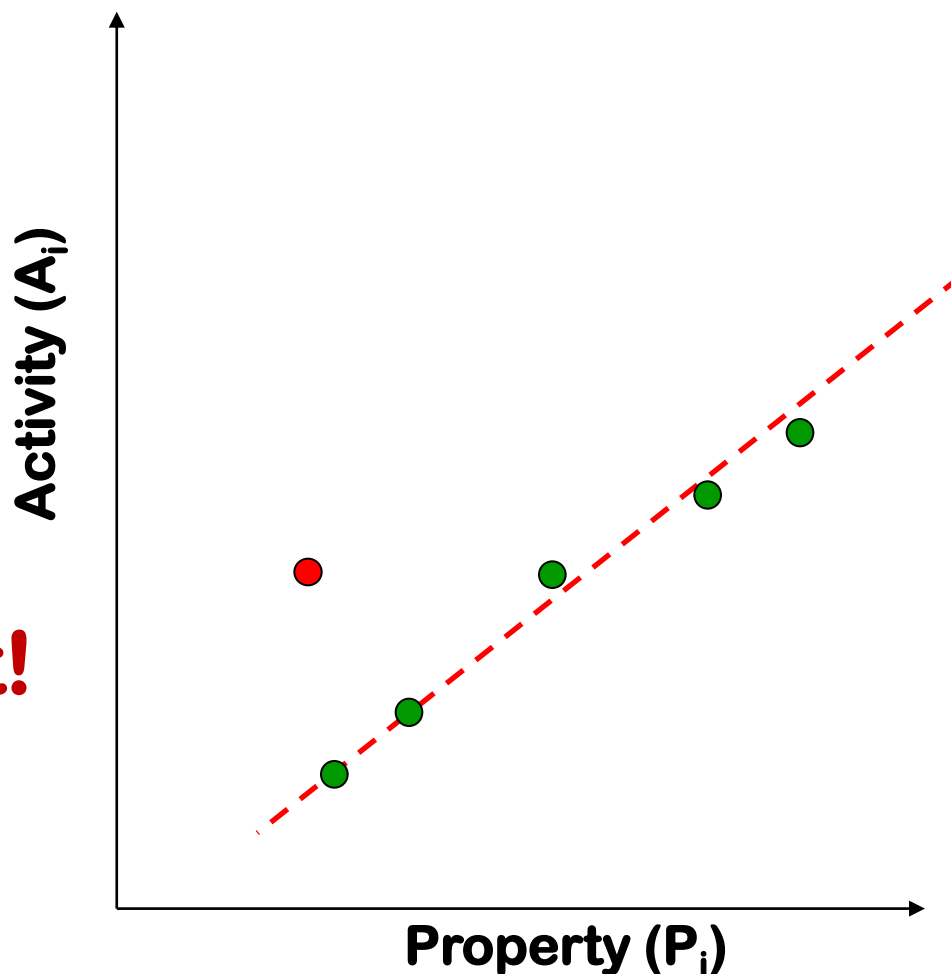
Outliers: an outlier is an observation that is numerically distant from the rest of the data.



How do we deal with them, usually?

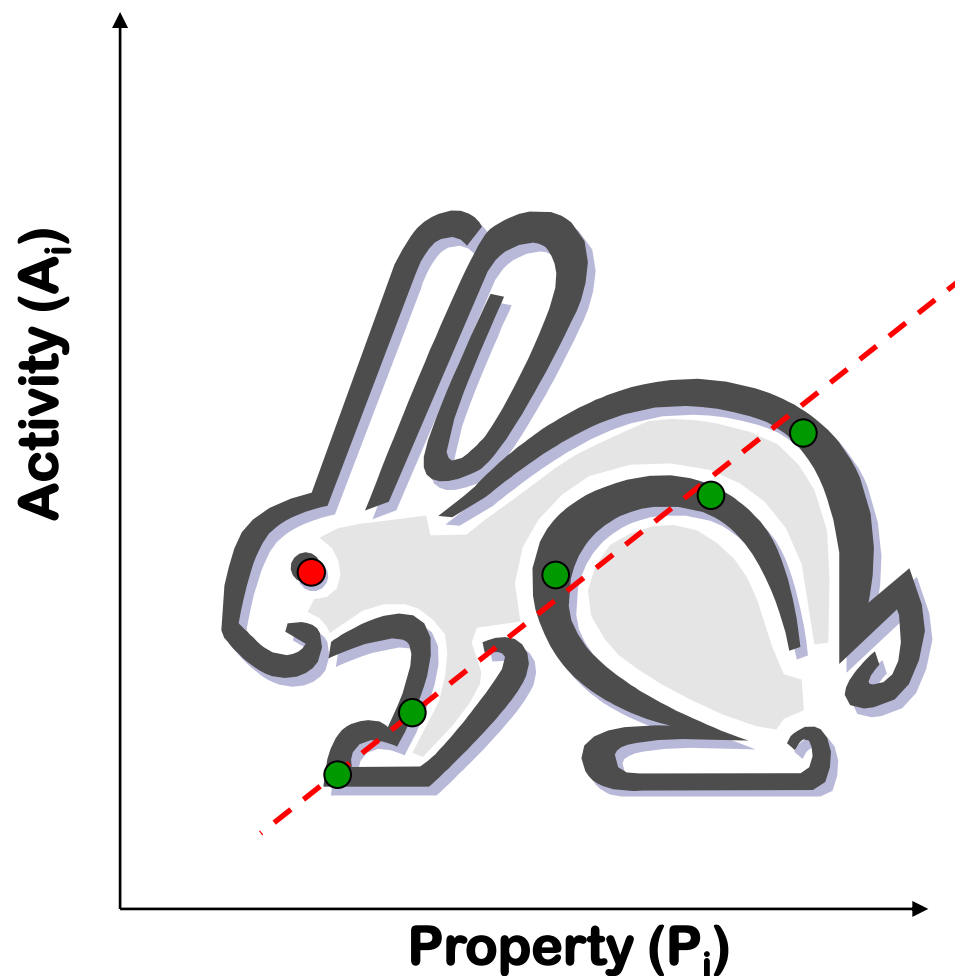


Remove it!
Shift it!





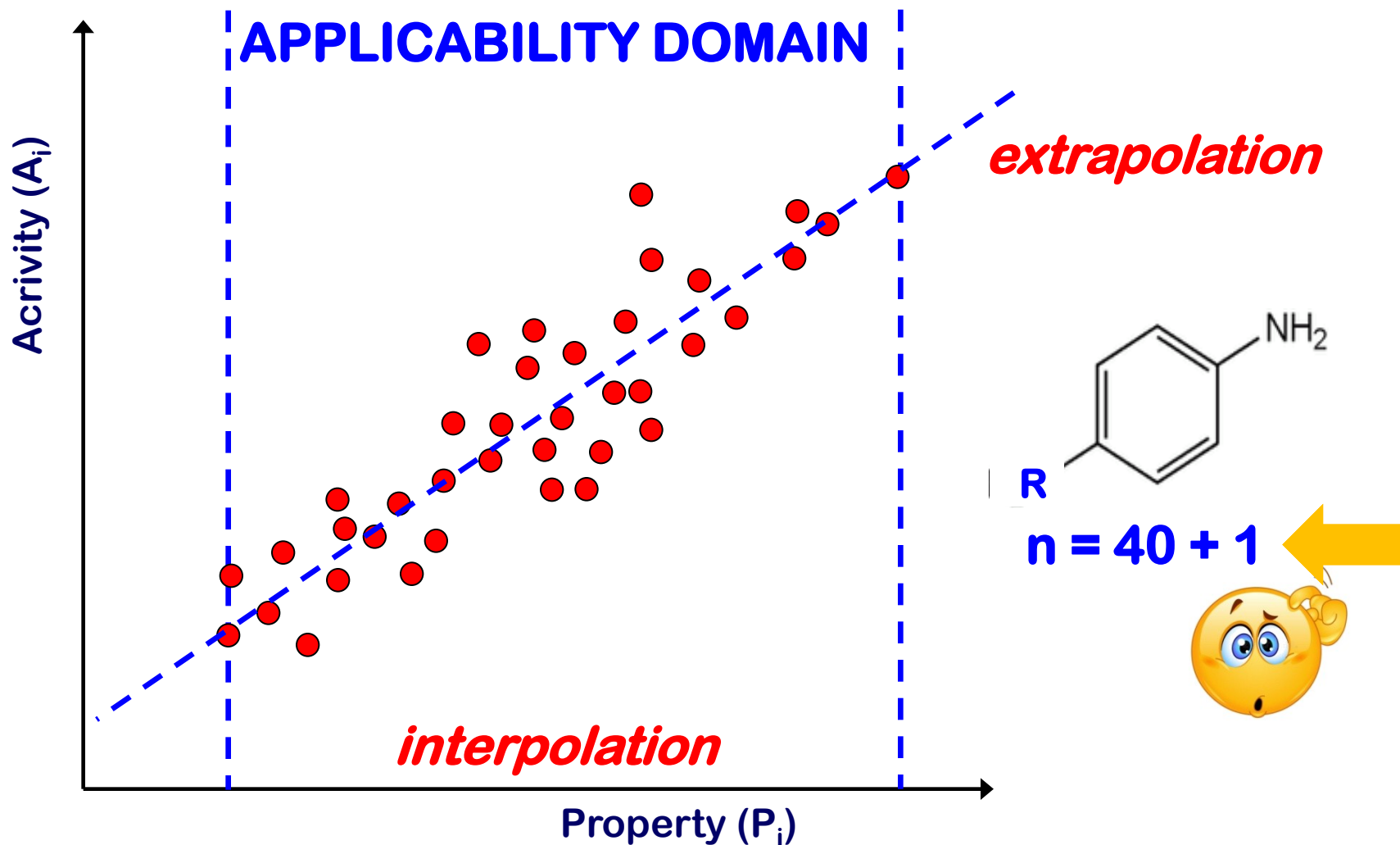
Be carefull...



... the rabbit (pattern) is out there!!!

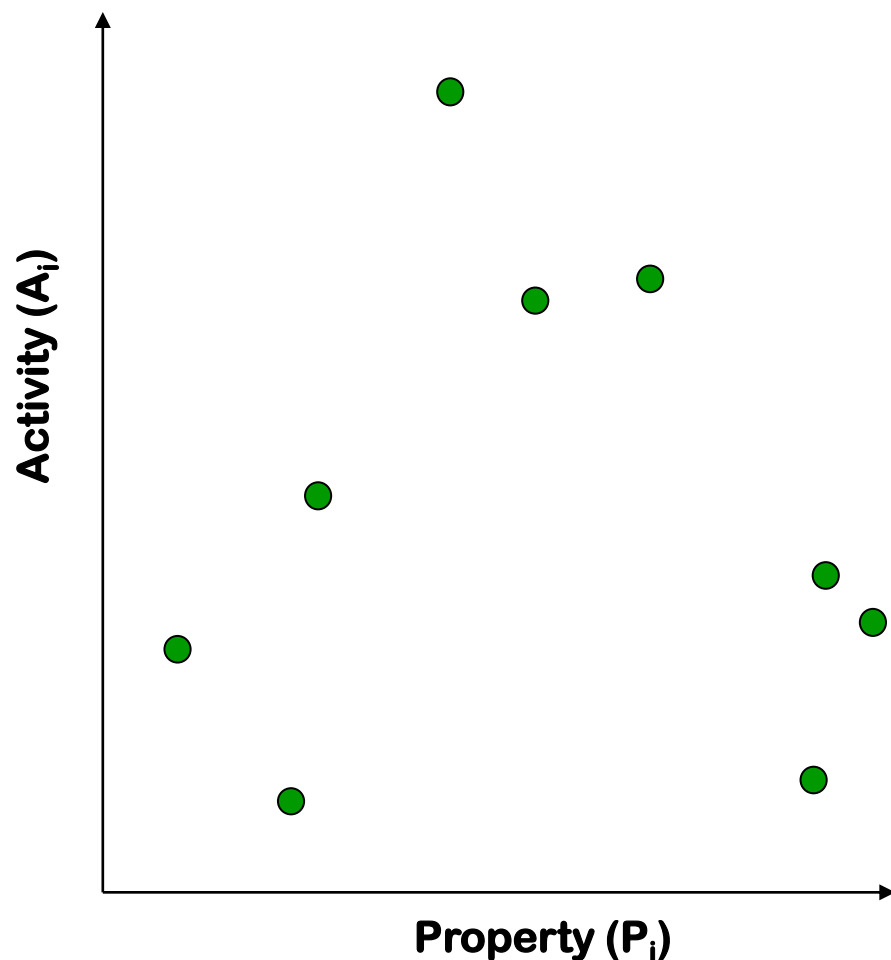


The scatter plot: the best place where explore (Q)SAR.





Cross-validation (CV) for detecting and preventing overfitting!

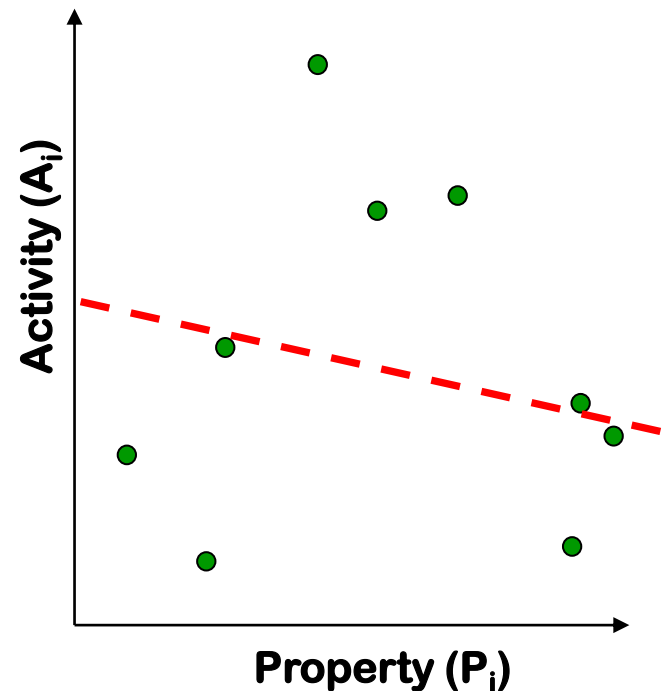


$$y = f(x) + \text{noise}$$

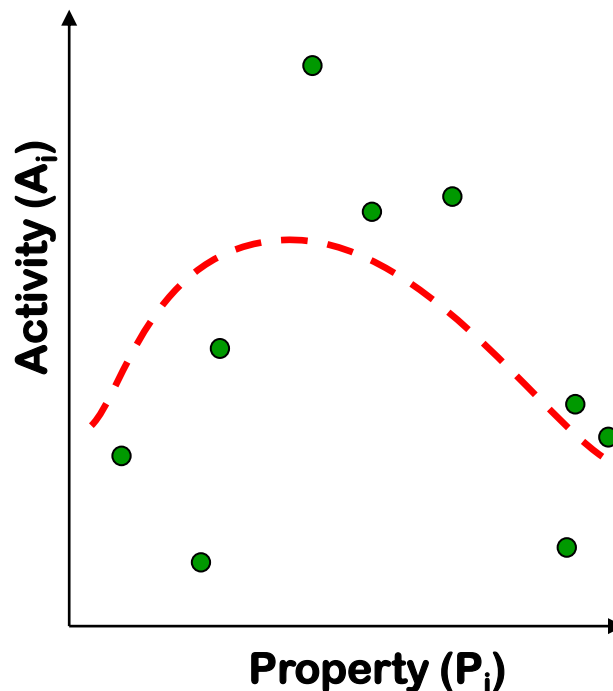
How we can deal with these data?

Let's consider three different approaches...

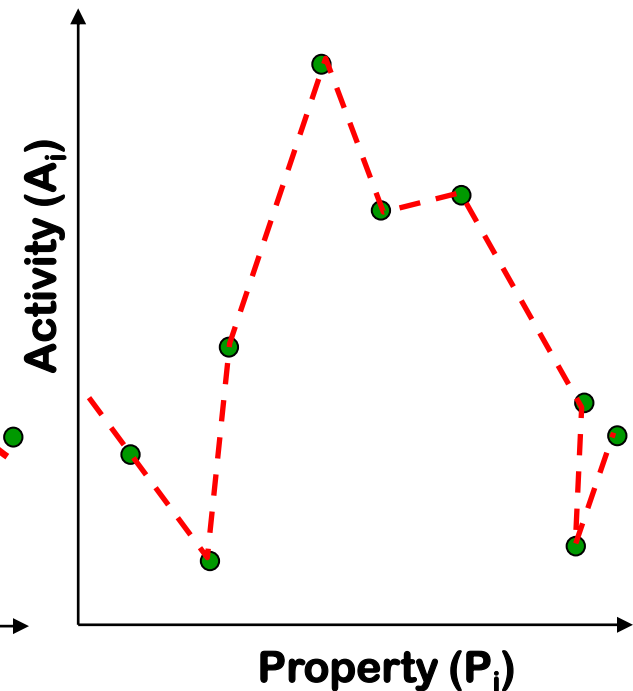
Linear



Quadratic



Joint-the-dots

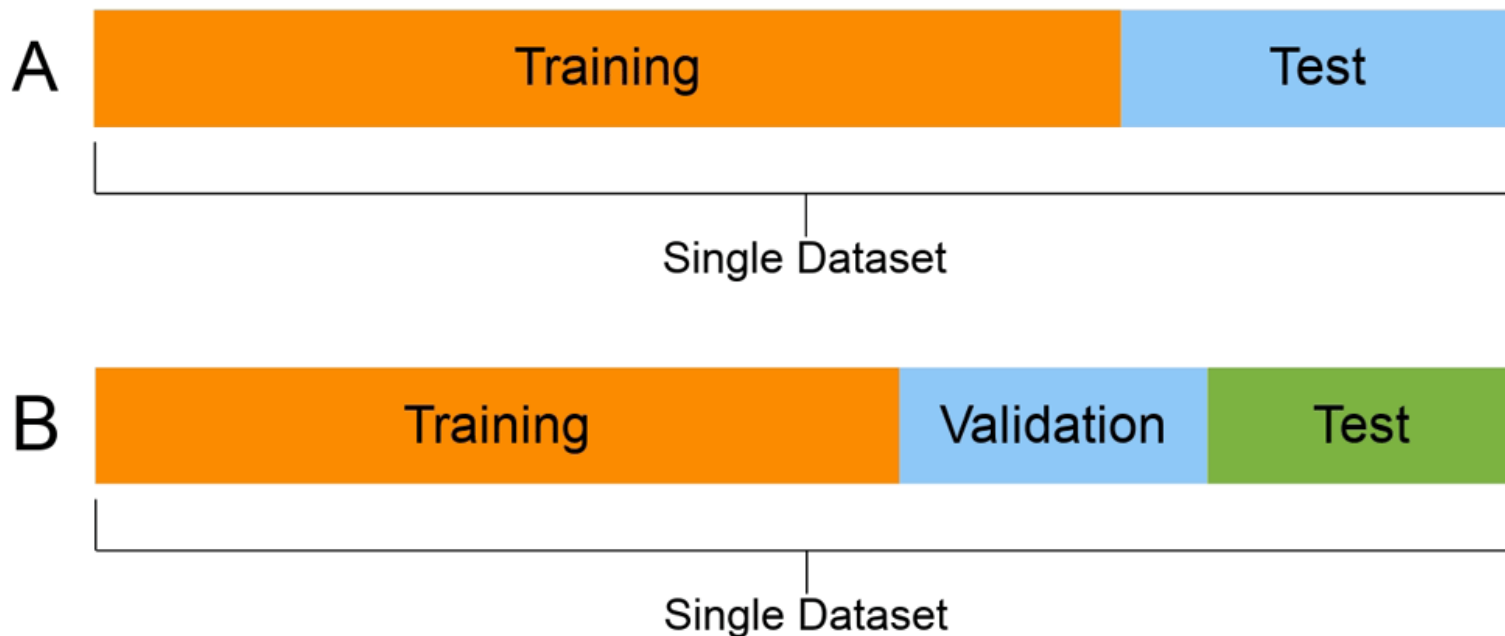


Also known as
*piecewise linear non
parametric regression...*
if that makes you feel
better!!!

**“How well are you going to predict future data
drawn from the same distribution?”**

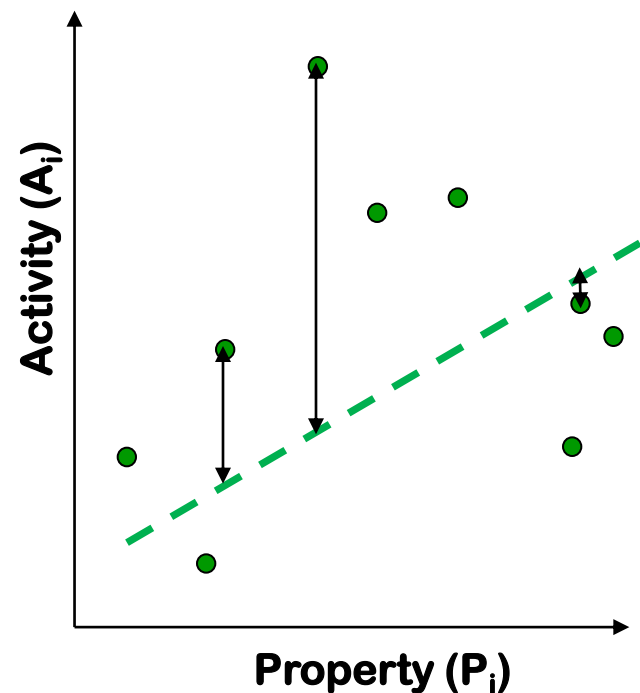


The “*test set*” method...looks this:





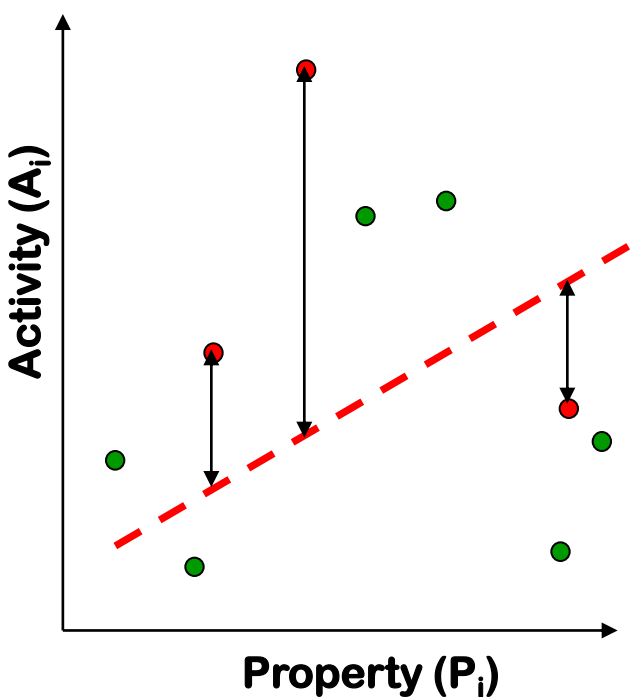
The “*test set*” method...looks this:



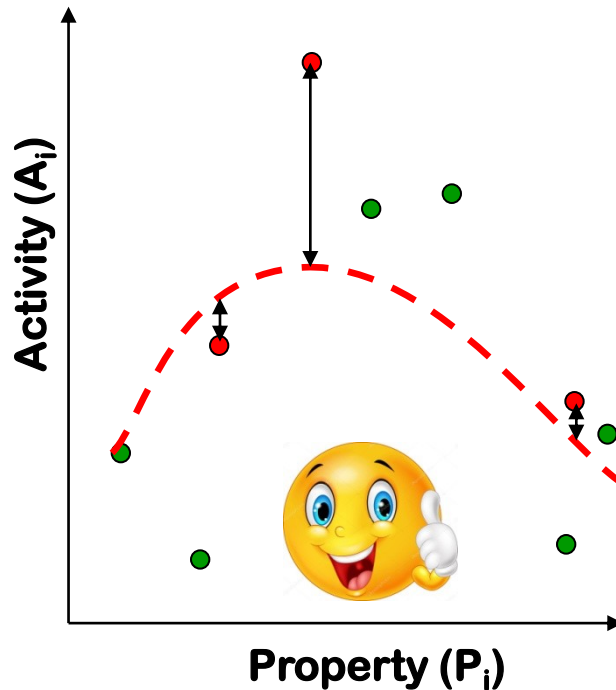
1. Randomly choose 30% of the data to be in a ***test set***;
2. The remainder is a ***training set***;
3. Perform your regression on the ***training set***;
4. Estimate your future performance with the ***test set***.

Mean Squared Error (MSE)

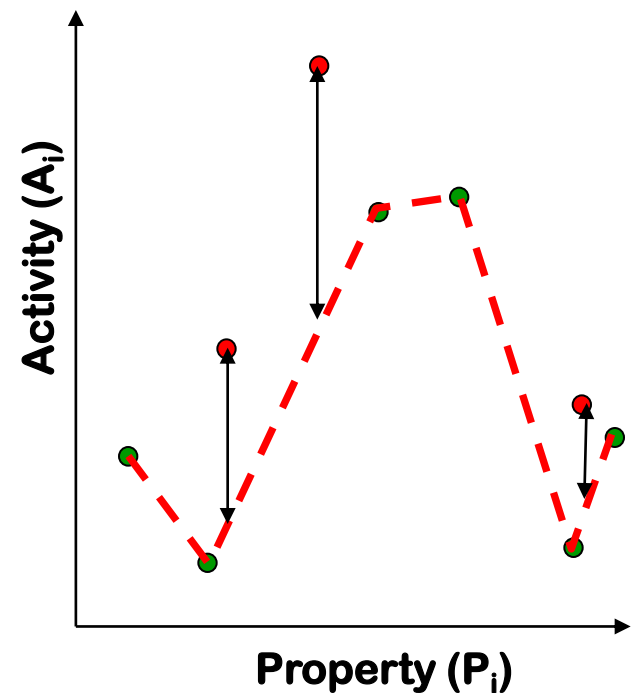
$$MSE = \frac{\sum (x_i - \bar{x})^2}{n}$$



MSE = 2.4



0.9



2.2

Good news:

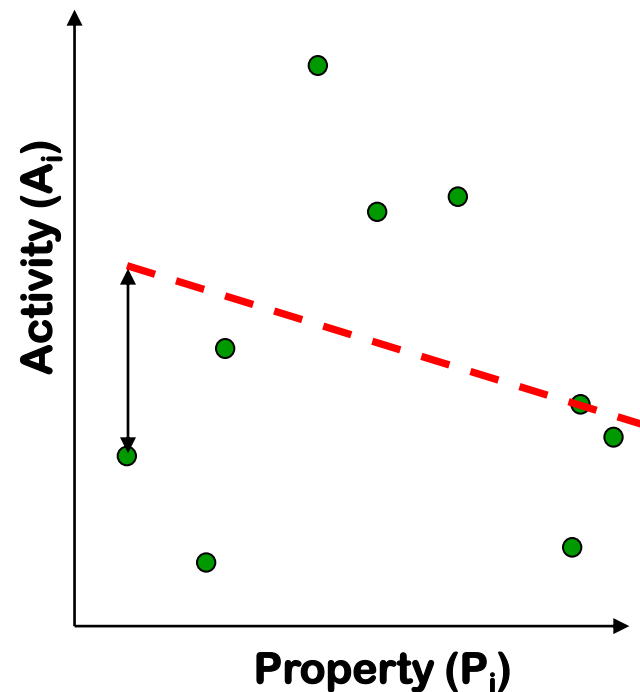
- Very very simple;
- Can then simply choose the method with the best “*test set*” score.

Bad news:

- Wastes data: we get an estimate of the best method to apply to 30% less data;
- If we don't have much data, our test-set might just be lucky or unlucky.



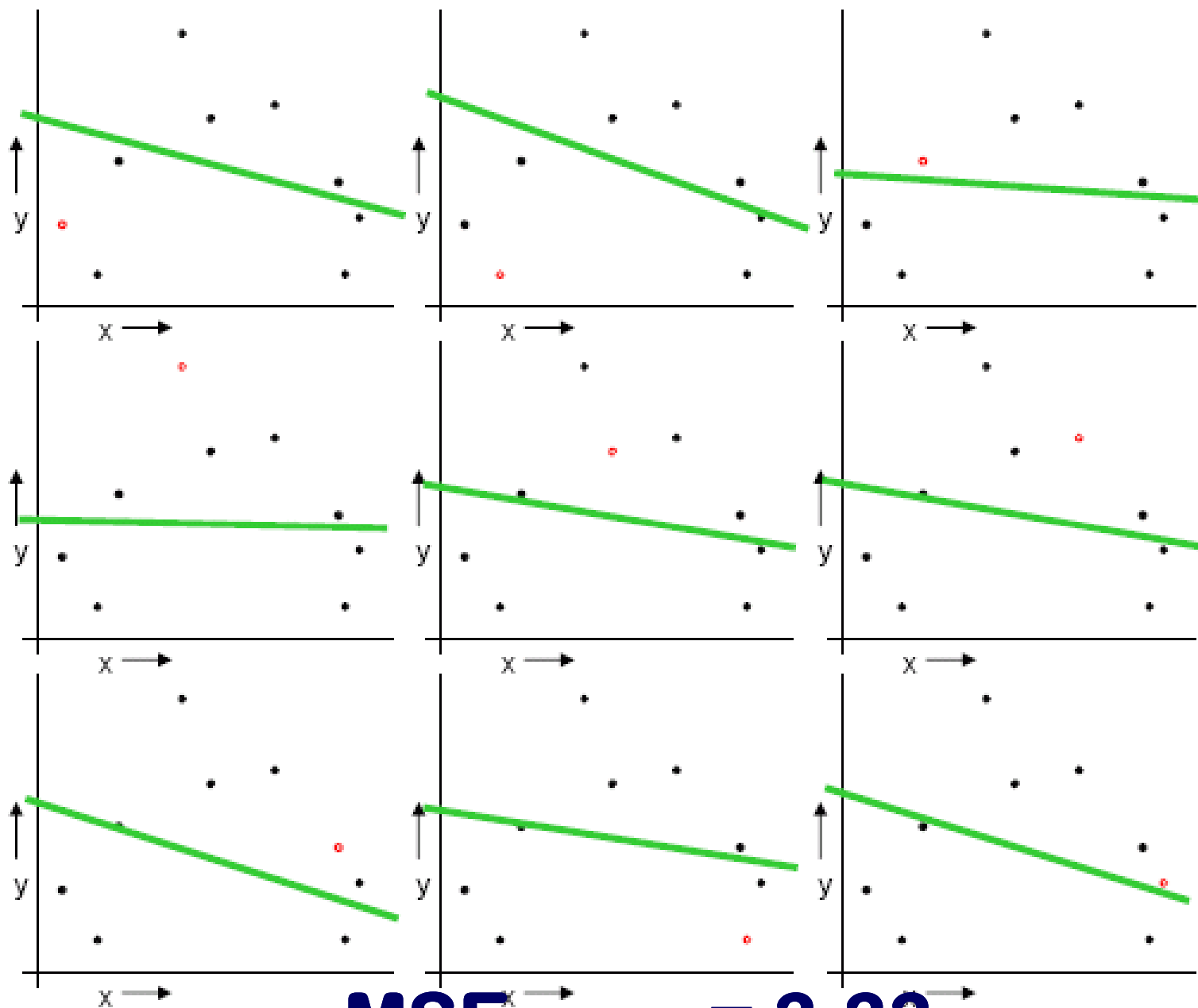
or “*LOOCV*” (Leave-One-Out Cross Validation) method... looks this:



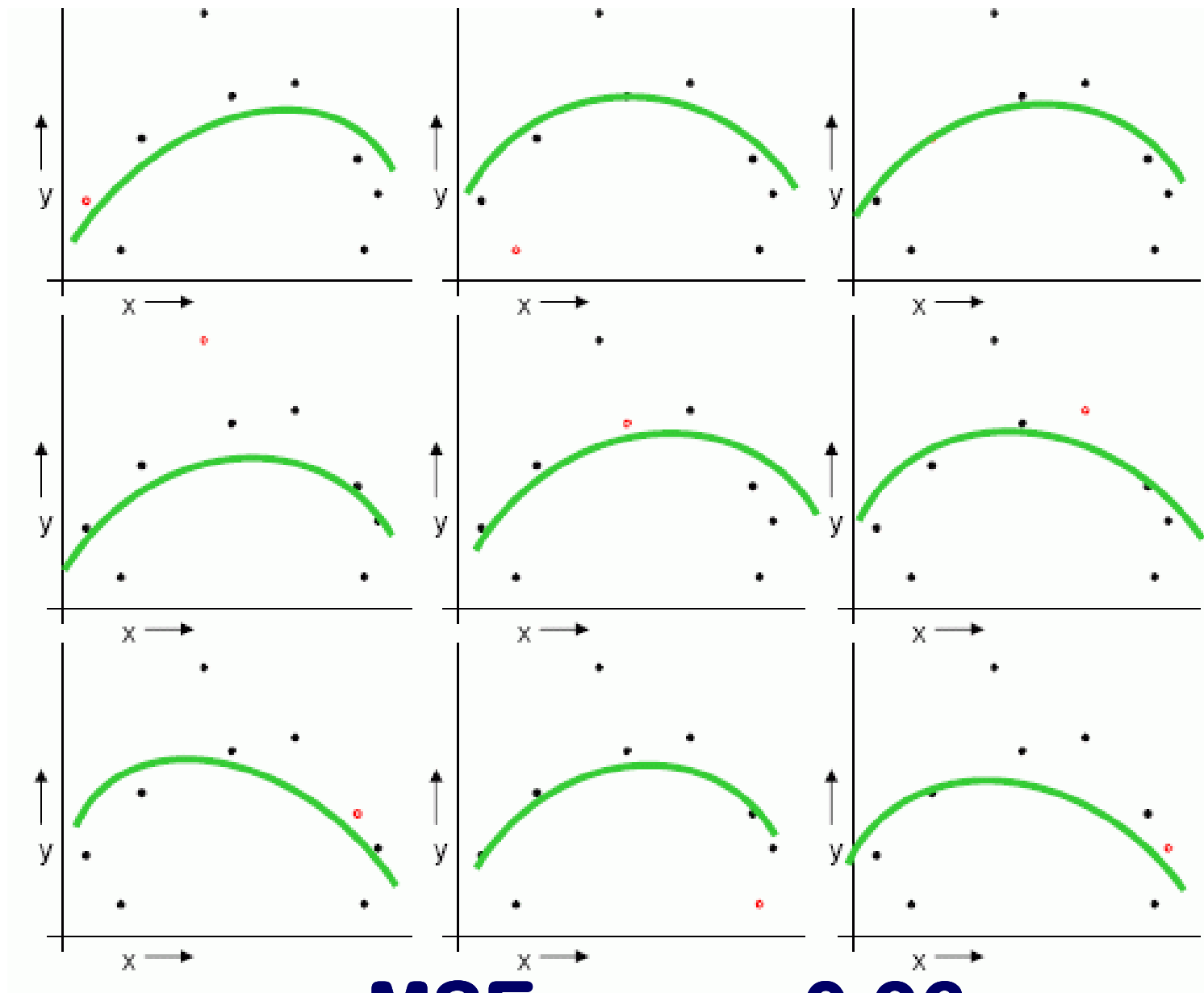
For each data consider this loop:

- 1. Select the first (x_i, y_i) data;**
- 2. Temporary remove (x_i, y_i) from the data set;**
- 3. Train on the remaining $n-1$ datapoints;**
- 4. Note your error (x_i, y_i) ;**

When you've done all points, report the mean squared errors (MSE).

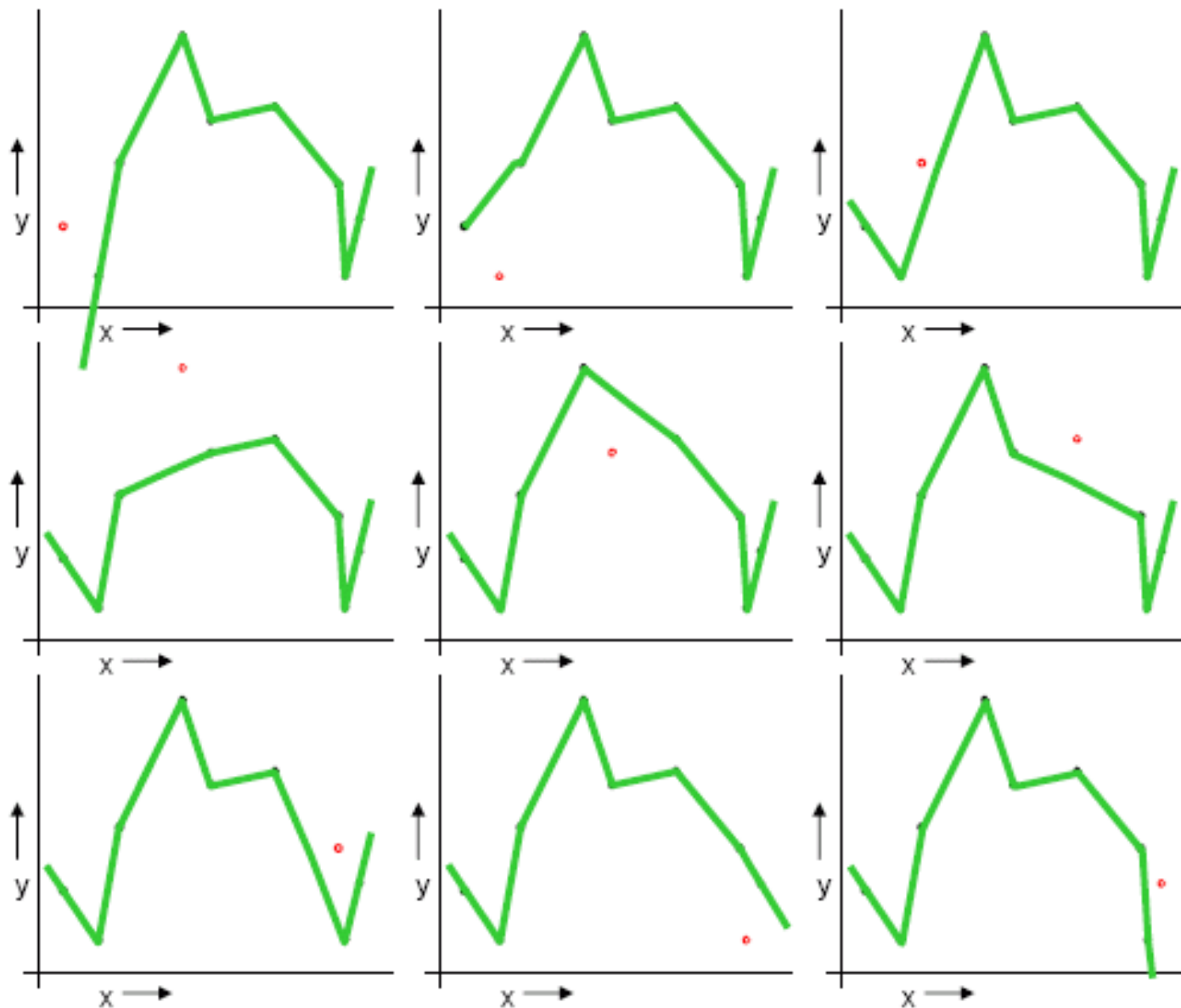


$$\text{MSE}_{\text{Loocv}} = 3.33$$



$$\text{MSE}_{\text{Loocv}} = 0.96$$





$$\text{MSE}_{\text{LOOCV}} = 2.12$$



So... which kind of validation?

	Downside	Upside
Test-set	Variance: unreliable estimate of future performance	Time cheap
Leave-one-out	Time expensive. Has some weird behaviour	Doesn't waste data



Here a possible work-flow:

