

Unit 1: Introduction to Computer Aided Drug Design and QSAR Basics

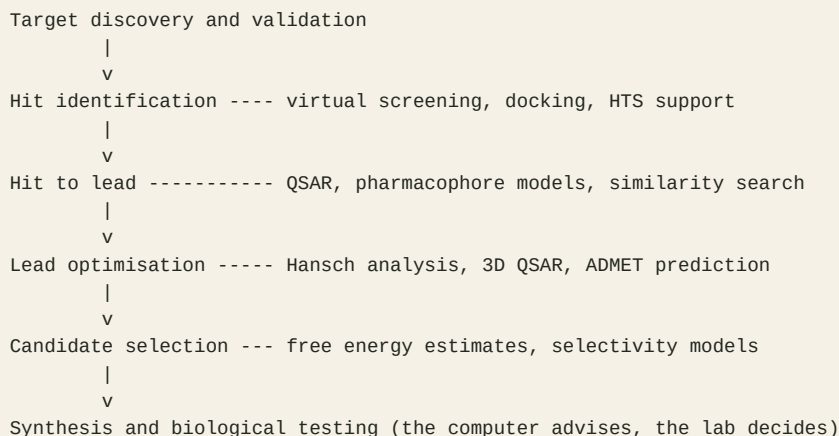
Free notes at pharmacolearn.fyi

Computer Aided Drug Design, subject code MPC203T, is a core Semester 2 subject of M.Pharm Pharmaceutical Chemistry. It carries 4 credits and 60 hours of theory, taught at 4 hours per week, for 100 marks. This subject teaches how computers are used to design drug molecules before, and sometimes instead of, making them in the laboratory. Unit 1 introduces computer aided drug design (CADD) and the foundations of quantitative structure activity relationships (QSAR): the physicochemical parameters, electronic, lipophilic and steric, that connect a structure to its activity. Unit 2 applies those parameters in Hansch analysis and Free Wilson analysis, shows how 2D QSAR equations are derived, moves to 3D QSAR and contour maps, and covers the statistical methods that decide whether a model is real or a mirage. Unit 3 is molecular modeling and docking: molecular and quantum mechanics, energy minimization, the difference between the global minimum and the bioactive conformation, and docking into the enzyme targets the syllabus names, DHFR, HMG CoA reductase, HIV protease and the cholinesterases. Unit 4 turns to molecular properties: prediction of ADMET behaviour, de novo and fragment based design, and homology modeling of proteins whose structures are unknown. Unit 5 finishes with pharmacophore mapping and virtual screening, the techniques that search whole databases for the next lead. A student who completes this subject should be able to read a QSAR equation and judge it, explain what a docking program actually does, and plan a sensible in silico screening protocol.

1.1 Why drug design needed computers

A modern drug discovery programme tests chemistry against biology at industrial scale, and it is slow and expensive: a new medicine typically takes well over a decade to reach patients, and most candidate molecules fail along the way, usually because they are not absorbed, are metabolised too fast, or turn out to be toxic. Synthesis is the bottleneck of classical medicinal chemistry. Every analogue in a series costs days of bench work, and a chemist can only explore the ideas that can physically be made and assayed. Computer aided drug design attacks that bottleneck by moving part of the exploration into the computer. A model can estimate the activity of a compound that does not exist yet, a docking program can test whether a proposed molecule fits its target before a flask is charged, and a screening protocol can search millions of structures overnight and hand the chemist a shortlist worth making. CADD does not replace experiment; it decides which experiments are worth doing. Two broad strategies run through the whole subject. **Ligand based design** works from known active molecules and infers the features activity requires, the home ground of QSAR and pharmacophore methods. **Structure based design** works from the three dimensional structure of the target protein and designs molecules to fit it, the home ground of docking, de novo design and homology modeling. Most real projects use both, and this course teaches them in that order.

1.2 The CADD toolbox: techniques and applications



The main techniques, each developed in later units, are:

- **QSAR:** mathematical equations linking activity to measurable or calculable properties of a series, used to predict the next analogues to make (Units 1 and 2).
- **Molecular modeling:** building and manipulating molecular structures in three dimensions, minimising their energy and exploring their shapes (Unit 3).
- **Molecular docking:** placing a ligand into the binding site of a target and scoring the fit, used for screening and for understanding binding (Unit 3).
- **ADMET prediction:** in silico estimates of absorption, distribution, metabolism, excretion and toxicity, used to kill weak candidates early (Unit 4).
- **De novo and fragment based design:** constructing new molecules inside a binding site, or growing them from small binding fragments (Unit 4).
- **Homology modeling:** building a model of a protein whose experimental structure is unknown, from the structure of a relative (Unit 4).
- **Pharmacophore mapping and virtual screening:** abstracting the essential features of active molecules and using them, or docking, to search large databases (Unit 5).

Applications follow from the list: prioritising compounds for synthesis, explaining why a series behaves as it does, finding new chemotypes for a known target, flagging ADMET liabilities before animal work, and rescuing projects where no crystal structure exists. The honest limitation belongs in the same breath: a model is only as good as the data and assumptions inside it, and every in silico prediction of consequence is eventually checked at the bench.

The CADD toolbox: two strategies, one loop

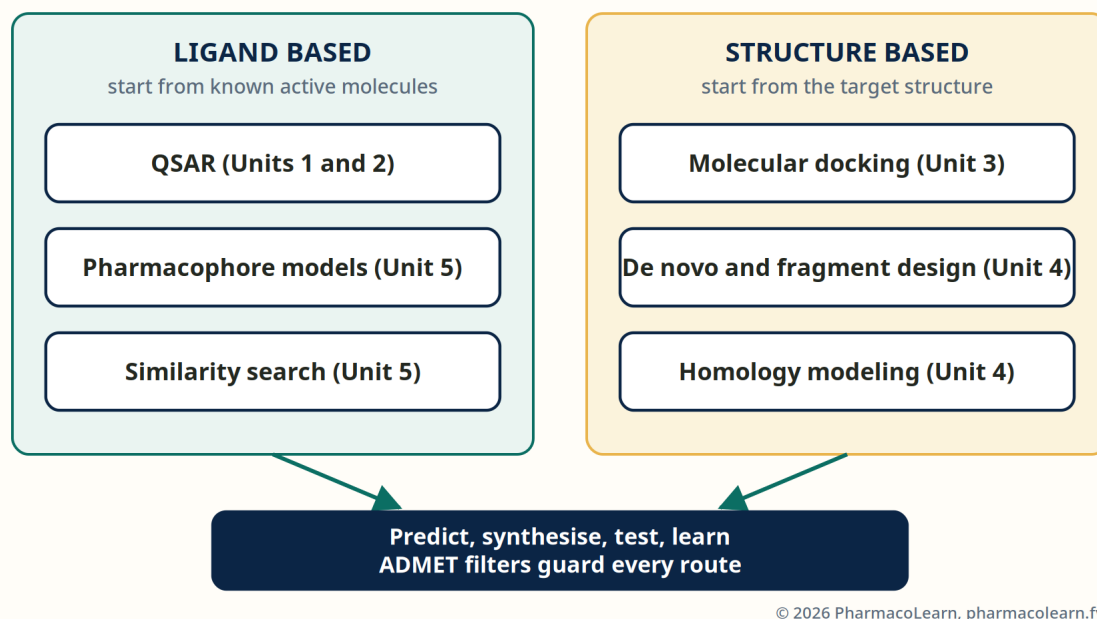


Figure: Ligand based methods learn from known actives, structure based methods from the target; both feed the same predict, synthesise and test loop, with ADMET filters on guard.

1.3 A short history of the subject

The idea that structure determines activity is older than the computer by a century. In 1868 Crum Brown and Fraser proposed that the physiological action of a compound is a function of its chemical constitution, the founding statement of structure activity thinking. Around 1900 Meyer and Overton independently linked narcotic potency to oil water partitioning, the first lipophilicity relationship. Hammett, in the 1930s, gave physical organic chemistry its sigma rho equation for electronic effects in benzene derivatives, the direct ancestor of QSAR parameters. The quantitative era opened in the early 1960s: Hansch and co workers combined lipophilic, electronic and steric parameters in multiple regression equations (the Hansch approach, 1962 to 1964), and Free and Wilson published their additive group contribution model in 1964. Through the 1970s and 1980s, molecular graphics and force fields made molecules visible and manipulable on screen, the first docking programs appeared, and in 1988 Cramer introduced CoMFA, the first practical 3D QSAR method. The 1990s added combinatorial libraries, high throughput screening and Lipinski's rule of five (1997), which defined the property space of oral drugs. The modern era runs on thousands of experimental protein structures, cheap computing, and machine learning methods layered on the classical ones. The exam relevant point is the logic of the history: each generation of technique answers the weakness of the one before, and the classical methods in this syllabus are still the interpretive core of the newer ones.

1.4 SAR versus QSAR: the basic idea

A **structure activity relationship (SAR)** is a qualitative account: in this series, a chloro substituent raises activity, a hydroxyl lowers it, moving the side chain kills it. SAR is read from tables of analogues and guides intuition. A **quantitative structure activity relationship (QSAR)** goes one step further and makes the relationship mathematical: the biological activity of a series of compounds is expressed as an equation in terms of their physicochemical properties or structural descriptors, for example $\log(1/C)$ equals some combination of $\log P$, sigma and a steric term plus a constant. Three consequences follow. First, the equation can **predict** the activity of compounds not yet synthesised, which is the economic point of the

exercise. Second, the coefficients are interpretable: a positive coefficient on log P says activity rises with lipophilicity in this series, a negative one on sigma says electron donating groups help, and so on. Third, the model can be **tested statistically**, and Unit 2 insists on that testing. Activity in QSAR is almost always handled in logarithmic form, log (1/C), where C is the molar concentration producing a standard effect (an IC50, EC50 or minimum effective dose), because free energy of binding is proportional to the logarithm of an equilibrium or potency term: equal steps in log activity correspond to equal steps in binding free energy, which is what substituent effects actually add up to.

1.5 Physicochemical parameters: the three families

QSAR parameters fall into three families, and the syllabus names a champion for each.

Family	What it captures	Syllabus parameters
Electronic	The electron pushing or pulling effect of a substituent on reaction centres and binding interactions	Hammett sigma constants, derived from ionisation of benzoic acids
Lipophilic	The tendency of a molecule or substituent to leave water for a fatty phase, governing transport and hydrophobic binding	log P of the whole molecule, pi substituent constant
Steric	The size and bulk of a substituent, governing fit into a pocket and access to a centre	Taft Es constant, molar refractivity (MR)

A fourth, informal group completes the working set: **structural and indicator parameters**, such as counts of hydrogen bond donors and acceptors, molecular weight, and indicator variables that simply record the presence (1) or absence (0) of a feature. The art of QSAR is choosing a small set of parameters that are not strongly intercorrelated, because two parameters carrying the same information destabilise the regression and confuse the interpretation.

1.6 Electronic parameters: the Hammett equation and sigma

Hammett studied the ionisation of substituted benzoic acids in water and defined the substituent constant **sigma** by comparing each substituted acid with benzoic acid itself: sigma equals log (K_X divided by K_H), where K_X is the ionisation constant of the substituted acid and K_H that of the parent. A substituent that withdraws electrons stabilises the carboxylate anion, increases ionisation and earns a **positive** sigma; an electron donating substituent earns a **negative** one. Illustrative textbook values: para nitro, sigma about plus 0.78 (strongly withdrawing); para chloro, about plus 0.23; hydrogen, zero by definition; para methyl, about minus 0.17; para methoxy, about minus 0.27 (donating by resonance). Meta and para positions have separate constants (sigma_m and sigma_p) because resonance donation reaches the para position but not the meta. The **Hammett equation** generalises the idea to any reaction or equilibrium influenced by the same substituents: log (k_X divided by k_H) equals rho multiplied by sigma. The reaction constant **rho** measures how sensitive that particular reaction is to electronic effects; rho near 1 by construction for the benzoic acid ionisation itself, positive rho means the process is helped by electron withdrawal, negative rho that it is helped by donation. A worked micro example: for a reaction series with rho equal to 1.00, a para nitro group changes log (k/k0) by 1.00 times 0.78, that is 0.78 log units, roughly a sixfold rate increase. In drug series the same logic applies to binding: if activity correlates with sigma with a positive coefficient, electron withdrawing substituents favour activity, a mechanistic clue about the interaction in the binding site.

1.7 Lipophilicity: log P, log D and the pi constant

Lipophilicity is the single most used parameter in drug design, because a drug must dissolve in water well enough to be carried, yet cross fatty membranes well enough to reach its target, and then often bind a hydrophobic pocket. It is measured by the **partition coefficient** P, the ratio of the concentration of the unionised compound in n octanol to its concentration in water at equilibrium, and quoted as **log P**. n

Octanol is the standard organic phase because its hydroxyl head and hydrocarbon tail make it a crude but serviceable model of a membrane. Two refinements matter. **log D**, the distribution coefficient, counts all species, ionised and unionised, at a stated pH (usually 7.4), and is the relevant number for ionisable drugs in the body. The **pi substituent constant** (Hansch and Fujita) isolates one substituent's contribution: π_X equals $\log P$ of the substituted parent minus $\log P$ of the unsubstituted parent. Worked example: $\log P$ of benzene is 2.13 and of toluene 2.69, so π for a methyl group is 2.69 minus 2.13, which equals 0.56; methyl adds about half a log unit of lipophilicity wherever it is attached. π values let a chemist price a substitution before making it, and they add: a compound's $\log P$ can be estimated by summing the parent value and the π values of its substituents, with corrections where substituents interact.

1.8 Determining log P: experimental and theoretical approaches

The syllabus explicitly asks for experimental and theoretical approaches to the parameters, and $\log P$ is the model case.

- **Shake flask method (experimental, the reference method):** the compound is shaken between pre saturated octanol and water, the phases are separated, and the concentration in each is measured, usually by UV spectroscopy or HPLC. It is accurate but slow, needs pure material, and struggles with very lipophilic or very hydrophilic compounds.
- **Chromatographic methods (experimental):** reversed phase HPLC retention correlates with $\log P$ after calibration with standards of known value; thin layer variants exist. Faster and usable for mixtures, but indirect.
- **Computational methods (theoretical):** fragment based programs, of which CLOGP is the classic, add up contributions of the molecule's fragments with correction factors; atom based methods assign a contribution to every atom type; property based methods use calculated surface and charge descriptors. They are instant and need no material, but their accuracy depends on how well the fragment library covers the chemistry in hand, and unusual structures can be badly wrong.
- **For sigma and the steric constants** the same pattern holds: sigma comes from measured ionisation constants (experimental) or is looked up from extensive tables for new work, and calculated electronic descriptors from quantum chemistry (charges, orbital energies) increasingly serve as theoretical substitutes.

The working rule: calculated values screen thousands of ideas, measured values anchor the model where it matters.

1.9 Steric parameters: Taft Es and molar refractivity

Bulk matters when a substituent must fit a pocket or shield a centre. **Taft's steric parameter Es** comes from the acid catalysed hydrolysis of substituted acetates, a reaction almost insensitive to electronic effects but very sensitive to crowding near the reaction centre. E_s is defined as $\log (k_X \text{ divided by } k_H)$ for that hydrolysis, with methyl set to zero; bulky groups slow the reaction and earn negative values, tert butyl being around minus 1.54 on the classic scale, a quantitative statement that it is much bulkier than methyl. **Molar refractivity (MR)** is a whole substituent size measure derived from the refractive index and density, MR equals a function of refractive index and molar volume; it scales with the substituent's volume and polarisability, so it carries both a steric meaning (bulk) and a hint of dispersion interaction strength. MR is positive and grows with size, roughly 1.03 for hydrogen and about 5.65 for a methyl group on the usual scaled units, and Hansch equations often use MR where E_s is unavailable. The two constants are not interchangeable: E_s measures bulk near an attachment point, MR measures total bulk, and choosing between them is part of model building.

1.10 Indicator variables and other working parameters

Some structural facts are yes or no, and QSAR handles them with **indicator variables**: 1 equals 1 if the molecule carries an ortho substituent, a free phenolic hydroxyl, a basic nitrogen able to be protonated, a particular ring system, and 0 otherwise. Their coefficients measure the activity bonus or penalty of that feature. Alongside them sit global descriptors used throughout design: molecular weight, polar surface area, counts of hydrogen bond donors and acceptors, and the molecule's ionisation constant pKa, which decides how much of the drug is unionised and membrane permeant at intestinal and plasma pH. These descriptors return in Unit 4 as the raw material of ADMET prediction and drug likeness rules.

1.11 Unit 1 summary table

Parameter	Symbol	What a positive value means	Typical source
Hammett constant	sigma	Substituent withdraws electrons	Ionisation of substituted benzoic acids
Reaction constant	rho	Process is favoured by electron withdrawal	Slope of Hammett plot
Partition coefficient	log P	Compound prefers octanol, lipophilic	Shake flask, HPLC, CLOGP calculation
Substituent lipophilicity	pi	Substituent adds lipophilicity	Difference of log P values
Taft constant	Es	Substituent is small near the centre (bulk gives negative)	Acid hydrolysis rates of acetates
Molar refractivity	MR	Substituent is large and polarisable	Refractive index and density, or tables

With the parameters in hand, Unit 2 does the arithmetic: building equations from them, testing the equations, and reading what they say.

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