PharmacoLearn

MPL203T Principles of Drug Discovery, M.Pharm Semester 2

Unit 2: Lead Identification and Protein Structure

Free notes at pharmacolearn.fyi

2.1 Lead identification: the search for chemical starting points

Lead identification (lead seeking) is the stage that converts a validated, assayable target into chemical matter: compounds with confirmed activity on the target, understood structure, and room to be improved. Its raw materials are compound collections: the accumulated synthetic output of the industry, natural product collections, and libraries built deliberately for screening. Its engines, the three the syllabus names, are combinatorial chemistry, which manufactures candidates at scale; high throughput screening, which tests them at scale; and in silico lead discovery, which searches them, and virtual libraries far larger, by computation before any bench work. All three feed on one prerequisite: a trustworthy assay, which is where this unit begins in practice.

2.2 Assay development for hit identification

An **assay** is the laboratory measurement that reports whether a compound acts on the target. Everything downstream is only as good as this measurement, so assay development is a discipline in itself. The assay must reproduce the relevant biology in miniature: for an enzyme target, the purified protein converting substrate to product; for a receptor, binding of a labelled ligand or a cellular signal downstream of activation; for a cell based question, a phenotype such as growth, death or reporter gene expression in the disease relevant cell.

Assay design choice	Options	Consequences
Biological level	Biochemical (isolated target) vs cell based (target in its cellular context)	Biochemical assays are clean, cheap and fast but blind to membrane permeability and cell toxicity; cell based assays are realistic but noisier and slower, and a hit may act anywhere in the pathway
Readout	Radioactivity, fluorescence, luminescence, absorbance, label free methods	Must be robust at screening scale, cheap per well, and free of interference by the test compounds themselves (coloured or fluorescent compounds are a classic false positive source)
Format	Plate based, miniaturized (384 or 1536 wells), automated liquid handling	Miniaturization cuts cost per test by orders of magnitude and makes million compound screens feasible
Quality control	Positive and negative controls on every plate, signal window, the Z prime statistic comparing control separation with control variability	A Z prime value comfortably above 0.5 marks an assay fit for screening; below that, hit lists are mostly noise

A developed assay is hardened by **counter screens**: parallel tests that discard compounds acting by nuisance mechanisms, such as inhibiting the reporter rather than the target. A **hit** survives this gauntlet: reproducible, concentration dependent activity, confirmed on retest from fresh material, of known structure. The hit rate of a well run screen is a small fraction of one percent of the library, and hits pass to the further filters of section 2.6.

2.3 Combinatorial chemistry

© 2026 PharmacoLearn | pharmacolearn.fyi

Combinatorial chemistry is the systematic synthesis of large sets of related compounds (libraries) by combining sets of building blocks in every, or nearly every, possible combination. If ten variants of one component react with ten of a second and ten of a third, the arithmetic gives a thousand products from thirty building blocks: chemistry by multiplication. Two execution styles dominate:

- **Parallel synthesis**: each compound is made separately, in its own well or vessel, as a single pure product of known identity. Libraries are smaller but every member is identified and immediately testable.
- **Split and mix (split and pool) synthesis**: resin beads are divided into portions, each portion reacts with one building block, the portions are pooled and mixed, and the cycle repeats. Each bead ends carrying one compound, built by its own route through the pools; the library is enormous (hundreds of thousands to millions) but the identity of the compound on a given bead must be decoded or the active mixture deconvoluted by resynthesizing sub libraries.

Combinatorial chemistry transformed lead identification by guaranteeing that a screen always had matter to test, and it pairs naturally with high throughput screening: one manufactures the library, the other interrogates it. Its historical lesson also belongs in the notes: size alone did not raise success, because libraries built for easy synthesis clustered in flat, greasy, similar chemistry that made poor leads. Modern libraries are designed for diversity and for drug like properties, applying the filters of Unit 3 at the drawing board.

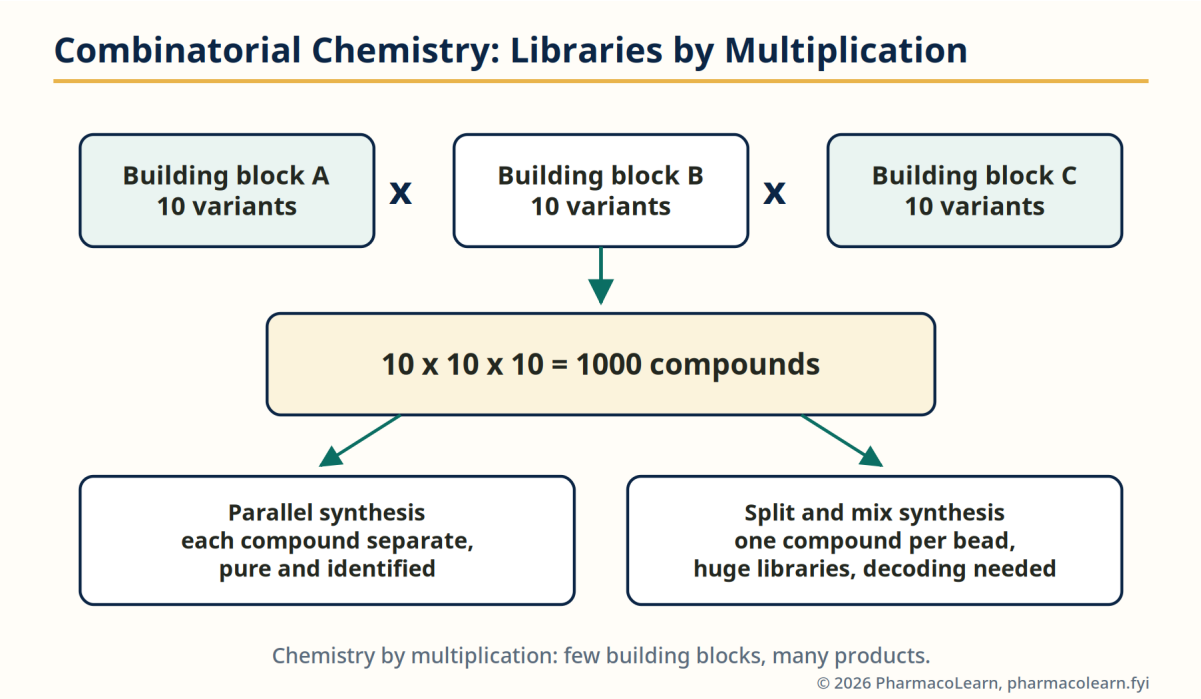

Figure: Combinatorial chemistry multiplies a few building blocks into a library of thousands, made by parallel synthesis or by split and mix.

2.4 High throughput screening

High throughput screening (HTS) is the automated testing of very large compound collections against the assay of section 2.2. Robotic liquid handlers dispense nanolitre volumes into miniaturized plates, incubators and plate readers process them in sequence, and software assembles the results into ranked activity lists. At full industrial scale a screen can test on the order of a hundred thousand compounds in a day, and a full collection in weeks, a rate unthinkable for the manual pharmacology of earlier units.

```
A HIGH THROUGHPUT SCREEN, WELL BY WELL

  Compound library (plates of single compounds in DMSO)
      |
      v
  Robotic dispensing into assay plates -- target + reagents added
      |
      v
  Incubation, then plate reader measures the signal
      |
      v
  Software: % inhibition per well, plate quality controls checked
      |
      v
  Primary hits -> retest in duplicate -> dose response curves
      |
      v
  Counter screens remove nuisance actives
      |
      v
  Confirmed hits -> hit to lead chemistry
```

HTS is powerful and credulous in equal measure, and the syllabus's pairing of it with assay development is the point. It reports whatever the assay measures, including artefacts: compounds that fluoresce like the product, quench the signal, aggregate into colloidal particles that sequester the protein nonspecifically (a notorious source of promiscuous false hits), or react covalently with assay components. Hit triage, with fresh powder retests, dose response curves, detergent controls and orthogonal assays measuring the same biology by a different readout, is therefore part of HTS rather than an afterthought. HTS also returns in Unit 3, inside rational design.

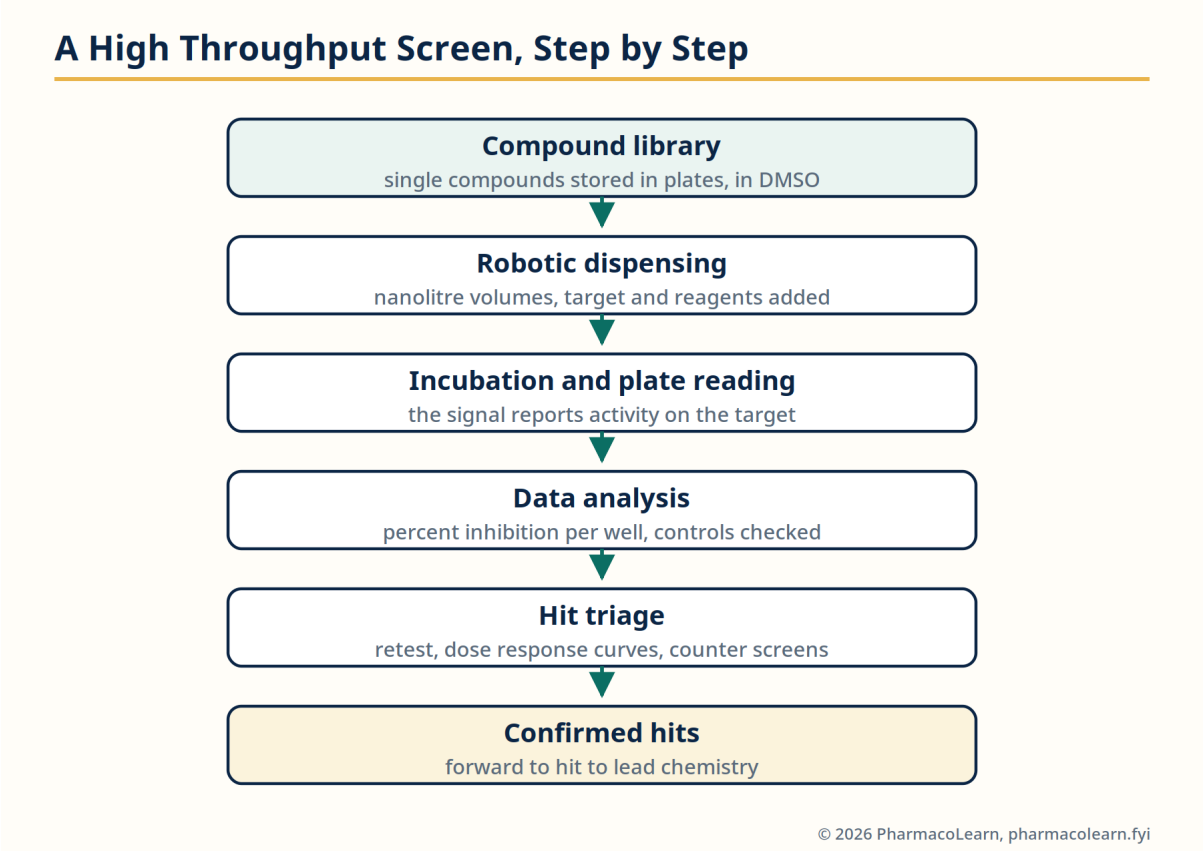

Figure: A high throughput screen, step by step, from stored library plates to confirmed hits after triage.

2.5 In silico lead discovery techniques

In silico lead discovery searches for leads by computation before, or instead of, bench screening. Its main techniques, developed fully in Units 3 and 4, are introduced here as lead seeking methods: **similarity searching**, which finds library compounds resembling a known active; **pharmacophore screening**, which searches for molecules presenting the same arrangement of chemical features as known actives; **molecular docking**, which predicts how each library member would sit in the target's binding site and ranks them by computed fit; and **property filtering**, which removes compounds whose predicted solubility, permeability or toxicity make them hopeless as leads whatever their activity.

The strategic value is leverage. A physical collection numbers in the millions; a virtual collection, enumerated on a computer, can be far larger, and in silico screening costs computer time rather than robotics and reagents. Its discipline is that every prediction is a hypothesis: in silico hits must still be bought or made and tested in the real assay, and the method earns its place by enriching that test set, delivering a far higher hit rate among the compounds actually assayed than random selection would.

2.6 From hits to leads: what a lead must be

Not every confirmed hit deserves a chemistry programme. The **hit to lead** step applies the criteria that separate a lead series from a screening curiosity:

- **Confirmed, tractable potency**: activity strong enough to improve (typically at least micromolar, with a clean dose response), in a series where analogues can be synthesized to explore structure activity relationships.
- **Selectivity at birth**: meaningful preference for the target over its closest relatives and over standard nuisance panels, or at least a structural reason to believe selectivity can be built.
- **No structural alerts**: freedom from reactive, unstable or promiscuously binding chemotypes that will poison optimization later.
- **Drug like physicochemistry**: solubility and lipophilicity in ranges compatible with the drug likeness rules of Unit 3; potency bought with excessive lipophilicity is a debt, repaid later as poor solubility and promiscuity.
- **Novelty and freedom to operate**: a series already patented by a competitor is scientifically interesting and commercially useless.

A series meeting these tests is declared a **lead**, and lead optimization (section 1.1) begins. The rest of this unit supplies the structural knowledge that both hit assessment and optimization increasingly rely on: the three dimensional structure of the target protein.

2.7 Protein structure: why discovery runs on it

Nearly every method in Units 3 and 4 consumes protein structure. Docking needs the coordinates of the binding site; pharmacophore models of a target need to know which residues offer hydrogen bonds; lead optimization explains each potency change by what the structure shows. When the structure of the target itself is unavailable, the structure of a relative may serve, and when none exists, computational prediction (sections 2.11 and 2.12) attempts to supply one. This half of Unit 2 therefore covers what protein structure is, how it is described, and how it is determined or predicted.

2.8 Levels of protein structure

Protein architecture is described at four levels:

Level	Meaning	What holds it together
Primary structure	The amino acid sequence, read from the N terminus to the C terminus	Peptide bonds; sequence alone, determined by gene sequencing and mass spectrometry
Secondary structure	Local regular folding of the chain: alpha helices, beta strands and sheets, turns and loops	Hydrogen bonds between backbone carbonyl and amide groups
Tertiary structure	The complete three dimensional fold of one polypeptide chain	Hydrophobic packing of side chains into a core, hydrogen bonds, ionic interactions, van der Waals contacts, and disulfide bonds between cysteines
Quaternary structure	The assembly of several folded chains (subunits) into one functional complex	The same non covalent forces as tertiary structure, acting between subunits, as in haemoglobin's four chains

The drug binding site is a tertiary (sometimes quaternary) feature: a pocket or groove whose shape and chemistry are created by residues far apart in sequence but adjacent in the fold. That single fact explains both why sequence alone rarely predicts druggability and why the experimental structures of sections 2.13 and 2.14 are so prized.

2.9 Domains, motifs and folds

Between the whole chain and the single residue sit three descriptive units the syllabus names. A **domain** is a segment of the chain that folds independently into a compact structure and usually carries one function: the kinase domain that phosphorylates, the SH2 domain that reads phosphotyrosine marks. Domains are the reusable bricks of protein evolution, and druggable pockets very often sit in or between domains. A **motif** (sequence motif) is a short conserved pattern of amino acids that signals a function: the zinc finger of section 1.10 is a motif, as is the Walker motif of ATP binding proteins; motifs are recognized from sequence and are the working vocabulary of bioinformatics. A **fold** is the overall three dimensional arrangement of secondary structures in a domain (for example the Rossmann fold of many nucleotide binding domains), and unrelated sequences can converge on the same fold while related sequences nearly always share one. The distinction matters practically: fold is conserved longer than sequence, which is the opening that threading (section 2.12) exploits.

2.10 Computational prediction of protein structure: the problem

Experimental structure determination is slow and sometimes impossible (membrane proteins, the largest target class, resist both crystallography and NMR). **Computational prediction** attempts to build a usable model from sequence alone. The methods divide by how much help the sequence receives from known structures. If a related protein of known structure exists, **homology modelling** copies and adapts it. If no relative is detectable by sequence, **threading** tests whether the sequence could adopt any known fold. Both are syllabus methods; both produce models, not structures, and a model's value depends on how it is validated and what question it is asked to answer. A model good enough to suggest which residues line a pocket, and to design the experiment that tests the suggestion, has earned its keep even if its details are approximate.

2.11 Homology modelling

Homology modelling (comparative modelling) builds a model of a target protein from the experimental structure of a homologous protein, the **template**, on the principle that structure is conserved more strongly than sequence: two proteins of clearly similar sequence fold the same way. The standard steps are:

```
HOMOLOGY MODELLING, STEP BY STEP

  1. Template search: find proteins of known structure related
     to the target sequence (sequence database searching)
      |
  2. Sequence alignment: line the target sequence up with the
     template so equivalent residues correspond
      |
  3. Model building: copy the template backbone for aligned
     regions; build the differing loops; place side chains
      |
  4. Refinement: energy minimization settles clashes and strains
      |
  5. Validation: geometry checks, eg the Ramachandran plot,
     should show a protein like model; implausible regions
     are realigned and rebuilt
```

Accuracy tracks relatedness. With a close template (high sequence identity over the whole chain) the backbone of the model approaches experimental usefulness and docking into it is routine; with a distant one, only the conserved core is trustworthy and the loops, often including the binding site rim, carry large errors. The alignment is the step where most error enters: a single shifted residue in the alignment misplaces every atom built on it, which is why experienced modellers spend their effort there and validate before they dock.

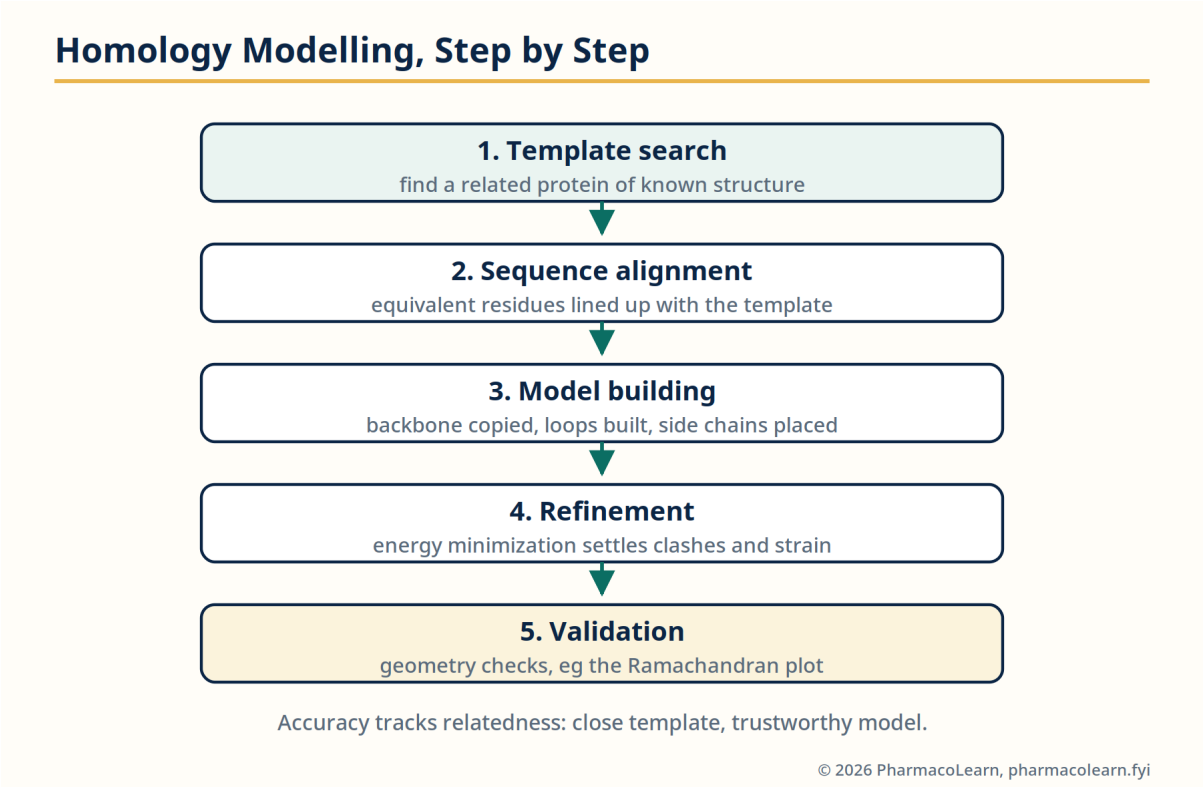

Figure: Homology modelling in five steps. The sequence alignment is the foundation: errors made there cannot be repaired by later refinement.

2.12 Threading and fold recognition

Threading answers the question homology modelling cannot: when no sequence relative with known structure can be found, does the target sequence fit any known **fold** at all? The target sequence is threaded, in imagination, through each fold in a library of known structures, and each trial fit is scored by how happily the sequence sits in that environment: hydrophobic residues buried, polar residues exposed, favourable residue pair contacts. The best scoring fold is adopted as the predicted architecture, and a model is built on it.

Threading trades resolution for reach. It can identify the correct fold family for a protein with no detectable sequence relatives, because structure tolerates sequence change better than alignment statistics tolerate it, but it cannot deliver atom level confidence: side chain placement and loop detail remain speculative. Its discovery use is therefore triage and hypothesis generation, for example predicting that a new target is a kinase fold protein and should be screened and modelled as one, rather than supplying coordinates for precision docking.

2.13 NMR spectroscopy in protein structure determination

Nuclear magnetic resonance determines protein structures in **solution**, the state closest to physiology. In outline: the protein is labelled with the NMR active isotopes nitrogen 15 and carbon 13, a battery of multidimensional experiments assigns each resonance to a specific atom, and experiments that detect through space proximity (the nuclear Overhauser effect) yield thousands of approximate distances between atom pairs. Computational folding then finds the family of structures consistent with those distances, angles and couplings, and the structure is reported as an ensemble of conformers rather than one rigid pose, honestly reflecting molecular motion.

NMR needs no crystals, works in solution, and reports dynamics and weak interactions, which is why it also serves ligand screening directly: methods that watch ligand signals report binding and can locate it on the protein. Its limits are size and labour: proteins much above roughly 30 to 40 kilodaltons broaden into unreadability, and throughput is low. In discovery, NMR structures are most prized for small domains and for the dynamic questions rigid crystal structures cannot answer.

2.14 X-ray crystallography in protein structure determination

X-ray crystallography remains the source of most protein structures in the public databases and the workhorse of structure based design. The protein is persuaded to form an ordered **crystal**, a regular lattice of many copies; X-rays scattered by the crystal produce a diffraction pattern; and computation converts that pattern into an electron density map into which the atomic model is built and refined. The map's sharpness is quoted as **resolution**: at resolutions around 2 angstroms or better, individual atoms and ordered water molecules are placed with confidence sufficient for drug design.

Three features make crystallography a discovery engine rather than merely a measurement. First, **soaking and co crystallization**: crystals of the target with a lead bound show exactly how the drug sits and which interactions it makes or misses, turning each potency puzzle of lead optimization into a structural instruction. Second, throughput: once a crystallization system exists, structures of dozens of analogues can follow in the design cycle. Third, the Protein Data Bank archive supplies the templates for homology modelling and the folds for threading, so crystallography underpins even the computational methods. Its limits are the crystal itself: some proteins refuse to crystallize, flexible regions vanish from the map, and the lattice can hold a conformation the cell never uses, which is why crystallographic claims are cross checked against solution methods where the answer matters.

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