

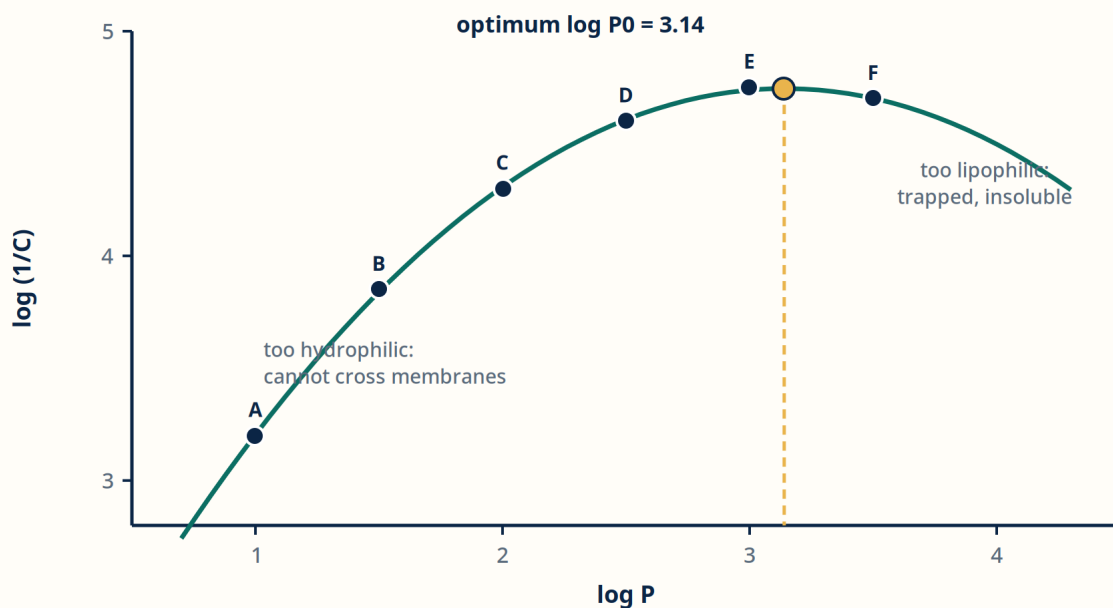
2.1 The Hansch approach: putting the parameters together

Hansch's insight was that biological activity is the product of a journey and an arrival. The molecule must first travel, through aqueous and lipid compartments in series, to the target (governed above all by lipophilicity), and must then interact with the target (governed by electronic and steric fit). The **Hansch equation** therefore combines the parameter families in one regression, of the general form: $\log(1/C)$ equals a coefficient times $\log P$, plus a coefficient times σ , plus a coefficient times a steric term, plus a constant. Each term reads like a sentence. The $\log P$ term prices transport and hydrophobic binding. The σ term prices the electronic demands of the interaction. The steric term prices fit. The constant absorbs the activity of the parent scaffold. Because the parameters were designed to be independent, the equation can in principle separate effects that a qualitative SAR tangles together, and tell the chemist, for the next analogue, whether to raise lipophilicity, donate electrons, or cut bulk. Hansch analysis needs a congeneric series, a single scaffold with varied substituents, consistent biological data measured the same way, and enough compounds to support the number of fitted terms, the working rule being at least five compounds per parameter.

2.2 The parabolic model and optimum lipophilicity

Activity does not rise with $\log P$ forever. A very hydrophilic compound cannot cross membranes; a very lipophilic one cannot dissolve, gets trapped in the first membrane it meets, binds plasma proteins and fat, and never arrives. Hansch therefore added a squared term: $\log(1/C)$ equals minus a times $(\log P)^2$, plus b times $\log P$, plus the other terms and the constant. Plotted against $\log P$ at fixed other parameters, the equation is an upside down parabola with an optimum, $\log P_0$, at its peak, found by setting the derivative to zero: $\log P_0$ equals b divided by $2a$. The bilinear model is a refinement often used in its place, rising linearly then falling linearly with a break near the optimum, but the parabola is the syllabus model and the teaching point stands: there is a best lipophilicity for a series acting by this mechanism, often near $\log P$ 2 for centrally acting drugs that must cross the blood brain barrier, and designing beyond the optimum in either direction wastes synthesis. A related practical tool is the **Craig plot**, which scatters substituents on axes of π against σ so the chemist can pick a test set of substituents spread across all four quadrants, varying lipophilicity and electronics independently, instead of testing ten variations of the same idea.

The Hansch parabola: activity peaks at an optimum log P



© 2026 PharmacoLearn, pharmacolearn.fyi

Figure: The worked example of section 2.3 plotted: activity rises with log P, peaks at the optimum log P of 3.14, and falls as compounds grow too lipophilic to reach the target.

2.3 Worked example: deriving a 2D QSAR equation

Consider six analogues of one scaffold with measured log P and activity log (1/C), the kind of small teaching set used to show the method (real projects use far more compounds):

Compound	log P	Observed log (1/C)
A	1.0	3.20
B	1.5	3.85
C	2.0	4.30
D	2.5	4.60
E	3.0	4.75
F	3.5	4.70

Step 1, fit a straight line. Least squares regression of activity on log P gives: log (1/C) equals 0.600 log P plus 2.883, with n equal to 6, r equal to 0.926, r squared equal to 0.857, s equal to 0.257 and F equal to 23.9. Step 2, look at the residuals, observed minus predicted: minus 0.283, plus 0.067, plus 0.217, plus 0.217, plus 0.067, minus 0.283. The pattern is systematic, negative at both ends and positive in the middle, the fingerprint of a curve forced into a line. Step 3, add the squared term, as section 2.2 prescribes: log (1/C) equals minus 0.336 (log P) squared plus 2.111 log P plus 1.429, with n equal to 6, r squared equal to 0.999 and s equal to 0.011; the residuals are now tiny and patternless. Step 4, read the model. The optimum is log P equals 2.111 divided by (2 times 0.336), which is 3.14, and the predicted peak activity is about 4.74, matching compound E, after which activity falls, exactly as compound F shows. Step 5, use and respect the model. It predicts that a new analogue at log P 3.1 should be near the peak of the series, and that money spent pushing log P to 4 would be wasted; it says nothing reliable about log P 5, outside the fitted range, because extrapolation is where QSAR models go to die. The example is deliberately small; the

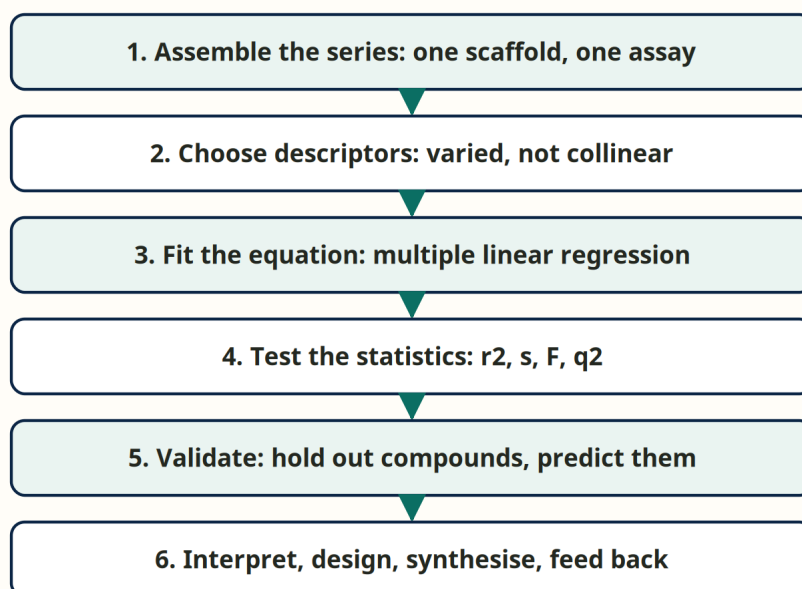
statistical cautions of sections 2.9 and 2.10 apply with full force.

2.4 The steps of building a QSAR model

1. Assemble the series same scaffold, same assay, reliable data
|
v
2. Choose descriptors parameters that vary and are not collinear
|
v
3. Fit the equation multiple linear regression
|
v
4. Test the statistics r^2 , s , F , cross validated q^2
|
v
5. Validate hold out compounds, predict, compare
|
v
6. Interpret and design predict new analogues, synthesise, retest

Every step has a failure mode this course repeats like a refrain: mixed assays poison step 1, correlated descriptors poison step 2, too many terms poison step 3, skipped validation poisons step 5, and ignoring the model's range poisons step 6.

Building a QSAR model: six steps, six failure modes



© 2026 PharmacolLearn, pharmacollearn.fyi

Figure: The QSAR workflow of section 2.4; the validation steps, 4 and 5, are the ones that separate a model from a mirage.

2.5 Free Wilson analysis: activity as a sum of contributions

Free and Wilson (1964) took the opposite route to Hansch. Ignore physicochemical parameters entirely; assume instead that **each substituent at each position contributes a fixed amount to the activity**, whatever the rest of the molecule, and that contributions add. The model is written: $\log(1/C)$ equals the sum over positions and substituents of (indicator for that substituent times its contribution) plus μ , the activity of the unsubstituted parent. Each compound is coded by a row of ones and zeros recording which

substituents it carries, and regression solves for the contributions. A small worked miniature: suppose the parent of a series has $\log(1/C)$ equal to 4.00, regression assigns a chloro group at position R1 a contribution of plus 0.42, and a methoxy group at position R2 a contribution of minus 0.18. A compound carrying both is predicted at 4.00 plus 0.42 minus 0.18, that is 4.24, and a compound with chloro at R1 alone at 4.42. The output is a league table of substituents at each position, ranked, which is exactly the shopping list a chemist wants. The method's price: it can only talk about substituents it has seen, at positions it has seen, and it says nothing about why a contribution is positive. It also needs a well populated data matrix; a substituent appearing once gets a contribution fitted to a single measurement.

2.6 Hansch and Free Wilson compared, and the mixed approach

Aspect	Hansch analysis	Free Wilson analysis
Basis	Physicochemical parameters ($\log P$, σ , steric)	Indicator variables for substituents
Explains mechanism	Yes, coefficients interpret transport and binding demands	No, ranks contributions without explaining them
Predicts new substituents	Yes, if their parameters are known	No, only substituents present in the training set
Data needs	Parameter values for each substituent; fewer compounds per idea	Many compounds to populate each position and substituent
Fails when	The series spans more than one binding mechanism	Additivity breaks down (substituents interact)

The two are mathematically related: if substituent contributions happen to be linear functions of Hansch parameters, a Free Wilson model reduces to a Hansch model, and de novo, the founders showed the approaches are special cases of one framework. The **mixed approach** therefore combines them: Hansch parameters for what they can explain, plus indicator variables for structural features (a hydrogen bond donor here, a ring closure there) that parameters miss. Modern descriptor based QSAR is the mixed approach grown up: hundreds of calculated descriptors in place of the classic handful, with the same duties of validation.

2.7 3D QSAR: CoMFA and CoMSIA

2D methods never ask where in space a substituent sits. **3D QSAR** does. In **Comparative Molecular Field Analysis (CoMFA, Cramer 1988)**, the molecules of a series are first aligned in space on a common frame, a step on which everything depends. A three dimensional grid is laid around them, and at each grid point a probe atom measures two fields: the steric field (van der Waals interaction energy) and the electrostatic field (Coulomb interaction with a unit positive charge). Thousands of grid values become the descriptors, and **partial least squares (PLS)** regression, built to handle many correlated descriptors, fits them to activity. **CoMSIA** (Comparative Molecular Similarity Indices Analysis) extends the idea with smoother similarity based fields, adding hydrophobic and hydrogen bond donor and acceptor fields alongside steric and electrostatic, and is less sensitive to exactly where the grid cut offs fall. Both methods answer the question 2D QSAR cannot: not just which substituent, but where in space bulk, charge, or a hydrogen bond is wanted.

2.8 Reading contour maps

The product of a 3D QSAR run is a **contour map** drawn around a reference molecule. By the standard CoMFA colour convention, **green** contours mark regions where added steric bulk raises activity and **yellow** regions where bulk lowers it; **blue** contours mark regions where a positive charge, or an electron poor group, is favoured, and **red** regions where negative charge or an electron rich group is favoured. CoMSIA

adds colours for hydrophobic and hydrogen bonding fields. Reading a map is the design act itself: a green contour beyond a methyl group says grow the molecule there; a yellow shell around a ring says the pocket is tight; a red contour near a carbonyl says the target wants electron density at that point in space. Contour map analysis carries two warnings. The map is only as good as the alignment, and a wrong alignment manufactures persuasive nonsense. And the map interpolates within the chemistry tested; it cannot describe a binding feature no training molecule ever probed.

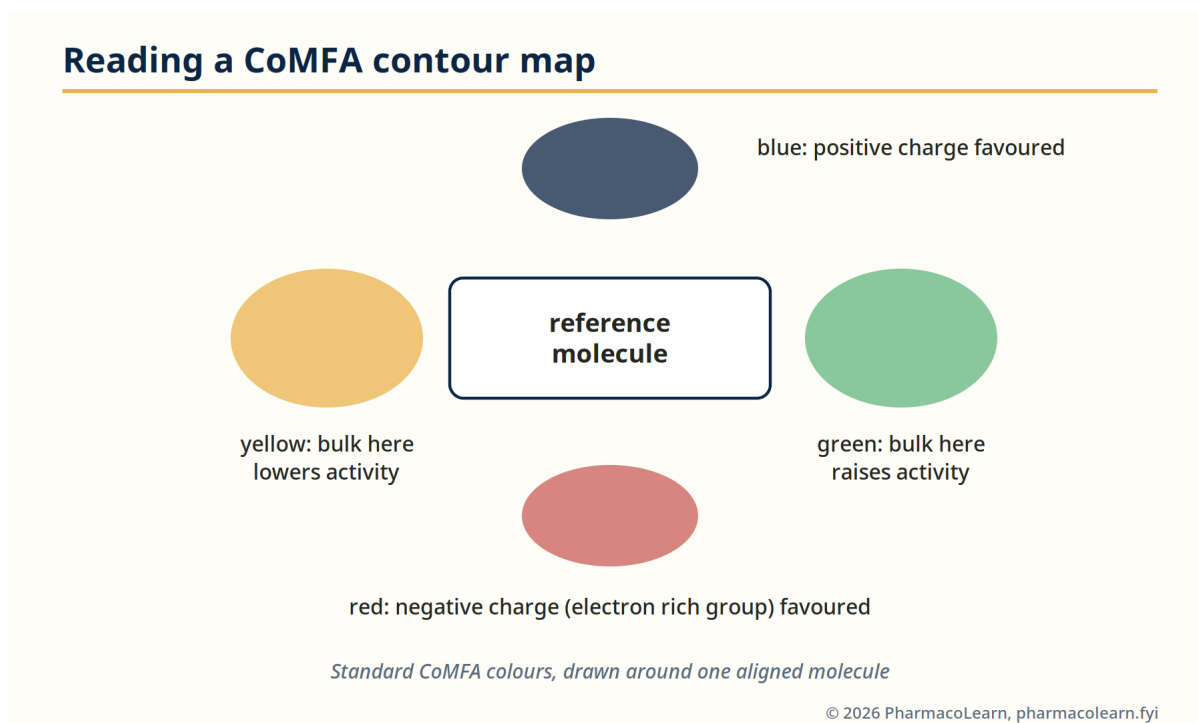


Figure: A CoMFA map in words: green wants bulk, yellow forbids it, blue wants positive charge, red wants negative charge, each in a defined region of space.

2.9 Statistical methods used in QSAR analysis

The fitting engine of classical QSAR is **multiple linear regression (MLR)**, which finds the coefficients minimising the sum of squared residuals, and of 3D QSAR, **partial least squares (PLS)**, which compresses thousands of field descriptors into a few latent components before regressing. Around the fitting sit the validation methods, which matter more:

- **Cross validation:** compounds are repeatedly held out, the model is rebuilt without them, and their activities are predicted. In **leave one out** cross validation each compound is predicted by a model that never saw it. The cross validated correlation, q^2 , measures predictive power rather than fit.
- **Test set validation:** a portion of the series, often 20 to 30 percent, is never used in fitting at all and serves as the honest exam at the end.
- **Y randomisation:** the activity values are shuffled and the model rebuilt; if a shuffled model performs as well as the real one, the real one was chance.
- **Applicability domain:** the region of descriptor space the training set covers; predictions are trusted only inside it.

2.10 Statistical parameters and their importance

Parameter	Meaning	Healthy sign
n	Number of compounds	At least about 5 per fitted term
r, r squared	Correlation, and fraction of activity variance explained by the fit	r squared above about 0.8 for a classical model, but never sufficient alone
s (standard error of estimate)	Typical size of a residual, in log units	Small compared with the activity range; near assay error is ideal
F ratio	Variance explained per term against residual variance	Large, and statistically significant at the stated level
q squared	Cross validated predictive correlation	Above 0.5 is the usual working threshold for a useful model
PRESS	Sum of squared prediction errors under cross validation	Falls as q squared rises; the raw material of q squared

The importance of the statistical parameters is precisely that they stop self deception. A high r squared can be bought cheaply by adding terms until the equation memorises the data, which is why q squared, computed on compounds the model did not fit, is the number an examiner, and a project team, should ask for first. A model that fits beautifully and predicts badly is worse than no model, because it spends synthesis money confidently.

2.11 Advantages, disadvantages and classic errors

Advantages of QSAR, fairly stated: it extracts mechanism flavoured guidance from data already paid for; it predicts before synthesis and so saves the scarcest resource, chemist time; it is cheap, fast and uses no animals; and its equations communicate, a Hansch coefficient is an argument the whole team can inspect.

Disadvantages: it works only within a series and a mechanism; it cannot invent a new scaffold; its parameters idealise a messy binding event; and 3D methods stand or fall on the alignment. The classic errors deserve naming because each has burnt real projects: **chance correlation**, fitting noise when descriptors outnumber information; **collinear descriptors**, which make coefficients swap signs between runs; **overfitting**, too many terms for the compound count; **extrapolation** beyond the fitted ranges; **activity cliffs**, where a tiny structural change collapses activity and no smooth equation can follow; and **dirty data**, mixed assays, wrong structures and misreported values, against which no statistic protects. Unit 2's discipline, parameters with meaning, statistics with suspicion, is the antidote.

*More free M.Pharm notes and unit wise PDFs at pharmacolearn.fyi
Free video lectures on the PharmacoLearn YouTube channel*