PharmacoLearn

MPL203T Principles of Drug Discovery, M.Pharm Semester 2

Unit 5: QSAR Statistics, 3D QSAR and Prodrug Design

Free notes at pharmacolearn.fyi

5.1 The statistical contract of QSAR

A QSAR equation is a statistical claim before it is a chemical one, and Unit 5 supplies the machinery for judging it. The contract has three clauses. First, **enough data**: a working rule is at least five compounds per fitted parameter, since each added term will improve the fit to the data in hand whether or not it captures anything real. Second, **honest significance**: the statistics of section 5.2 must show the fit exceeds what chance correlation among random numbers would produce. Third, **demonstrated prediction**: the equation must predict compounds it was not fitted on, tested by the validation discipline of section 5.5. An equation meeting all three is a model; one meeting only the first two is a description.

5.2 Regression analysis

Regression analysis, in QSAR almost always multiple linear regression, finds the coefficients of the Hansch type equation that minimize the squared differences between observed and calculated activities. The output is judged by a standard battery:

Statistic	Meaning	Healthy sign
r, the correlation coefficient (and r squared)	Fraction of the variation in activity the equation accounts for	r squared above about 0.8 for a working model, read with the other statistics, never alone
s, the standard deviation of the residuals	Typical size of the equation's error, in activity units	Comparable to the experimental error of the biological measurements themselves; it cannot honestly be much smaller
F, the variance ratio	Whether the explained variance dwarfs the unexplained, given the numbers of compounds and parameters	Large, and significant at a strict level; punishes equations that buy fit with parameters
Residual pattern	The signs and sizes of individual errors across the series	No structure: a curved residual pattern is the equation confessing a missing term, as when a linear log P fit leaves a parabola in its errors

Regression also carries the standard warnings of the method: it assumes the parameters are measured without major error and independent enough to be separable, and with many candidate parameters and few compounds it will find a flattering equation by chance, the **chance correlation** problem that the five to one rule and cross validation exist to restrain.

5.3 Partial least squares analysis

Partial least squares (PLS) is the multivariate method that succeeds regression where regression fails: when the descriptors are numerous and highly correlated with one another. Instead of using the raw descriptors, PLS constructs a small number of **latent variables**, weighted combinations of many descriptors, each chosen to capture the maximum remaining covariance with activity, and the model regresses activity on a handful of these components. Two consequences follow. The method tolerates descriptor counts far exceeding compound counts, precisely the situation in 3D QSAR, where each molecule contributes thousands of field values (section 5.6). And the number of components is chosen by cross validation, stopping where prediction, not fit, stops improving, so restraint is built into the procedure.

© 2026 PharmacoLearn | pharmacolearn.fyi

5.4 Other multivariate statistical methods

Around regression and PLS stands a family of methods the syllabus groups as the other multivariate techniques, each answering a different question of the data:

- **Principal component analysis (PCA)** compresses correlated descriptors into a few uncorrelated components that carry most of the information, used to survey a series, spot clusters and outliers, and see the shape of a data set before modelling it.
- **Cluster analysis** groups compounds or substituents by similarity, guiding the selection of a diverse training set that spans the chemistry instead of duplicating it.
- **Discriminant analysis and classification methods** (including k nearest neighbours) predict categories, active or inactive, rather than a numerical potency, suiting the pass or fail data of large screens.
- **Artificial neural networks** fit flexible nonlinear relationships and can capture interactions linear equations miss, at the price of transparency: a network may predict well while explaining nothing, and it over fits small data sets eagerly.

The choice follows the data: few clean parameters and one mechanism favour regression; thousands of correlated fields demand PLS; category answers call on the rest.

5.5 Validation of QSAR models

Validation is where statistical QSAR earns its predictions. **Internal cross validation** repeatedly refits the model with part of the data withheld (in leave one out validation, one compound at a time) and predicts the withheld values; the resulting cross validated r squared, written q squared, estimates predictive power, and a value above about 0.5 is the customary minimum claim. **External validation** is stricter: compounds set aside before any modelling, the test set, are predicted once by the finished model, and the errors are reported without appeal. Two further guards complete the discipline. **Y randomization** scrambles the activity values and refits: if scrambled data yield comparable statistics, the original model was chance correlation and is discarded. And the **applicability domain** states the chemical territory the model covers, defined by the training set's structures and parameter ranges; predictions outside it are extrapolations and are labelled as such. A QSAR presented without q squared, a test set and a domain is, by the standards of this unit, an anecdote with subscripts.

5.6 3D QSAR: from substituent constants to molecular fields

Classical QSAR describes a molecule by substituent numbers; it cannot see shape. **3D QSAR** describes each molecule by the **fields** around it: the steric (van der Waals) and electrostatic potentials that a probe senses at thousands of points of a grid enclosing the aligned molecules. Alignment is the method's foundation and its vulnerability: all molecules must be placed in the binding relevant conformation and orientation, by docking, by pharmacophore matching or by atom superposition, because the field comparison is meaningless unless equivalent positions correspond. With alignment done, PLS (section 5.3) relates the thousands of field values to activity. The result is read not as an equation but as **contour maps**: volumes around the average molecule where more bulk, less bulk, positive or negative charge raise or lower predicted activity. The map is a design drawing laid over the chemistry, and it points synthesis at the unexploited favourable regions.

5.7 CoMFA

CoMFA (comparative molecular field analysis), introduced by Cramer and colleagues in 1988, is the founding 3D QSAR method. Each aligned molecule is placed in a regular three dimensional lattice, and at

every lattice point a probe atom computes the steric energy (a Lennard Jones potential) and the electrostatic energy (a Coulomb potential with atomic charges) of its interaction with the molecule. The resulting table, thousands of field values per compound, is analysed by PLS, and the coefficients are drawn back onto the grid as contour maps with a conventional reading: green contours mark where added steric bulk is predicted to raise activity and yellow where it lowers it; blue marks where positive charge favours activity and red where negative charge does. CoMFA's weaknesses follow from its machinery: the steep Lennard Jones wall forces arbitrary energy cut offs, results move with grid placement and alignment choices, and only steric and electrostatic stories can be told. Its historical importance is secure: it converted QSAR from substituent arithmetic into three dimensional design guidance.

5.8 CoMSIA

CoMSIA (comparative molecular similarity indices analysis) keeps the CoMFA architecture and repairs its rough edges. Instead of interaction energies with a probe, it computes, at each grid point, **similarity indices**: how similar each atom of the molecule is to a probe in steric bulk, electrostatic character, hydrophobicity, and hydrogen bond donor and acceptor capacity, using smooth Gaussian functions of distance rather than the hard cut off potentials of CoMFA. The practical gains are smoother, less grid sensitive contour maps and two extra fields classical CoMFA lacks, so the maps can now show where a hydrophobic group or a hydrogen bond donor belongs, the very vocabulary of pharmacophore thinking in section 3.9. The methods are read together in practice: where CoMFA and CoMSIA agree on a region, the design instruction is trusted; where they differ, the difference itself localizes modelling uncertainty. The comparison is summarized:

Feature	CoMFA	CoMSIA
Quantity computed on the grid	Steric and electrostatic interaction energies with a probe atom	Similarity indices for steric, electrostatic, hydrophobic, hydrogen bond donor and acceptor properties
Distance dependence	Lennard Jones and Coulomb potentials, with energy cut offs	Gaussian functions, smooth, no cut offs needed
Sensitivity to grid position	Noticeable	Lower
Output	PLS model and steric/electrostatic contour maps	PLS model and contour maps in five property fields
Alignment requirement	Same: critical for both	Same: critical for both

5.9 Prodrug design: the basic concept

A **prodrug** is a pharmacologically inactive, or far less active, derivative of a drug that is converted in the body, by enzymes or by chemistry, back to the active parent. The design inverts the usual problem. Sections 1.1 to 5.8 find and perfect the active molecule; prodrug design accepts the active molecule as it is and engineers its delivery through the body by temporary chemical disguise. The disguise (the **promoiety**) is chosen to repair a named defect of the parent, poor solubility, poor absorption, instability, bad taste, short action, or failure to reach a tissue, and it must come off at the right place and rate. Two construction families are distinguished: **carrier linked prodrugs**, where an inert carrier group is attached and later cleaved (usually an ester), the common case; and **bioprecursor prodrugs**, where the molecule itself is metabolically transformed into the active species, as levodopa is decarboxylated into dopamine. The syllabus then names the goals one by one, and they are taken in its order.

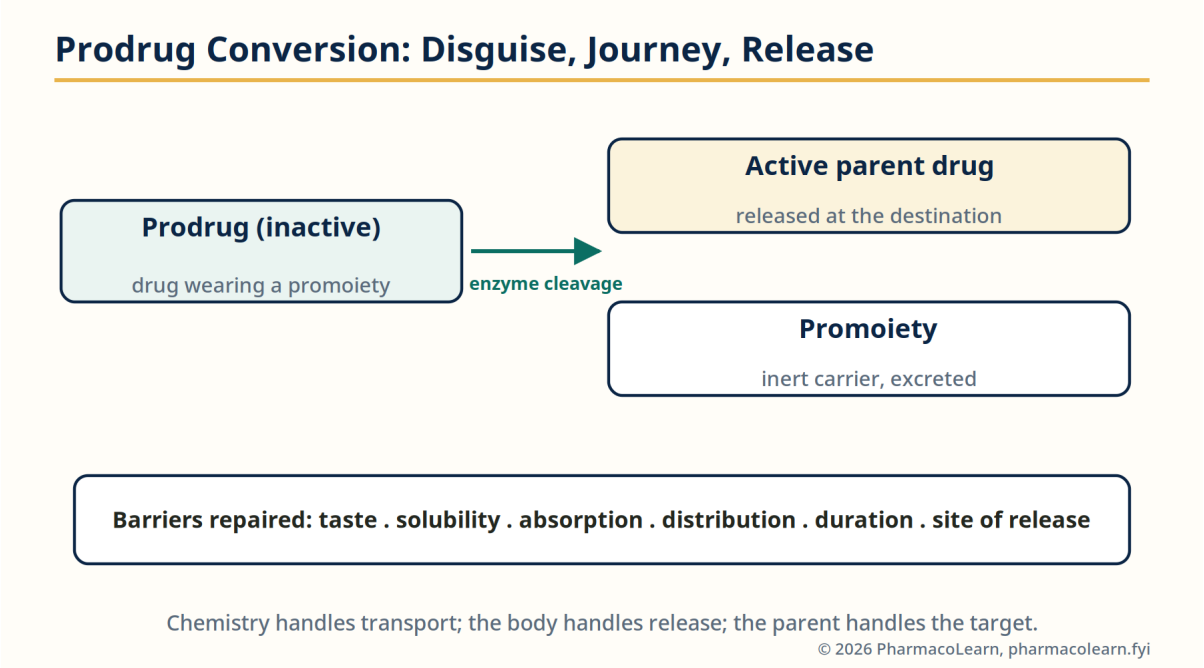

Figure: Prodrug conversion: the disguised drug makes the journey, the promoiety is cleaved, and the active parent drug is released at the destination.

5.10 Prodrugs to improve patient acceptability

A drug a patient refuses is a drug that fails. Acceptability prodrugs mask properties that make dosing unpleasant. The classical case is **taste**: chloramphenicol is intensely bitter, and its palmitate ester is essentially tasteless because it does not dissolve, and therefore is not tasted, in the mouth; intestinal enzymes hydrolyse it to the parent. The same logic reduces the pain of injections (an irritating parent presented as a bland ester in an oily depot), the odour or gastric irritation of others, and generally trades a formulation problem the patient experiences for a cleavage step the body performs silently.

5.11 Prodrugs to improve drug solubility

Insolubility defeats dosage forms, above all injections, which demand a true solution. The solubility prodrug attaches an ionizable, water loving promoiety, classically a **phosphate ester**, to the parent: the prodrug dissolves freely, is administered, and phosphatases in blood and tissues release the parent. **Fosphenytoin**, the phosphate prodrug of phenytoin, is the standard example: freely water soluble where phenytoin is not, it can be given intravenously or intramuscularly without the aggressive solvents and the tissue damage of the older formulation. Chloramphenicol sodium succinate plays the same role for parenteral chloramphenicol. The design constraint is conversion kinetics: cleavage must be fast and complete enough that dosing the prodrug is equivalent to dosing the parent, and the equivalence must be demonstrated, not assumed.

5.12 Prodrugs to improve absorption and distribution

More prodrugs exist for this goal than any other, because absorption is where lipophilicity earns its keep. The ester strategy raises membrane permeability by masking a polar, ionized group, then surrenders the ester to plasma and tissue esterases:

- **Enalapril** is the ethyl ester of enalaprilat; the parent diacid is too polar to absorb well, the ester crosses the gut, and hydrolysis in the liver releases the ACE inhibitor.
- **Oseltamivir** is the ethyl ester of its active carboxylate; the ester is absorbed, the acid is not, and the antiviral is delivered by the same trick.

- **Valaciclovir** couples valine to aciclovir so that an intestinal peptide transporter carries the prodrug across, multiplying the aciclovir exposure the parent achieves; here the promoiety recruits a transporter rather than merely hiding polarity.
- **Levodopa** solves a distribution problem: dopamine cannot cross the blood brain barrier, but levodopa, an amino acid, is carried across by the neutral amino acid transporter and decarboxylated to dopamine in the brain, the bioprecursor pattern of section 5.9 at its most famous.

Distribution prodrugs in general aim the parent at a tissue: increasing brain entry, prolonging residence, or, in the next section, releasing only at a chosen site.

5.13 Site specific delivery and sustained action through prodrugs

Site specific delivery arranges for conversion to happen at the target and nowhere else. **Sulfasalazine** is the textbook design: 5 aminosalicylic acid, the anti inflammatory active in ulcerative colitis, is joined by an azo bond to sulfapyridine; the intact molecule passes the upper gut largely unabsorbed, and colonic bacteria reduce the azo bond exactly where the disease is, releasing the active agent in the colon. Enzyme targeted strategies generalize the idea, exploiting enzymes enriched at the site (including, experimentally, antibody directed enzyme prodrug therapy in cancer, where an antibody delivers an activating enzyme to the tumour before the prodrug is given). **Sustained action** uses the opposite kinetic arrangement: the prodrug is built to convert slowly, so it acts as a chemical depot. Long chain fatty acid esters of antipsychotics such as fluphenazine decanoate, injected in oil, hydrolyse over weeks, converting a daily oral adherence problem into a monthly injection, the pharmacokinetic repair of a social problem.

```
PRODRUG DESIGN, DECISION FLOW

  Parent drug has a named barrier: taste? solubility?
  absorption? distribution? duration? site?
      |
      v
  Choose a promoiety that repairs that barrier
  and a bond the body can cleave (usually an ester)
      |
      v
  Check conversion: where, by what enzyme, how fast,
  and is the promoiety itself harmless?
      |
      v
  Check stability: shelf life in the bottle,
  survival to the conversion site
      |
      v
  The prodrug is a new drug: full testing follows
```

5.14 The rationale of prodrug design

The **rationale** assembles the preceding sections into one argument. Begin with a parent whose pharmacology is right and whose pharmaceutical or pharmacokinetic behaviour is wrong in a specific, nameable way; the prodrug is justified only when the defect is in delivery, not in the target interaction, since no promoiety repairs a weak pharmacophore. Choose the promoiety to remove exactly that defect and no more, because every addition changes solubility, stability and conversion together. Arrange conversion to occur at the intended site, at a rate matched to the dosing plan, by enzymes reliably present in patients. Then the prodrug is not a trick but a division of labour: chemistry handles transport, biology handles release, and the parent handles the target.

5.15 Practical considerations of prodrug design

The practical register closes the unit, and it is where elegant designs are audited:

- **Conversion reliability**: the activating enzymes vary between species, which has misled many animal based programmes, and between patients, through genetics, age, disease and interacting drugs; a prodrug activated by a variable enzyme inherits that variability as variable dosing.
- **Promoiety safety**: the released carrier is a chemical dose in its own right and must be innocuous at the exposure involved; history's cautionary examples come from promoieties, such as formaldehyde releasing groups, that were not as innocent as assumed.
- **Stability balance**: the bond must survive manufacture, shelf storage and the journey to the conversion site, yet cleave promptly on arrival; too stable and the drug passes through unconverted, too labile and it is a parent drug with extra steps.
- **Incomplete conversion and competing pathways**: if cleavage is partial, or metabolism diverts the prodrug elsewhere, exposure to the parent becomes unpredictable, and the programme must measure parent, not prodrug, as the active currency.
- **Regulatory reality**: a prodrug is a new chemical entity, requiring its own full characterization and safety evaluation; the shortcut it offers is pharmacological, not administrative.

Read with Unit 4, this closes the subject's arc: targets are found and validated, leads are found and measured, relationships are made quantitative and three dimensional, and where the winning molecule still cannot reach its target, chemistry disguises it for the journey.

Frequently Asked Questions

Q: What is the difference between target identification and target validation? Identification names a molecule that a drug could act on, based on its place in disease biology. Validation proves the harder claim: that modulating this molecule, in the right direction, will actually change the disease in a patient at tolerable cost. Evidence climbs a ladder from association, through human genetics and pharmacological and transgenic experiments, to clinical proof. Most costly discovery failures are validation failures: excellent chemistry against targets that turned out not to matter.

Q: A knockout mouse lacking the target shows no protection from disease. Is the target dead? Not automatically. A germline knockout removes the protein for the animal's whole development, and other pathways can compensate, masking a role that acute blockade would reveal; the protein may also matter only in one tissue or disease stage. The result weakens the target and demands a second test, an inducible or conditional knockout, knockdown by siRNA or antisense, or a tool compound, before the programme is killed. Validation is a weight of evidence, not a single experiment.

Q: Combinatorial chemistry and high throughput screening are often named together. How do they differ? They are the supply and the test of lead identification. Combinatorial chemistry manufactures large libraries of related compounds by combining building blocks systematically, by parallel synthesis or split and mix methods. High throughput screening tests large collections, from any source, against the assay, using robotics and miniaturized plates. One makes candidates at scale; the other interrogates candidates at scale; a library from combinatorial chemistry is one of the things HTS screens.

Q: When is homology modelling used, and when threading? Both predict protein structure from sequence. Homology modelling applies when a related protein of known structure exists: the target sequence is aligned to that template and the model is copied and adapted from it, with accuracy rising with sequence identity. Threading is the fallback when no sequence relative can be found: the sequence is fitted against a library of known folds and scored for compatibility, identifying the likely fold family but not atomic

detail. Homology gives coordinates for design; threading gives architecture for triage.

Q: NMR and X-ray crystallography both determine protein structure. What decides between them? Crystallography needs the protein to crystallize, and then delivers high resolution structures, including drug complexes, at relatively high throughput; most structures used in design come from it. NMR works in solution, needs no crystals, reports conformational dynamics, and can detect weak ligand binding directly, but is limited to smaller proteins and is slow and expert intensive. They are complements: crystallography for the design cycle, NMR for solution behaviour, small domains and binding studies.

Q: What, in one sentence each, is the core difference between traditional and rational drug design? Traditional design finds a molecule by its effect in a whole system first and learns its mechanism later, if ever: effect before explanation. Rational design starts from knowledge of the disease mechanism and the target, designs the molecule to fit that knowledge, and tests the stated hypothesis: explanation before effect. Modern programmes are hybrids, rational in choosing the target and assay, empirical in running screens, rational again in optimizing what the screens return.

Q: Is Lipinski's rule of five a law that a drug must obey? No. It is a statistical screen distilled from marketed oral drugs: absorption and permeation problems become more likely beyond molecular weight 500, log P 5, 5 hydrogen bond donors and 10 acceptors. In drug likeness screening it flags risk, one violation warning and two arguing against a series, but genuine drugs violate it, notably natural products, non oral drugs and transporter substrates. Deleting every violator would delete real medicines; ignoring it entirely fills pipelines with large, greasy failures.

Q: Rigid, flexible and manual docking: what changes between them? The amount of motion allowed in the search. Rigid docking moves a fixed ligand conformation through the site as one body: fastest, and it misses poses that need the ligand to bend. Flexible docking explores the ligand's rotatable bonds during placement, usually with the protein held rigid or nearly so, and is the standard working method. Manual docking is placement by a scientist at a graphics screen, slowest of all, used to build and inspect hypotheses for the few compounds that matter most. Speed falls and realism rises along the series.

Q: Hansch and Free Wilson analysis both predict activity. When is each preferred? Hansch analysis relates activity to physicochemical parameters (log P, sigma, steric terms) of the substituents, so it can estimate a substituent never yet made and it reveals an optimum, such as the best log P. It needs parameter values and a single mechanism across the series. Free Wilson analysis uses only the series' own data, adding each substituent's measured contribution to the parent activity, so it needs no property assumptions but can only recombine substituents already synthesized, and fails where substituents interact. Fujita's mixed approach combines the two where each is strongest.

Q: Why would a designer deliberately make an inactive molecule, a prodrug? Because the active molecule often cannot complete the journey: too insoluble to inject, too polar to absorb, too bitter to swallow, excluded from the brain, or gone in hours. The prodrug is the active drug in a temporary disguise chosen to repair that one delivery defect: enalapril's ester gets the ACE inhibitor absorbed, levodopa's amino acid form gets dopamine's precursor into the brain, sulfasalazine's azo bond holds release until the colon. Body enzymes remove the disguise at the destination, and the freed parent does the pharmacology.

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