

5.1 The pharmacophore concept

Strip a family of active molecules down to the arrangement that activity cannot live without and you have a **pharmacophore**: the spatial arrangement of features, hydrogen bond donors and acceptors, charged groups, aromatic rings, hydrophobic regions, that a molecule must present, in the right geometry, to bind its target. The word descends from Ehrlich's side chain ideas and was sharpened by generations of medicinal chemists comparing active analogues; the modern statement, in the IUPAC sense, is that a pharmacophore is the ensemble of steric and electronic features necessary for optimal interaction with a specific target and for triggering or blocking its response. Two properties of the concept matter for everything after. A pharmacophore is an **abstraction**: it names features and distances, not atoms, so chemically different scaffolds can match the same pharmacophore, which is why the method finds new chemotypes. And it is a **hypothesis**: built from the molecules that taught it, falsifiable by the next assay.

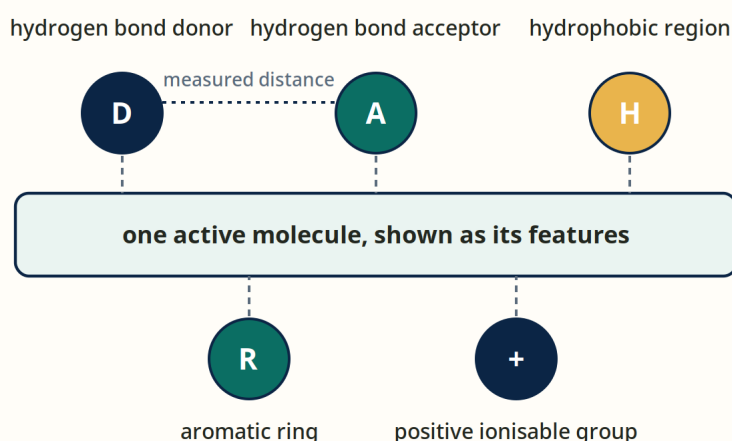
5.2 Pharmacophore features and their identification

The working feature set is short and standard:

- **Hydrogen bond donor**: a hydroxyl, amine or amide NH, drawn at the donor heavy atom, often with a projected point where the partner should sit.
- **Hydrogen bond acceptor**: a carbonyl oxygen, ether oxygen or ring nitrogen, again with a direction, because hydrogen bonds have geometry.
- **Positive and negative ionisable features**: a basic nitrogen protonated at physiological pH, a carboxylate, placed where a salt bridge is wanted.
- **Hydrophobic feature**: an aromatic ring or an alkyl cluster that must sit in a nonpolar pocket.
- **Aromatic ring feature**: where flat, pi stacking contact is the requirement, distinct from mere bulk.
- **Excluded volumes**: regions where atoms are forbidden because the protein is there, added when a model keeps proposing clashes.

Identification proceeds by aligning the active molecules and asking what they share: the features that recur, at consistent mutual distances and angles, across actives while absent or misplaced in inactives, form the model. Tolerances are part of the model; a feature sphere of generous radius admits near matches, a tight one admits only close family.

A pharmacophore: features in a fixed geometry



Features, plus the distances and angles between them, are the model

© 2026 PharmacoLearn, pharmacolearn.fyi

Figure: Five common feature types held at measured distances: the abstraction that lets new scaffolds match an old binding idea.

5.3 Building the model: ligand based and structure based

A **ligand based pharmacophore** is derived from a set of active molecules alone, as in section 5.2: the method of choice when the target structure is unknown, and dependent on the alignment of the actives, which in turn depends on their conformations, the subject of section 5.4. A **structure based pharmacophore** is derived from a protein ligand complex or an empty binding site: the pocket's donors, acceptors and hydrophobic patches (the cavity analysis of section 4.5) are converted directly into features, with excluded volumes read off the protein surface. The two routes are complementary and often combined: a structure based model disciplined by which of its features the known actives actually use. Either way, the model is then scored against databases of known actives and inactives; a model that retrieves the actives while rejecting the inactives has earned screening work.

5.4 Conformational search in pharmacophore mapping

A pharmacophore is a statement about three dimensional geometry, so mapping one from flexible molecules raises the question of section 3.4 in a sharper form: which conformation of each active molecule is the binding one? **Conformational search used in pharmacophore mapping** therefore generates, for every training molecule, an ensemble of low energy conformations by systematic rotation, Monte Carlo sampling or dynamics, and the mapping process tries the ensemble, not a single shape, looking for a feature arrangement that some conformation of every active can present. The discipline is energetic: conformations far above the global minimum (section 3.5) are implausible binding candidates and are cut, which both speeds the search and keeps the model honest. A pharmacophore built from strained, impossible conformations will screen for molecules that cannot adopt it, an elegant route to an empty hit list.

5.5 Virtual screening: the idea and the cascade

Virtual screening is the in silico counterpart of high throughput screening: instead of robotically assaying a physical library, a computational protocol scores a database of structures and ranks them, so that the fraction physically tested is the fraction most likely to be active. Databases run to millions of commercially available or enumerated compounds; the output is not drugs but **hits**, weak, real binders to be confirmed

and optimised. Because no single method is both fast enough for millions and accurate enough to trust, practical screening is a **cascade**: cheap filters first, expensive physics last, each stage shrinking the list by orders of magnitude. The enrichment that justifies the whole exercise is measured against random selection: a good protocol concentrates the true actives into the top fraction of its ranking far beyond chance.

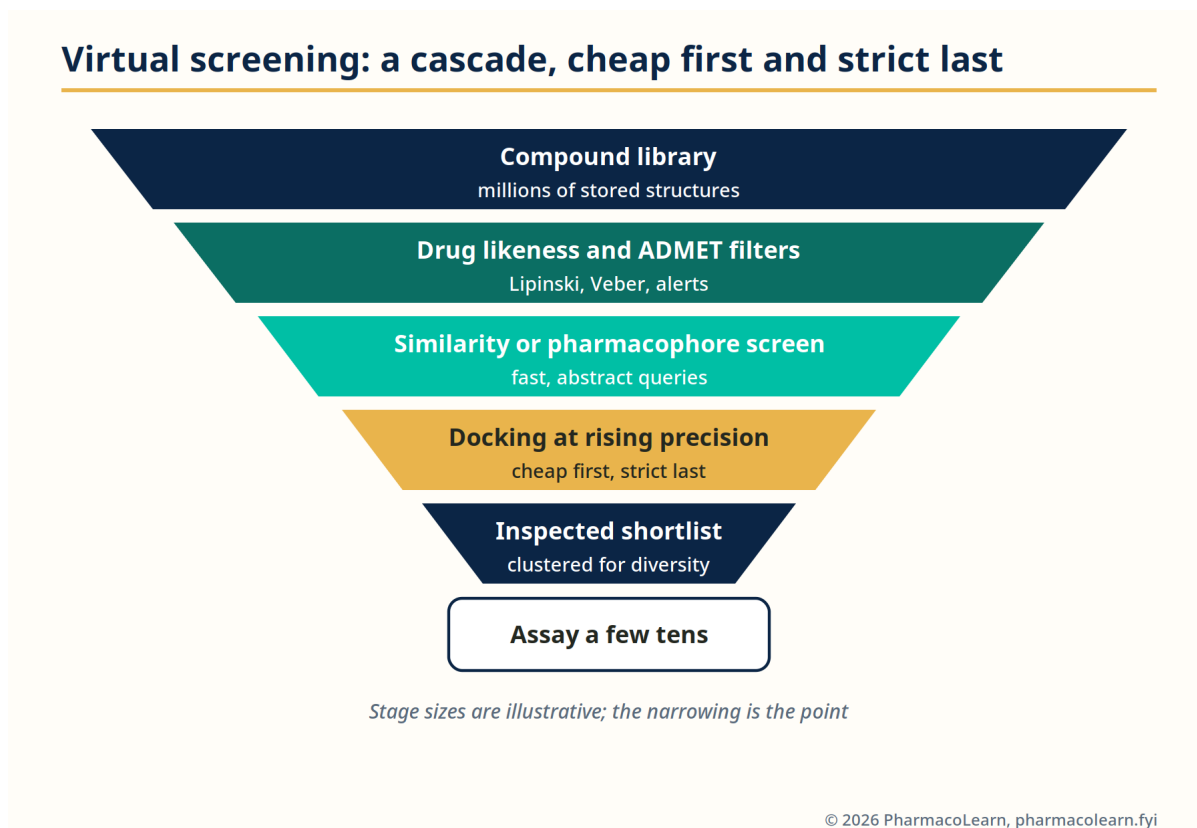


Figure: The screening cascade spends cheap computation on millions so that expensive computation, and finally bench assays, are spent on tens.

5.6 Similarity based methods: fingerprints and the Tanimoto coefficient

The oldest screening idea is the **similarity principle**: molecules that look alike tend to act alike. Similarity searching takes a known active as the query and ranks a database by resemblance. Resemblance is computed on **fingerprints**, bit strings in which each bit records the presence of a structural feature, a bond path, an atom environment or a substructure. The standard comparator is the **Tanimoto coefficient**: the number of bits set in both fingerprints divided by the number set in either, ranging from 0, nothing shared, to 1, identical fingerprints. A worked example: molecule A sets 12 bits, molecule B sets 10, and 7 bits are set in both; the Tanimoto coefficient is 7 divided by (12 plus 10 minus 7), that is 7 divided by 15, about 0.47. Working thresholds depend on the fingerprint, but similarities around 0.85 and above on common fingerprints usually mean close analogues, while the interesting scaffold hops hide lower, among the 0.6 to 0.8 neighbours that share the pharmacophore but not the skeleton. The method's weakness is written into its principle: it finds more of what you already know, and it is blind when activity turns on a feature the fingerprint does not encode.

5.7 Pharmacophore based screening

Pharmacophore based screening searches a database of precomputed conformations for molecules that can present the model's features within their tolerances and outside its excluded volumes. Because the query is abstract, the hits are structurally diverse by construction: a benzene query never appears, only the

geometry of features, so a new scaffold that happens to hold a donor, an acceptor and a ring at the model's distances surfaces alongside the usual analogues. This is the standard route to **scaffold hopping**, changing the core of a series while keeping its binding logic, valuable when the old core is patented, toxic or intractable. The screen's quality is the model's quality: models built on a single chemotype, with features set too loosely, retrieve thousands of false positives, and models set too tightly retrieve only the training family, so model validation on actives and decoys, as in section 5.3, precedes any serious screen.

5.8 Structure based virtual screening protocols

A **structure based in silico virtual screening protocol**, the syllabus's closing topic, runs docking (Unit 3) as the screen. A defensible protocol reads as a checklist:

```
Database of compounds, prepared: 3D, tautomers, charges, duplicates out
|
v
Drug likeness and property filters (Lipinski, Veber, ADMET alerts)
|
v
Fast docking of the whole library, lenient scoring
|
v
Top fraction re-docked at higher precision (extra precision mode)
|
v
Consensus scoring and interaction checks (key hydrogen bonds made?)
|
v
Visual inspection of the shortlist, clustering for diversity
|
v
Order or synthesise a few tens of compounds, assay, confirm hits
```

Each step is a decision a student should be able to defend: the protein preparation and redocking validation of section 3.6; the choice of filters and the actives they might wrongly kill; the enrichment expected at each cut; and the final, essential step off the computer, the assay, without which a virtual hit is only a well formatted opinion. Known traps complete the syllabus: docking into the wrong protonation state, trusting scores as energies, screening databases whose stored stereochemistry is undefined, and forgetting that frequent hitters, reactive and aggregating compounds, game assays as happily as they game scoring functions.

5.9 What this subject adds up to

The five units form one workflow. Parameters turn structures into numbers (Unit 1); QSAR turns numbers into equations and teaches suspicion of them (Unit 2); modeling and docking give the numbers a three dimensional stage (Unit 3); property prediction and the design methods decide what is worth building, including building the target itself when nature has not published it (Unit 4); and pharmacophores with virtual screening scale the whole logic to millions of molecules (Unit 5). The computer's verdict is always a shortlist and a hypothesis. The laboratory still casts the deciding vote, but it votes on a much better ballot.

Frequently Asked Questions

Q: What is computer aided drug design (CADD) in one line? A: CADD is the use of computational techniques, QSAR, modeling, docking, property prediction and virtual screening, to decide which molecules are worth making and testing in drug discovery.

Q: What is the difference between SAR and QSAR? A: SAR is a qualitative account of how structural changes affect activity, while QSAR expresses activity as a mathematical equation in physicochemical parameters, so it can predict untested compounds and be statistically tested.

Q: Why does QSAR use log (1/C) instead of the raw activity? A: Because binding free energy is proportional to the logarithm of potency, so substituent effects add up linearly on the log scale, which is what regression equations assume.

Q: What does a negative coefficient on (log P) squared in a Hansch equation mean? A: It means activity is parabolic in lipophilicity: it rises to an optimum log P₀ and then falls, because very lipophilic compounds dissolve poorly and get trapped in membranes instead of reaching the target.

Q: How do Hansch and Free Wilson analysis differ? A: Hansch analysis regresses activity on physicochemical parameters such as log P and sigma and can predict new substituents, while Free Wilson analysis assigns a fixed activity contribution to each substituent seen in the series and can only rank substituents already tested.

Q: Why is q squared more important than r squared for judging a QSAR model? A: r squared measures fit to the compounds used in building the model and can be inflated by adding terms, whereas q squared comes from cross validation on held out compounds and measures whether the model actually predicts.

Q: What is the difference between the global minimum and the bioactive conformation? A: The global minimum is the lowest energy shape of the free molecule, while the bioactive conformation is the shape it adopts in the binding site; the energy gap between them is a strain the binding interactions must repay.

Q: What do a docking search algorithm and a scoring function each do? A: The search algorithm explores the ligand's positions, orientations and conformations in the binding site, and the scoring function estimates the quality or binding strength of each pose so the poses can be ranked.

Q: What is the point of redocking when validating a docking protocol? A: Redocking removes a ligand from its experimental complex and docks it back; if the program cannot reproduce the known pose, usually within about 2 Angstrom RMSD, its poses for unknown ligands should not be trusted.

Q: When is homology modeling reliable enough to use for drug design? A: As a working rule, models built from templates with more than about 50 percent sequence identity are generally reliable in the core, models at 30 to 50 percent need care and validation, and below about 30 percent the alignment errors usually dominate.

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