UNIVERSITÀ DI PISA

SCUOLA
NORMALE
SUPERIORE

UNIVERSITÀ DI PISA

SCUOLA NORMALE SUPERIORE

National Ph.D. in Artificial Intelligence

XXXVII cycle

Modeling drug-biosystems interactions at multiple scales through AI methods

Candidate
Francesco Carli

Supervisors
Prof. Francesco Raimondi
Prof. Pietro Liò

Thesis submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Artificial Intelligence

some people are just born to fight, I think
it's not that they're born brave,
it's not that they're born strong,
it's just that the universe has decided that this one,
this one will have grit
and fire and steel in their blood,
and it'll be tested,
this cosmic mettle of theirs.
they'll face trial after trial,
be broken and damaged in countless ways,
but this one was born to fight.
maybe it's not the life they would have chosen,
maybe they'd love to lay down their arms,
but they were born to fight.
it's what they know,
it's what they do best,
it's all they can do.

Summary

Modern drug discovery and precision medicine face the persistent challenge of integrating information across scales, from biochemical drug–protein interactions to cellular responses and ultimately patient outcomes. Bridging these levels of biological organization remains a major barrier to translating preclinical findings into therapies. Artificial intelligence (AI) offers powerful tools to model drug–biosystem interactions, but their impact depends on methods that are accurate, interpretable, generalizable, and accessible. This thesis addresses these needs by developing computational frameworks spanning biochemical, cellular, and patient scales.

At the *biochemical scale*, we developed *BindSight*, a modular framework for drug–target interaction prediction unifying data curation, representation learning, model evaluation, and deployment. It incorporates scaffold-aware splitting, protein promiscuity stratification, and a two-phase prediction scheme: rapid library-wide screening followed by TabPFN re-scoring to balance efficiency with generalization. Central to *BindSight* is a CLIP-style architecture embedding proteins and compounds in a shared latent space, supporting heterogeneous molecular and protein representations, and accommodating advanced loss functions with distributed training.

At the *cellular scale*, we introduced *CellHit*, an interpretable framework that predicts drug responses from transcriptomic profiles of cancer cell lines and extends them to patient tumors. By training on large pharmacogenomic resources (GDSC, PRISM) and aligning them with patient bulk RNA-seq through Celligner, the framework uncovered transcriptional programs underpinning drug sensitivity and recovered known drug–target relationships. Incorporating LLM-curated mechanism-of-action pathways enhanced predictive power. To promote accessibility, *CellHit* has been released as open-source software and deployed as a publicly available web server.

At the *patient scale*,we applied our models to over 10,000 patient transcriptomes from The Cancer Genome Atlas (TCGA), we successfully recovered a majority of approved drug-indication pairs and providing strong *in silico* validation. Importantly, we bridged the gap from computational hypotheses to experimental confirmation through prospective wet-lab experiments, which validated the novel vulnerabilities predicted by our models in pancreatic and glioblastoma cell lines.

In sum, this thesis demonstrates how AI can model drug–biosystem interactions across biochemical, cellular, and patient scales. By combining predictive performance with interpretability, biological grounding, and accessibility, it offers methodological advances, experimentally supported insights, and open resources to accelerate drug discovery and translational medicine. While BindSight is a domain-agnostic tool for drug–target interaction prediction, the subsequent cellular and patient scale work focuses specifically on oncology applications.

Acknowledgements

Questi ringraziamenti sono l'ultima parte che ho scritto di questa tesi e, probabilmente, anche la più difficile. Forse perché li scrivo adesso che sono lontano da tutti voi, e mi mancate tutti come l'aria. Imparando dai matematici, vi metto in ordine alfabetico:

Un grazie sincero al mio supervisor, Francesco. Mia madre dice sempre che non ho un carattere semplice, e tu hai sicuramente avuto modo di accorgertene. Ti ringrazio per la pazienza, per la fiducia e per il tanto, tanto, tanto lavoro che abbiamo condiviso insieme. Un grazie anche tutto il resto del gruppo di bioinformatica!

~~Alby~~ Sukuna: da quel giorno in cui ti ho visto verde nella cucina di casa CC ho capito che saremmo stati brothers 4e. Sei una persona stupenda e in questi anni ti ho visto crescere moltissimo. Ti auguro di raggiungere tutti i tuoi traguardi (il winter arc non finisce mai) e per te ci sarò sempre. Sei un grande ma hai comunque gli occhi neri.

Edo: mio maestro spirituale supremo in questi 3 anni di PhD pagato dai contribuenti pubblici. Illuminami per sempre con la tua luce.

Eli: parte dello zoccolo duro del polvani core, tutti i giorni lì a combattere. Adesso che sto finendo dico solo una cosa: evviva. Avanti tutta amico mio.

Marin: con te ho condiviso tanta scienza quanto drama emotivo e mi manca da morire il momento quotidiano del confessionale. Con te non ci si può sentire soli e porti (troppo) ottimismo ovunque tu vada. Sei unico. Detto questo devi imparare l'italiano e imparare a guidare.

Marta: Ci siamo sempre un po' capiti a metà. Alla fine mi hai insegnato tanto sui miei stupidi intenti. Grazie

Micky: tre anni passati insieme tra brainrot, discorsi filosofici, tramonti, viaggi, cene intime, palestre, playlist trap e di Katy Perry. Casa CC è stata il nostro prime e non avrò mai più un coinquilino così...ma tu rimarrai per sempre mio fratello. Sei il mio Harry preferito.

Pasquale: sei probabilmente una delle persone più brillanti che io abbia mai conosciuto. Mi hai insegnato tantissimo e sono veramente onorato di aver condiviso fatiche, frustrazioni e gioie accademiche con te giorno dopo giorno. Che sei una persona brillante probabilmente lo sai già. Sei però anche un grande amico, con una sensibilità che spesso passa inosservata. Adesso vedi di diventare enorme.

Sofi: sei un pandoro. Ma sei anche molto cute e ti voglio molto bene. Andrà tutto bene e anche per te ci sarò sempre. Rispondimi ai meme su Instagram.

Vitto: grazie per la tua attenzione autentica verso gli altri e per la cura che metti in ogni cosa e in ogni persona che ti sta accanto. Le nostre lunghe conversazioni su cosa conta

davvero e su come restare fedeli a noi stessi mi hanno lasciato tanto. Alla fine però hai ceduto anche tu al politically incorrect, non sarai mica...

Un grazie enorme anche a Fabio, Marco, Ross, Capo, Fede, Gianmarco, Luca, Zeno, Bruna, Giorgio e Francesco. Le nostre gite al mare, gli aperitivi in terrazza e i momenti di convivialità in mensa hanno reso questi tre anni passati insieme veramente indimenticabili. Ci si rivede ogni anno per la cena di Natale!

Un grande saluto anche a tutto il personale della Scuola, dalla mensa alla portineria (in particolare Nadia e Giovanna!). Abbiamo condiviso tanta quotidianità in quella che è diventata per tre anni la mia casa.

Un grazie anche a Vale, Papo, Gio, Ale T., Ale M. e Edo. Siete i miei bros straight outta pianura padana e, pur essendo lontano (ora ancora più lontano), so quanto ci tenete a me e siete le mie radici a casa. Un grazie anche a Gaia, ti lascio il patatino prezioso, trattamelo bene, so dove abiti. Un grazie anche a Ester, ci saremmo visti 4 volte nella nostra vita (la 5a probabilmente mentre leggi questi ringraziamenti) ma per me sei una OG, slay.

Un grazie a tutte le persone incontrate nel mio peregrinare per universita' e laboratori: zia Fede, Ludo e Gabry (in Normale mi hanno sempre considerato "un Bocconiano"). Zia Pia, Pier, Sara e Francesca, la mia famiglia torinese quando ho avuto la brillante idea di prendere una laurea in matematica dopo economia. E Rossano Schifanella, il mio supervisor a UniTO: senza di lui, non avrei fatto un dottorato. Grazie a Ricky, il mio manager che prima o poi riuscirà a riportarmi in Italia, e a Claudia: non saremo mai sul brevetto di Biorek (forse meglio così). Thanks also to the great Doc Leo Celi: I'm spreading the cult of burpees across the world.

Un bacio grande va alla zia Cristina e allo zio Luciano che ci sono stati nel momento in cui ne avevo più bisogno. Senza di loro probabilmente questa tesi oggi non ci sarebbe.

Un grande, grande abbraccio va a mio fratello Giulio, Mery, la piccola Marta e al piccolissimo Giorgio. La vostra vicinanza e il vostro affetto serve sempre a riportarmi per terra, ricordandomi l'affetto di casa e il valore delle cose semplici.

Per ultima, ma non per importanza, voglio ringraziare mia madre Ivana. Se questo traguardo deve andare a qualcuno, va sicuramente a te. Dici sempre di essere orgogliosa di me, ma io lo sono almeno altrettanto di te. Sei il mio esempio più grande e mi hai insegnato dedizione, perseveranza e impegno, tenendo in piedi da sola la famiglia attraverso ogni fatica e dolore. Non seguirò sempre i tuoi consigli (perché un po' metti ansia, eh), ma farò del mio meglio per mettere a frutto tutto l'impegno che ci hai messo, tutto il bene che mi hai voluto e mi vuoi, tutto il tempo che hai sacrificato per me senza mai chiedere niente in cambio.

Voglio chiudere questi lunghissimi ringraziamenti con una riflessione, forse rivolta principalmente a me stesso. Nel corso della mia vita, molte delle mie convinzioni e molti dei miei valori sono cambiati. Ma una cosa è rimasta: l'idea di lasciare questo mondo un po' migliore di come l'abbiamo trovato. E se allora ero un ragazzo con il tempo aperto davanti, oggi, pur con uno sguardo diverso, sono ancora a caccia di vento.

Contents

Part I

Introduction

Chapter 1

Introduction

1.1 Motivations and Challenges

1.1.1 Rise of AI

From Symbolic AI to Foundation Models

Artificial intelligence has undergone significant transformations over the past decades. Research this field began with rule-based systems and progressed to methods capable of learning and identifying patterns in data. This transition from *symbolic AI* to *machine learning*, and more recently to *foundation models*, has provided new opportunities to address complex problems in biology.

Expert Systems and Limitations

Early AI relied on explicit rules crafted by human experts. Systems such as *MYCIN* showed that computers could solve domain-specific problems using these rules [43, 97, 249]. However, they suffered from the *knowledge acquisition bottleneck*: every new task required manual encoding of rules, and performance collapsed when faced with scenarios outside the predefined knowledge base [44].

Rise of Machine Learning

The introduction of *machine learning* shifted this paradigm. Instead of relying on hand-written rules, algorithms began learning patterns directly from data [241]. Classical approaches, including decision trees, support vector machines, and ensemble methods, proved effective across many applications [38, 72, 86]. Yet, these methods depended heavily on *feature engineering*, requiring experts to select the most relevant variables.

Deep Learning and Representation Learning

Deep learning addressed this limitation through *representation learning*, allowing neural networks to automatically extract features from raw data. Importantly, learned representations span multiple levels of abstraction [120]. Figure 1.1 illustrates this concept with a visualization of neurons in a multimodal network, highlighting that some units respond to abstract concepts like *art style* or *person traits* [118]. A milestone was the 2012 ImageNet competition, where deep convolutional networks drastically reduced error rates in image recognition [193, 200].

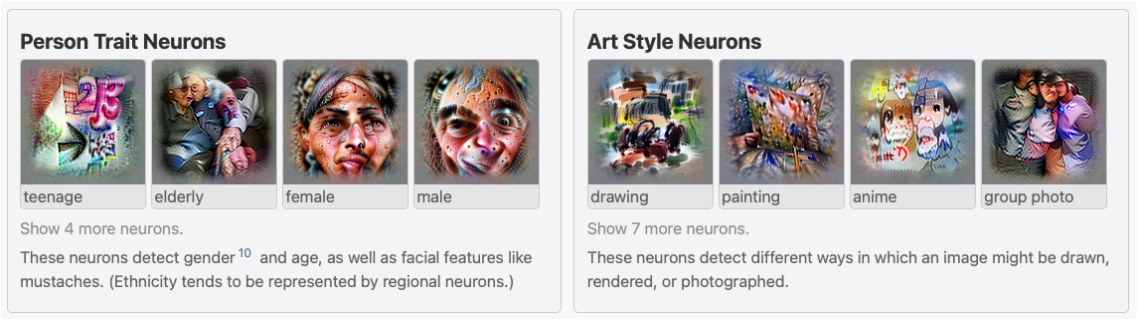

Figure 1.1: **Feature learning in multimodal neurons.** Visualization from Goh et al. [118] showing neurons in a multimodal network that activate for abstract concepts like *person traits* (e.g., age, gender) or *art styles* (e.g., drawing, anime). This demonstrates how deep learning models can discover high-level semantic features directly from data, a key element of representation learning

Foundation Models in NLP

Natural language processing experienced a similar leap with the *Transformer* architecture [368]. This enabled the emergence of *foundation models*, large-scale networks pre-trained on massive text corpora and adapted to diverse downstream tasks [36]. Breakthroughs included *BERT*, which captured bidirectional textual context [82], and the *GPT* family, culminating in GPT-3's ability to perform tasks from instructions alone [42, 281]. These advances follow *scaling laws*, where performance improves with larger datasets, more parameters, and greater computational power [176], as illustrated in Figure 1.2.

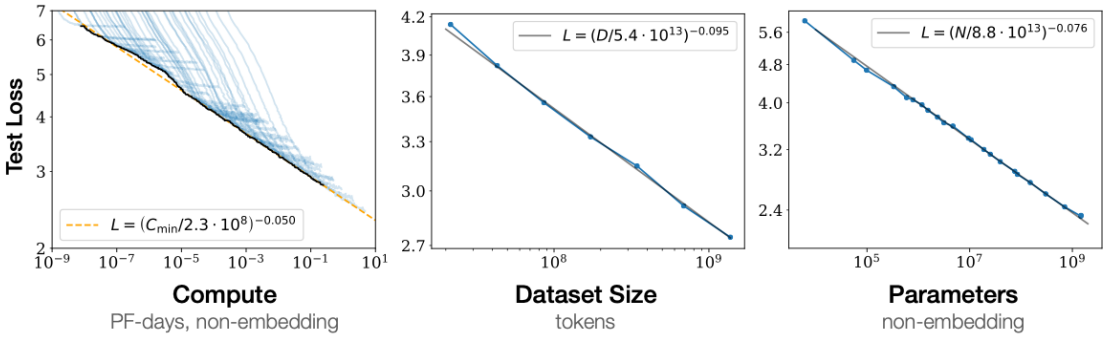

Figure 1.2: **Scaling laws in neural language models.** Visualization from Kaplan et al. [176] showing that test loss decreases predictably as a power law with increased compute (left), dataset size (center), and model parameters (right). These relationships underpin the rapid progress of large-scale foundation models.

Enabling Infrastructure

Three developments enabled this progress. First, the internet provided unprecedented quantities of training data. Second, *GPUs*, originally built for graphics, proved highly effective for neural network computation [263]. Third, specialized hardware such as *TPUs* further increased efficiency [171]. Coupled with distributed computing, these innovations made deep learning feasible for real-world applications.

1.1.2 Promise of AI in the Bioinformatics Field

Biological data challenge

The field of biology is currently confronted with a significant challenge: despite the generation of vast quantities of data, fundamental questions regarding diseases and their treatments remain unanswered [2]. High-throughput *omics* technologies produce massive datasets, but turning this data into valuable knowledge takes too long [335]. Artificial intelligence (*AI*) can help us find patterns and make discoveries.

Data size and complexity

A major obstacle is the size and complexity of biological data [335]. Genomics, transcriptomics, proteomics, and metabolomics each capture distinct aspects of biological systems, generating heterogeneous data that are difficult to integrate [55, 130]. For instance, whole genome and exome sequencing studies identify thousands of disease-associated variants, yet most are located in *noncoding regions* with poorly understood functions [48, 107, 247]. To connect molecular alterations with disease, we need more advanced analytical methods [8].

Causality over correlation

Crucially, we must move beyond correlations to establish *causal* links. Identifying which molecular changes actively cause disease, rather than merely co-occurring with it, requires computational methods able to integrate diverse data types and infer mechanistic relationships [130]. Such mechanistic understanding is central to advancing *precision medicine*.

Cancer research

In oncology, AI models trained on gene expression data have uncovered new cancer subtypes and biological programs. These advances deepen our understanding of tumor behavior and enable predictions of treatment response [51, 382].

Protein structure prediction

The breakthrough of *protein structure prediction* illustrates AI's potential. For decades, predicting a protein's 3D shape from its amino acid sequence was intractable. With AlphaFold, AI achieved near-experimental accuracy for most proteins, solving a problem that resisted fifty years of research [172].

Drug discovery acceleration

Drug discovery stands to benefit profoundly. Traditional development requires more than a decade and billions of dollars, with high failure rates in clinical trials [270]. AI now accelerates each stage: identifying targets from multi-omics data, designing molecules with desired properties, and predicting clinical outcomes [322, 364]. Deep learning even enables computational design of novel compounds that progress from in silico generation to laboratory testing within months [410]. Furthermore, AI supports patient stratification, biomarker discovery, and optimized clinical trial design [51, 358].

Rare disease diagnosis

AI is particularly promising for *rare diseases* [2]. These conditions affect few patients, present heterogeneous symptoms, and are difficult to diagnose. AI can analyze genomic data rapidly, mine medical records, and suggest diagnoses that might otherwise take years [17]. For critically ill infants, genome analysis assisted by AI reduces diagnostic times from weeks to hours, enabling earlier treatment [66]. Initiatives such as the 100,000 Genomes Project show how AI-enhanced genomics could become routine in healthcare [68].

Towards personalized medicine

The convergence of AI and biology holds the potential to transform medicine. By combining genomic, imaging, pathology, and electronic health record data, we may transition from *reactive* treatment to *preventive* care, and from uniform approaches to truly *personalized medicine* [357].Figure 1.3 highlights key applications of AI across the clinical pathway, from embryo selection to hospital risk prediction. Despite persistent challenges, including interpretability of AI decisions and integration into clinical workflows, the trajectory is clear: AI is becoming a central tool in biology, converting data

into knowledge and knowledge into effective therapies [357, 358, 364].

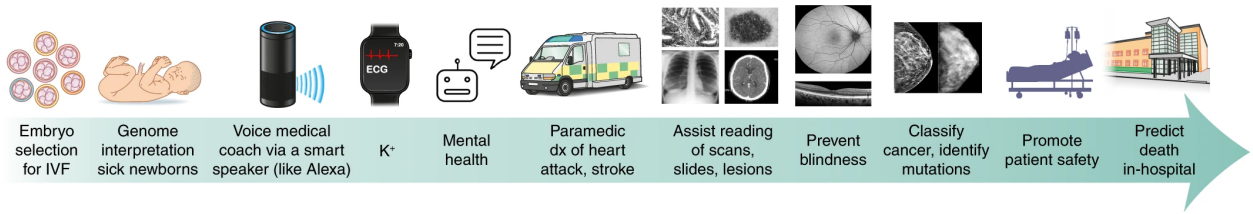

Figure 1.3: **Applications of AI across the clinical pipeline.** Visualization from Topol [357] showing AI's role in multiple areas of medicine, from embryo selection and genomic interpretation to diagnostic imaging support, patient monitoring, and in-hospital risk prediction. These advances reflect the growing potential of AI to enable precision and preventive medicine.

1.1.3 Superhuman Tasks

AI in biology vs. early AI

After discussing the potential of AI in biology, we must recognize a key feature: in this field, AI is tasked with solving problems that humans are unable to perform. This contrasts with early AI applications where computers learned to do tasks humans already could do. This difference changes how we utilize and assess AI in biological research.

Human-comparable tasks

Deep learning first succeeded with tasks humans can perform. Computer vision models learned to recognize objects, significantly reducing error rates on ImageNet [192], and later models even surpassed human accuracy in image classification [134]. Similarly, *language models* learned to translate and understand text, another domain where human benchmarks existed [368]. For these tasks, success was measurable because humans could perform them, and standards for good performance were clear.

Biological complexity

Biology is different. Humans cannot look at an amino acid sequence and know how the protein will fold, cannot mentally test millions of molecules to find drugs, and cannot track how thousands of genes regulate each other. These tasks are not only complex but fundamentally impossible for humans to execute. This opens the space for AI to operate at *superhuman* capacity [63, 240].

Superhuman protein prediction

One prominent example is protein structure prediction. Historically, determining a protein's conformation required long experimental procedures [333, 400] . Computational prediction seemed insurmountable due to the vast configuration space. *AlphaFold2* and other models overturned this assumption, producing predictions nearly as accurate as experiments [20, 172]. Its architecture (Fig. 1.4) integrates evolutionary, template, and sequence-based representations through a deep learning model, achieving superhuman-level accuracy in predicting protein folds. Here, AI is not replicating human performance but performing tasks beyond human capability.

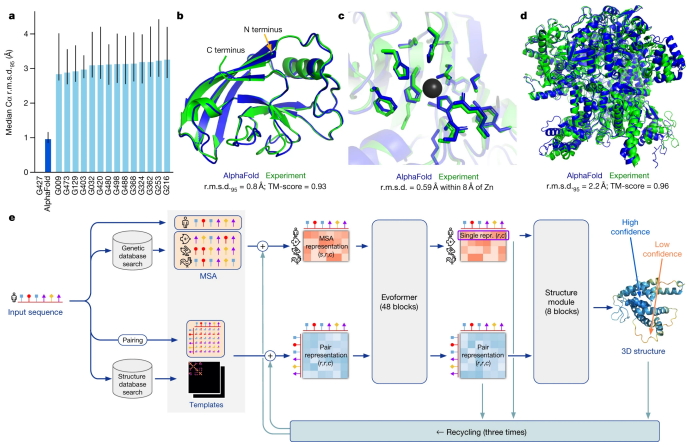

Figure 1.4: **AlphaFold2 architecture and performance.** (**a**) Median C_α r.m.s.d. of AlphaFold predictions compared with experimental structures, showing a clear performance gain over previous methods. (**b–d**) Structural comparisons of AlphaFold predictions (blue) and experimental models (green) at different scales, illustrating near-atomic accuracy. (**e**) Overview of the AlphaFold2 pipeline: input sequences are processed via multiple sequence alignments (MSAs), template searches, and pair representations, which are iteratively refined in the Evoformer and structure modules to generate high-confidence 3D protein structures. From Jumper et al. [172]

Superhuman drug discovery

A second example comes from drug discovery. Using *graph neural networks*, researchers computationally screened over 100 million compounds, something infeasible in the laboratory [336]. The approach (1.5) led to the discovery of *halicin* , a novel antibiotic with an unexpected mechanism of action. This illustrates AI's ability to explore vast chemical spaces inaccessible to human methods.

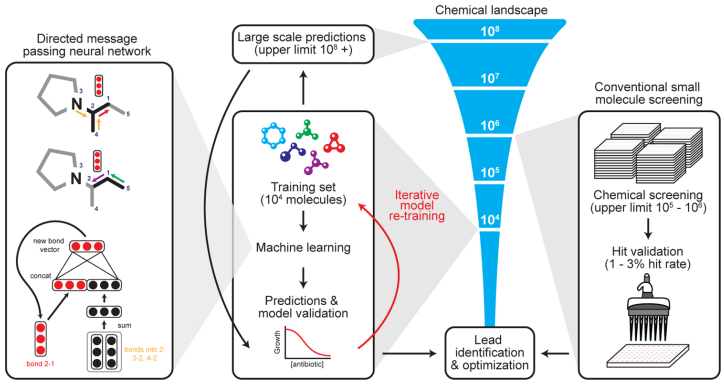

Figure 1.5: **AI-driven antibiotic discovery pipeline.** Graph neural networks (left) encode molecular structures as bond-based representations. A training set of $\sim 10^4$ molecules is used to build predictive models, which are iteratively re-trained and scaled up to screen over 10^8 candidate compounds. Predictions are validated experimentally, leading to lead identification and optimization (bottom). Compared to conventional screening (right), which is limited to 10^5–10^6 molecules with a $\sim$1–3% hit rate, the AI approach enables exploration of much larger chemical spaces and the discovery of novel antibiotics such as halicin. From Stokes et al. [336].

Superhuman single-cell analysis

A third example is single-cell analysis, where models such as *scGPT* [76] operate at unprecedented scale. As illustrated in Fig. 1.6, *scGPT* is first pretrained on a massive cell atlas of over 33 million single cells spanning diverse tissues (panel d), using a masked-attention transformer backbone (panel c) to jointly embed genes and expression levels (panel b). The model is then fine-tuned for downstream tasks such as cell type annotation, clustering, batch correction, perturbation prediction, and network inference. This large-scale training results in clear separation of cell populations in embedding space (panel e) and predictions that rival or surpass expert analyses. Beyond *scGPT*, *CellFM*, trained on 100 million cells with 800 million parameters, operates at a scale beyond human capacity [407]. Processes that once took months of expert analysis can now be automated [392].

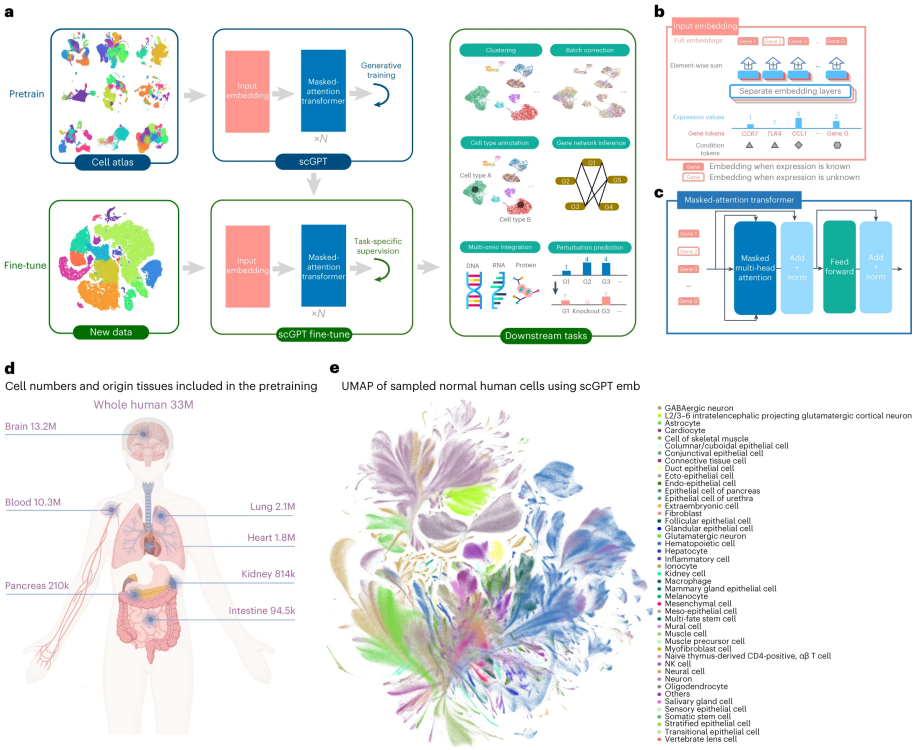

Figure 1.6: **Architecture and applications of scGPT for large-scale single-cell analysis.** (**a**) Overview of pretraining on a large cell atlas and fine-tuning for downstream tasks such as clustering, batch correction, perturbation prediction, and gene network inference. (**b**) Input embedding layers encode both gene identity and expression, allowing flexible handling of unknown values. (**c**) The masked-attention transformer backbone processes the input to learn cell and gene representations. (**d**) Cell numbers and tissues used in pretraining, highlighting coverage of over 33M cells from diverse organs. (**e**) UMAP visualization of single-cell embeddings, demonstrating clear separation of major cell types. From Cui et al. [76].

Verification challenges

Superhuman tasks create new challenges. How do we verify correctness when humans cannot assess the answers? Often, there is no definitive ground truth, or validation requires expensive and lengthy experiments. For example, checking *AlphaFold* predictions may take months of laboratory work, and testing new drug targets requires significant resources. Moreover, the patterns identified by models may be too complex for

humans to interpret directly.

Limits of current models

Being highly skilled at one task does not imply broad reliability. *AlphaFold* excels at static structures but struggles with *dynamic proteins* and disordered regions [201]. Drug screening models can find promising compounds but cannot fully predict their behavior in vivo. A gap persists between computational predictions and biological reality.

Human–AI synergy

AI's value lies in enabling exploration of areas previously unreachable, augmenting rather than replacing human scientists. As illustrated in Fig. 1.7, next-generation research workflows combine human creativity and domain expertise with AI's computational power, automation, and reasoning capabilities. These hybrid systems support everything from data integration and hypothesis generation to automated experimentation, creating a continuous feedback loop between human insight and machine-driven exploration. By shifting routine analysis and large-scale experimentation to AI, researchers are empowered to focus on creativity and strategy, opening new frontiers for biological discovery [109].

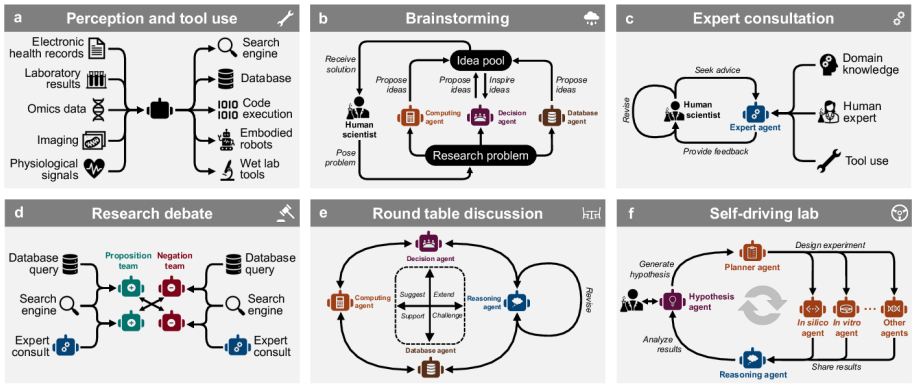

Figure 1.7: **Human–AI synergy in scientific research.** Conceptual illustration of collaborative research ecosystems where humans and AI agents co-create scientific knowledge. AI systems enhance human capabilities by integrating complex data, generating hypotheses, reasoning over evidence, and automating experiments, while humans provide context, creativity, and interpretation. Together, they enable a continuous cycle of discovery at scales unattainable by humans or machines alone. From Gao et al. [109].

1.1.4 The Structure of Biological Data

Multi-scale organization

To leverage artificial intelligence effectively in biology, we must understand what makes *biological data* unique. Biology works at many different scales: atoms form molecules, molecules form proteins, proteins form complexes, complexes form organelles, organelles form cells, cells form tissues, and tissues form organs and organisms [204, 315]. As illustrated in Fig. 1.8, these scales are deeply interconnected, with processes at one level influencing and constraining behavior at others. This multi-level organization creates both opportunities and challenges for artificial intelligence. Crucially, each level exhibits distinct behaviors that not only stand alone but also shape the functioning of other levels [204, 314]. A single DNA mutation can alter the functioning of an entire organism, as classically demonstrated for sickle-cell disease [152, 272]. Understanding these cross-scale dependencies remains one of biology's greatest challenges.

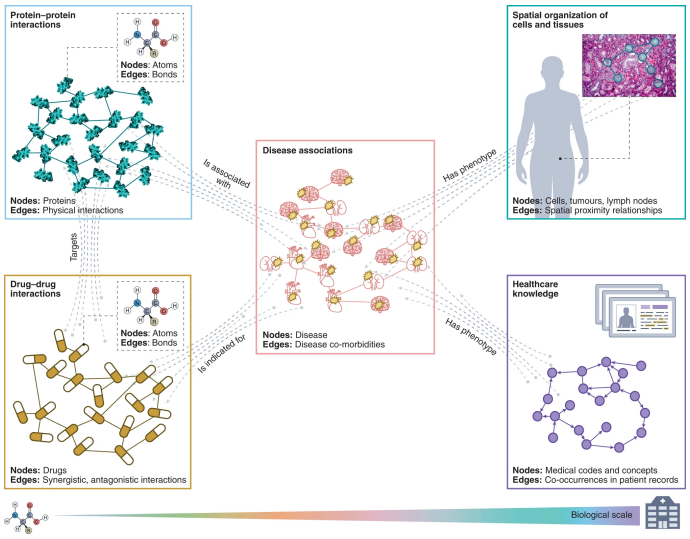

Figure 1.8: **Biology as an interconnected hierarchy of scales.** Biological systems span multiple levels of organization—from atoms and molecules to cells, tissues, organs, and whole organisms. These scales are tightly interlinked, with changes at the biochemical level propagating to influence physiology and disease, and higher-level states feeding back to shape molecular activity. AI models must capture these dynamic, bidirectional relationships to fully understand biological complexity. From Li et al. [204].

Data richness and noise

Modern technologies now allow us to measure biology in great detail, but this creates new challenges. Single-cell sequencing can measure individual cells, however, because cells contain tiny amounts of material, signal amplification introduces technical noise [40, 197]. The resulting data are both valuable and noisy, requiring methods that can recover the *biological signal* despite noise [221].

Diversity of data types

The issue is further complicated by the variety of available data formats. Genomics measures DNA, transcriptomics measures RNA, proteomics measures proteins, metabolomics measures small molecules, and epigenomics measures DNA modifications. Combining these heterogeneous layers can deepen our understanding of disease, but each technology presents distinct challenges [179].

AI for integration

Deep learning provides tools to integrate such diverse data types. Variational autoencoders and related probabilistic models can learn joint representations and help with missing data [113, 219]. More broadly, benchmarking studies show that joint dimensionality-reduction methods can transform heterogeneous datasets into unified representations [49].

Bias and representation gaps

Despite large data collections, biological datasets often lack representativeness. For example, the TCGA cancer database includes roughly 83% of samples from European ancestry [255], while African populations account for less than 2% of analyzed genomes worldwide [389]. Such biases are significant because models trained on unbalanced data often perform poorly on underrepresented groups [59, 279, 331].

Benchmarking pitfalls

Evaluating artificial intelligence models in biology is challenging. Academic benchmarks often use clean, well-organized data that differ substantially from real-world conditions; when models trained on curated datasets encounter clinical or experimental data with missing values, batch effects, or quality issues, their performance frequently degrades [256, 305, 367, 402, 409]. Data leakage between training and test sets can in-

flate results [178], while *shortcut learning* [81, 114] and hidden stratification obscure true performance on clinically relevant subgroups [256]. A similar problem affects protein–ligand binding data: public resources such as ChEMBL [112] and BindingDB [215] are skewed toward a few well-annotated protein families, leaving a long tail of poorly characterized targets [261]. This hub-target concentration inflates performance on random splits but reduces generalization to unseen proteins and ligands [160, 261].

Benchmarking pitfalls

Evaluating artificial intelligence models in biology is challenging. Academic benchmarks often use clean, well-organized data that differ substantially from real-world conditions; when models trained on curated datasets encounter clinical or experimental data with missing values, batch effects, or quality issues, their performance frequently degrades [256, 305, 367, 402, 409]. Figure 1.9 illustrates *shortcut learning*, where models rely on spurious correlations—such as image backgrounds, hospital identifiers, or dataset-specific artifacts—rather than biologically meaningful features [81, 114]. Data leakage between training and test sets can further inflate reported accuracy [178], while hidden stratification obscures performance gaps on clinically relevant subgroups [256]. A similar problem affects protein–ligand binding datasets: public resources such as ChEMBL [112] and BindingDB [215] are skewed toward a few well-annotated protein families, leading to hub-target concentration that inflates performance on random splits but reduces generalization to unseen proteins and ligands [160, 261].

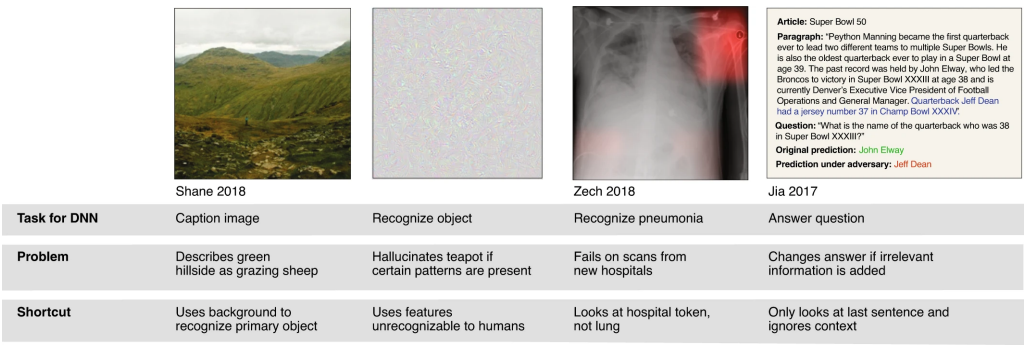

Figure 1.9: **Shortcut learning and dataset biases in AI models.** Examples of models exploiting spurious correlations: (*left*) image captioning systems infer content from backgrounds; (*middle*) medical imaging models rely on hospital-specific tokens; (*right*) NLP models change answers when irrelevant text is added. Such shortcuts inflate benchmark performance but fail to generalize, highlighting the need for robust evaluation strategies. From Geirhos et al. [114].

Key challenges

These issues are not incidental but intrinsic to biological data. In summary, successful artificial intelligence models must address several intertwined complexities, including the presence of multiple scales of organization, the integration of diverse data types, and the prevalence of substantial noise and missing values. They must also contend with biases in data collection, limited sample sizes for specific conditions, and evaluation difficulties that can obscure true model performance.

Future directions

To advance in this field, we require improved methods and higher-quality data. Models must be capable of handling incomplete or biased inputs and provide reliable measures of *uncertainty* [85, 190, 259]. Equally important are diverse datasets that accurately represent all human populations [279, 331]. Progress depends on rigorous evaluation

protocols, with strict reporting standards and external validation [67, 75, 217]. By explicitly addressing these unique characteristics of biological data, we can unlock the full potential of artificial intelligence in biomedicine.

1.1.5 Prior Knowledge and the Bitter Lesson

In computational biology, an important question arises: should we incorporate existing biological knowledge into our models, or should we allow them to learn solely from data? This question significantly impacts how we develop AI systems for biology.

Integrative analysis initiatives

Numerous large-scale initiatives demonstrate the value of *integrative analysis*, systematically assembling and jointly interrogating heterogeneous measurements to extract biological signal. The Cancer Dependency Map (DepMap) [360] exemplifies this strategy by coupling genome-scale perturbation screens with multi-omic features to reveal context-specific vulnerabilities in cancer cells. Likewise, the cBioPortal [52, 108] harmonizes tumor genomics with clinical annotations across many studies, enabling cohort-level pattern discovery and hypothesis generation. The Open Targets platform [257] (Fig. 1.10) adopts a similar principle in the context of drug discovery, integrating genetic association data, molecular profiling, pharmacological evidence, and literature mining to systematically prioritise therapeutic targets across diverse diseases. At broader scales, integrative resources such as TCGA/PCAWG pan-cancer syntheses, the ENCODE encyclopaedias of regulatory elements, and the GTEx cross-tissue eQTL atlas show how curating and analyzing diverse modalities together can yield mechanistic hypotheses, clinically relevant stratifications, and actionable targets [123, 151, 311, 351]. Collectively, these efforts operationalize the premise that many open biological questions can be advanced not only by producing new data, but also by rigorously organizing, harmonizing, and integrating existing knowledge across modalities and studies [130].

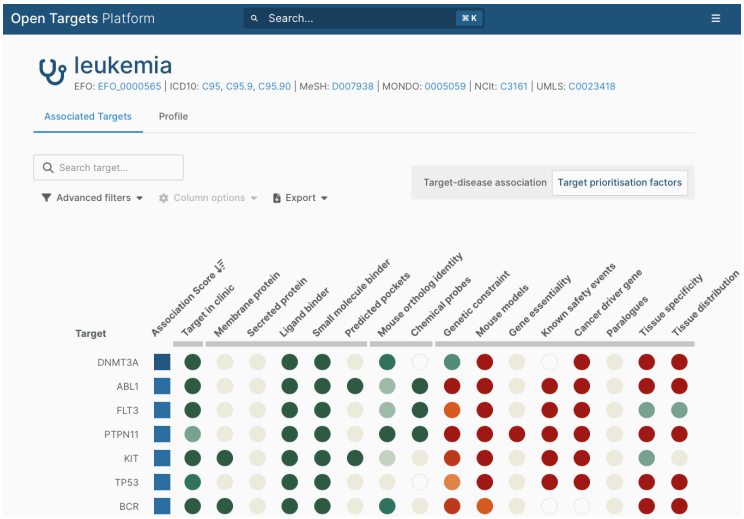

Figure 1.10: **Integrative prioritization of therapeutic targets in the Open Targets platform.** The platform aggregates and scores evidence from diverse modalities, including genetic associations, molecular characterization, pharmacology, and literature mining, to rank potential drug targets for specific diseases. Shown is a target prioritization view for leukemia, illustrating how integrative resources enable systematic hypothesis generation and translational decision-making. From `platform.opentargets.org`

Knowledge as Inductive Bias

Building on such integrative resources, we can design models that encode biological prior knowledge directly into learning algorithms. Network-based approaches operationalize prior knowledge by propagating signals over protein interaction, regulatory, and metabolic graphs to prioritize genes and targets, stratify tumors, and predict drug effects [24, 74, 121]. In this setting, models leverage curated pathways and interactomes rather than re-learning them from scratch, which improves sample efficiency and yields outputs that map onto interpretable modules. This practice aligns with the broader machine learning principle that incorporating domain-specific priors can serve as an *inductive bias*, constraining the hypothesis space toward functionally plausible solutions and thereby improving generalization in data-scarce settings [28]. A concrete example is shown in Fig. 1.11, where a graph convolutional model integrates drug–protein networks to predict adverse polypharmacy side effects [413]. Other methods, such as network-based stratification, cluster patients by mutations affecting shared network regions, revealing clinically meaningful subtypes and pathway-level mechanisms [65, 143].

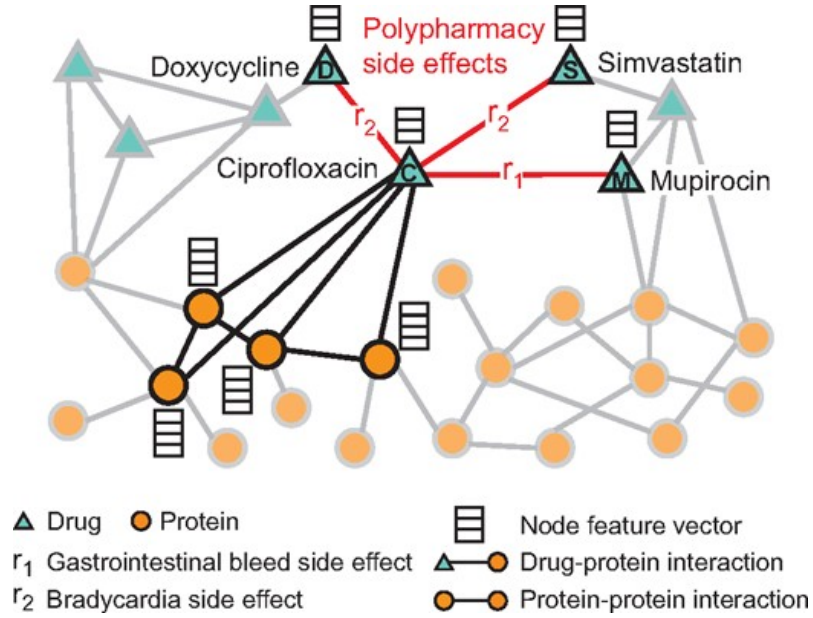

Figure 1.11: **Graph-based modeling of drug interactions.** A heterogeneous network encodes protein–protein interactions, drug–protein targets, and drug–drug interactions. Here, a graph convolutional model predicts adverse polypharmacy side effects by reasoning over network structure and integrating drug and protein feature vectors. Such knowledge-infused models leverage curated biological relationships to improve interpretability and predictive power. From Zitnik et al. [413].

Limits of Current Knowledge

Using existing knowledge presents specific challenges. Our understanding of biology is incomplete and sometimes incorrect. The pathways we currently comprehend represent only a small fraction of how cells function [319]. If we integrate our current understanding into models, we risk overlooking discoveries that might challenge our established beliefs. Many significant advancements in biology have emerged from disproving previous ideas [388]. Taken together, the promise of network-informed models and the incompleteness of current biological knowledge argue for a tempered strategy. Priors are valuable as inductive biases when data are scarce, but because curated path-

ways and interactomes are partial and sometimes wrong, models should benefit from priors without depending on them.

The bitter lesson

This stance aligns with the *bitter lesson* from machine learning: sustained progress has tended to come from general methods that scale with data and compute rather than from hand-engineered domain structure [343]. Simple methods that utilize large amounts of data and computing power often outperform complex methods that incorporate human knowledge. Chess programs stopped using human-designed rules and learned by playing against themselves. Image recognition moved from hand-designed features to learning features from data. Language models replaced grammar rules with training on vast amounts of text. The bitter lesson seems to be: "Let go of preconceived notions and allow models to learn directly from the data." Scaling laws indicate that models improve with increased data, parameters, and computational resources [142, 177]. Why limit models with human knowledge that might be wrong when they can find better patterns themselves?

Biology's unique challenges

Nevertheless, biology has special challenges. Biological data is expensive to generate. Some diseases affect only a small number of people worldwide. New diseases have no historical data. In these cases, we require prior knowledge because we lack sufficient data for pure learning. Moreover, in biology the objective is not always solely predictive accuracy. Often, the ultimate goal is *mechanistic understanding*: to identify determinants, pathways, or causal mechanisms underlying a phenotype [213, 346]. In such scenarios, model interpretability is not merely a desirable property but a primary objective. Here, scaling model size and compute may offer diminishing returns: a large, opaque model might achieve high predictive performance while providing little insight into the underlying biology. Instead, simpler and explainable-by-design approaches, such as sparse linear models, rule-based learners, or causal graphical models, can be preferable because they enable hypothesis generation and experimental validation. In this context, the trade-off between accuracy and interpretability must be considered explicitly, with the choice of modelling strategy guided by whether the aim is to predict or to understand.

AlphaFold2: a hybrid approach

The solution is not to choose one approach or the other, but to combine them [218]. A superb example of combining data and prior knowledge is the AlphaFold2 model [172]. AlphaFold did not rely on explicit human-written rules about how proteins fold. Instead, it was trained on approximately 170,000 experimentally determined protein structures from the Protein Data Bank, supplemented with additional pseudo-labeled examples via self-distillation [172]. The model combines inductive priors, such as triangular and invariant point attention for residue geometry, with the use of multiple sequence alignments and structural templates, and physically inspired loss functions like the Frame Aligned Point Error (FAPE), leveraging powerful large-scale learning. This approach enabled it to learn folding principles directly from data, outperforming purely physics-based methods and subsequently generating a database of over 200 million predicted protein structures, vastly expanding the available structural knowledge [172].

Foundation models and flexibility

This combined approach makes sense for biology. *Foundation models* learn patterns from large datasets [172, 292], identifying relationships that we may not yet be aware of. Then, we adapt these models for specific problems, using prior knowledge as a guide but not a strict rule. The key is to remain flexible, allowing data to override our assumptions when necessary, while leveraging our knowledge when data is insufficient. We should

remain humble about our knowledge and open to what data can teach us. While prior knowledge is valuable when applied thoughtfully, it can obstruct discovery if we regard it as the absolute truth. The most effective AI systems in biology will blend learned patterns with biological understanding, utilizing established science while remaining receptive to new discoveries.

1.1.6 Enhancing the Drug Discovery and Development Pipeline with AI

All the topics we have discussed (AI development, superhuman tasks, biological data, and the knowledge-versus-data debate) are essential for understanding the focus of this thesis: how we can utilize AI in drug discovery and development. The drug development process is both costly and risky. It typically takes more than ten years, multiple stage analyses and billions of dollars to bring a new drug to market (see figure 1.12), and most drugs ultimately fail.[111, 271, 313, 391]. About 90% of drug candidates fail in clinical trials [132, 189]. Even small advancements in predicting which medications will be effective can save significant time and money, while also positively affecting the lives of many patients. It is essential to acknowledge that drugs interact with biological systems at multiple levels, ranging from binding to proteins to influencing the entire body [363]. Understanding all these interactions is crucial for predicting a drug's effectiveness and safety. AI excels at learning these *complex relationships* from diverse types of data, thereby enhancing the efficiency of drug development.

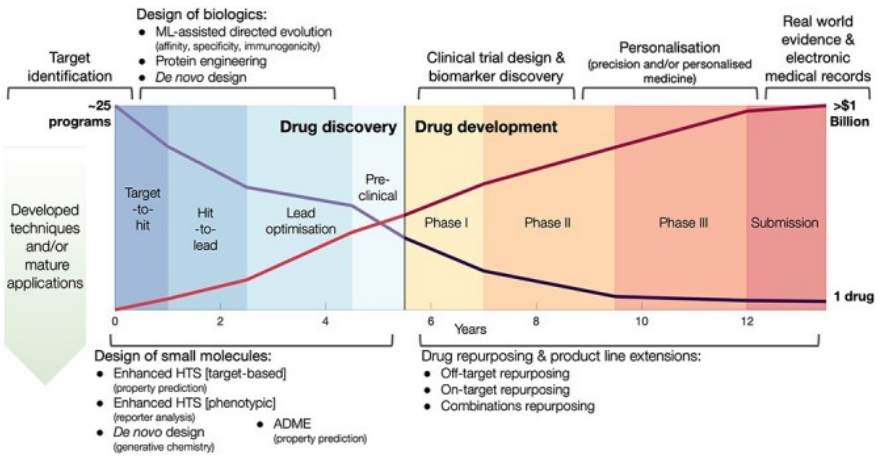

Figure 1.12: **Timeline and challenges of drug discovery and development.** The process spans over a decade, progressing from early-stage target identification and lead optimization through preclinical research, multiple phases of clinical trials, and regulatory approval, with costs often exceeding $1 billion per drug. The figure also highlights opportunities for AI and machine learning to accelerate progress at each stage, from molecular design to patient stratification and post-market analysis. From Gaudelet et al. [111]

Biochemical interactions and binding predictions

At the biochemical scale, drugs are small molecules that bind to proteins and change how they work. At this level, AI helps answer basic questions: Will this compound bind

to the target protein? How strongly? Will it bind to other proteins it should not? Traditional methods use physics-based simulations and docking, but these are slow and not always accurate [187]. Deep learning can learn binding patterns directly from structural data. Recent successes demonstrate that this approach is practical: machine learning has identified new compounds for proteins that were previously considered *undruggable* [228]. AlphaFold's protein structures enable us to screen compounds against proteins that were previously inaccessible for study [172].

Molecular design and optimization

AI models at this scale learn about molecular shape, charge distribution, and movement. They predict not only binding strength but also *selectivity*, the ease of synthesis, and whether the molecule possesses drug-like properties [310, 317]. Some models can design new molecules from scratch, optimizing multiple properties at once [321]. For example, AI found halicin, a new antibiotic, by screening millions of molecular structures [336].

Cellular response prediction

At the cellular scale, drugs alter how cells function. A drug may bind perfectly to its target, but it can still fail if it does not induce the appropriate changes in the cells. AI assists in predicting these cellular effects by analyzing data from experiments where numerous compounds have been tested on cells. Projects like the Connectivity Map [341] have measured the impact of thousands of compounds on gene expression across various cell types. AI models trained on this data can predict how new compounds will affect cells, find new uses for existing drugs, and suggest drug combinations.

Personalized treatment approaches

Cancer cell line databases, such as CCLE [25], the Genomics of Drug Sensitivity in Cancer (GDSC) [153] and PRISM [70], connect molecular features to *drug response*, helping to personalize cancer treatment [110]. Machine learning models integrate genomic, transcriptomic, and proteomic data to predict which patients are likely to respond to a specific drugs. Cell Painting uses images to capture drug effects that gene expression might miss [37]. Competition challenges demonstrate that AI can predict drug response, indicating its potential to work in real clinical settings [73]. Genetic screens help identify targets and understand drug resistance [361].

Drug distribution and toxicity

At the organism scale, successful pharmacotherapy requires delivering therapeutically sufficient drug concentrations to target tissues while minimizing *off-target toxicity*. *Physiologically based pharmacokinetic* (PBPK) models mechanistically describe absorption, distribution, metabolism, and excretion (ADME) across compartments to predict concentration–time profiles in plasma and organs, and are extensively used to extrapolate outcomes in pediatric, elderly, and impaired organ function populations [169, 170]. *Quantitative systems pharmacology* (QSP) extends PBPK by coupling pharmacokinetics to disease-relevant signaling and physiological pathways, facilitating mechanistic translation from exposure to pharmacodynamic response [353]. AI can enhances both frameworks by calibrating uncertain parameters and population variability from clinical and real-world datasets, generating computationally efficient surrogates for scenario exploration, and leveraging high-dimensional omics and imaging data to forecast safety risks such as drug-induced liver injury [277, 412].

Preclinical validation models

Animal models provide important organism-level data. Zebrafish let us test thousands of compounds for efficacy and toxicity in a whole organism [232]. Patient-derived xenografts in mice help us understand how human tumors respond to drugs [140]. AI can be leveraged combine these different data sources to predict clinical outcomes from preclinical experiments, potentially reducing *late-stage failures* [230].

Multi-scale integration

Connecting the various scales is where AI could demonstrate its true value. A drug's molecular structure influences how it interacts with cells, which in turn affects the overall response of the entire organism. AI models that comprehend these connections could outperform those that analyze each scale in isolation. By utilizing structural information, these models could make more accurate predictions about the effects on cells. Additionally, cellular data could assist in selecting molecules with the appropriate mechanisms. Organism-level models could help differentiate between drugs that fail to reach their target and those that simply do not work.

1.2 Research Questions and Contributions

The research questions addressed in this thesis span three scales of drug–biosystem interactions:

1.2.1 Molecular level

Is it possible to design a fast and accurate framework to identify new drug candidates for a given target?

We address this question through the design of *BindSight*, a modular framework that integrates the entire pipeline of ligand–target interaction prediction, from data curation and feature extraction to scalable inference. A key innovation is the adoption of a two-phase prediction strategy: a lightweight screening module that rapidly filters vast chemical libraries into a tractable subset, followed by a computationally intensive re-scoring stage that prioritises the most promising candidates. This design balances efficiency and accuracy, enabling both high-throughput virtual screening and more refined prioritisation steps. By supporting multiple protein and ligand representations, and by enabling precomputation of embeddings for similarity search, the framework ensures adaptability across targets and ease of deployment in lightweight environments such as Google Colab. This question is addressed in section 5.2.

Can we design a thorough data curation pipeline to address biases and inconsistencies typical of drug–target interaction data?

We answer this question by implementing a rigorous preprocessing pipeline tailored to the heterogeneity of chemogenomic resources such as BindingDB and Offensperger et al. [258]. The pipeline applies systematic filtering of raw records to enforce structural validity, metadata completeness, and consistent protein annotation. Chemical structures undergo normalization across tautomeric forms, protonation states, and stereochemistry, while multiple molecular formats are generated to support robust representation learning. Binding affinity values are harmonized across Ki, Kd, and IC_{50} measurements through conversion to $-\log_{10}(\mathrm{M})$ scales, with outlier detection and aggregation strategies reducing assay noise. Proteins are further enriched with UniRef clusters and Pharos metadata to support biologically informed stratification. Finally, intelligent negative sampling ensures balanced and biologically plausible training sets by excluding potential false negatives and enforcing chemical diversity. Collectively, this pipeline mitigates systematic biases and inconsistencies while preserving maximal coverage, thus providing a

reproducible and high-quality foundation for model training. This question is addressed in section 5.3.

Is it possible to set up a learning and validation strategy to effectively learn a classifier able to recover interesting new drug candidates for less annotated proteins?

We explore this question by introducing a stratified evaluation scheme and a multi-objective optimization strategy explicitly designed to handle the heterogeneity of protein annotation levels. Proteins are grouped into quantile-based promiscuity categories (Q_1–Q_3), reflecting their number of known ligands and thus their degree of prior characterization. To avoid scaffold leakage and ensure balanced distributions across protein families, binding outcomes, and promiscuity groups, we developed a greedy scaffold-aware splitting algorithm that yields unbiased train/validation/test partitions. On this foundation, we apply focal loss to mitigate extreme class imbalance and employ Optuna-based multi-objective hyperparameter optimization, simultaneously maximizing predictive performance on both data-rich (Q_3) and poorly characterized (Q_1) proteins while controlling overfitting across folds. This strategy allows us to rigorously assess whether predictive models generalize to less studied targets, providing a principled approach for identifying novel candidate ligands in areas of unmet biomedical need. This question is addressed in section 5.4.

1.2.2 Cellular level

To what extent can transcriptomic profiles alone capture drug sensitivity across cell lines and patients, compared to joint drug-cell feature models?

We approach this question by systematically comparing joint drug-cell models with per-drug models trained exclusively on transcriptomic features. Through the development of *CellHit*, we assess whether gene expression alone provides sufficient signal to capture drug response, or whether incorporating chemical descriptors yields measurable gains. By benchmarking both strategies on large-scale resources such as GDSC and PRISM, we establish the relative predictive power of transcriptomics in isolation, thereby clarifying the trade-off between simplicity, interpretability, and generalizability in drug sensitivity modeling. This question is addressed in chapter 6.1.

Can interpretable machine learning models recover general biological principles of drug sensitivity, such as core essential genes and tissue-specific dependencies, beyond nominal targets?

We address this question by leveraging model interpretation and explainable AI methods, combining SHAP values and permutation-based importance to systematically extract the transcriptional features driving drug response predictions. By analyzing the recovery of known targets and mechanisms of action, we assess whether models capture relevant biological signals rather than spurious associations. Beyond drug-specific effects, we investigate whether the most predictive genes align with experimentally defined tissue-dependent essential genes, demonstrating that interpretable models can reveal convergent biological principles of sensitivity. Together, these analyses allow us to evaluate the capacity of predictive frameworks to move beyond target recovery and un-

cover broader determinants of cellular response. This question is addressed in chapter 6.3.

How can prior knowledge (e.g., MOA pathways) be systematically incorporated into models to improve interpretability and accuracy without over-constraining discovery?

We tackle this question by creating a pipeline that utilizes large language models to curate associations between mechanisms-of-action (MOAs) and drugs. This process involves selecting relevant pathways related to drug responses in a scalable and systematic way, which allows us to construct structured biological priors. Prior knowledge obtained in this way is then exploited in two complementary directions. On the one hand, the systematic annotation of pathways to drugs enables the interpretation and validation of results emerging from standard models trained on genome-wide features. In this setting, gene-level importance scores, typically difficult to contextualize, are translated into pathway-level evidence, thereby providing more human-understandable and actionable insights. On the other hand, the curated pathways are directly integrated into predictive frameworks as inductive biases, giving rise to *MOA-primed models*. This dual strategy allows us to assess whether embedding prior knowledge can enhance both interpretability and accuracy. This question is addressed in chapter 6.2, 6.3 and 6.5

1.2.3 Translational level

How well can preclinical models generalize to patients, and what strategies best align cell-line transcriptomics with patient tumors to enable clinically relevant predictions?

We investigate the extent to which preclinical models can generalize to patient tumors and how alignment strategies enable clinically relevant predictions. By deploying drug-specific models on TCGA transcriptomes after alignment through a framework enhanced by the Celligner approach, we perform large-scale inference across ∼10,000 patient samples. This allows us to recover a good fraction of patient samples matching approved drug indications, while also surfacing systematic opportunities for drug repurposing and rational combinations. These contributions illustrate that aligning cell-line and patient transcriptomics effectively, along with robust modeling, bridges the translational gap and enables large-scale *in silico* predictions. This question is addressed in chapter 6.6.

What is the translational validity of AI-based predictions when prospectively tested in wet-lab assays?

To assess whether our predictive frameworks can inform real-world therapeutic decisions, we complement in silico validation with prospective experimental assays. Specifically, we test AI-derived predictions in pancreatic ductal adenocarcinoma (PDAC) and glioblastoma multiforme (GBM) models, two aggressive tumor types with limited treatment options. By validating subtype-specific and patient-specific drug sensitivities in matched cell lines and primary cultures, we directly examine the extent to which predictions translate into measurable biological responses. This prospective validation pro-

vides a rigorous benchmark for translational applicability, grounding computational results in experimentally verified evidence. This question is addressed in chapter 6.7.

How can we design accessible, end-to-end platforms that make advanced drug-response predictions usable by non-experts?

To address this question, we developed the CellHit webserver, a public resource that enables transcriptomics-based drug-response predictions through a fully automated pipeline. By integrating large-scale pharmacogenomic resources with robust preprocessing, alignment, and model inference, the platform allows users to upload bulk RNA-seq data and obtain predictions without coding expertise. Interactive visualizations, including low-dimensional embeddings, heatmaps, and gene-level attributions, make results interpretable and suitable for exploratory analysis, while open-source software modules ensure transparency, reproducibility, and reuse. These contributions demonstrate how predictive methods can be deployed as accessible, end-to-end platforms, broadening their utility beyond specialized computational settings. This question is addressed all across chapter 7.

What preprocessing and alignment strategies are required to ensure robustness when handling real-world transcriptomic data of varying quality and origin?

To address this question, we combined methodological advances in preprocessing, alignment, and embedding. We developed an enhanced version of *Celligner* that prevents information leakage and optimizes alignment quality through a neighborhood-consistency objective, thereby enabling robust comparison of patient tumors with cell lines. To ensure stability and reproducibility of samples low-dimensional embeddings, we introduced a *Parametric UMAP* model that projects new data into a fixed reference space without altering its global structure. Finally, we designed a *robust pre-processing stack* that couples batch correction via pyComBat with a machine-learning–based gene imputation module, allowing the harmonization of heterogeneous or incomplete expression profiles. Together, these strategies establish a reliable foundation for deploying predictive models in translational settings where input data are noisy, heterogeneous, and often incomplete. This question is addressed in chapters 7.1, 7.2 and 7.4.

What role do interpretability and built-in quality control play in fostering trust and adoption of predictive models in translational oncology?

To promote confidence in model outputs and encourage adoption in clinical and translational contexts, we integrate interpretability and diagnostic modules directly into our predictive framework. SHAP-based local explanations highlight gene-level drivers of drug sensitivity for each sample-drug pair, enabling users to trace predictions back to biologically plausible signals. Complementing this, kernel density estimation (KDE) diagnostics distinguish selective therapeutic effects from general cytotoxicity, ensuring that prioritized compounds display meaningful specificity. Finally, interactive heatmaps provide cohort-level quality control by revealing global patterns and outliers. Together, these built-in mechanisms transform raw predictions into transparent, context-rich outputs, allowing researchers and clinicians to critically assess reliability and biological plausibility, thereby lowering the barriers to practical deployment. This question is ad-

dressed in chapter 7.6.

1.3 Publications

We list below the journal papers that form the basis of this thesis published by the time of writing:

- [51] F. Carli, P. Di Chiaro, M. Morelli, C. Arora, L. Bisceglia, N. De Oliveira Rosa, A. Cortesi, S. Franceschi, F. Lessi, A. L. Di Stefano, et al. Learning and actioning general principles of cancer cell drug sensitivity. *Nature Communications*, 16(1): 1654, 2025.

- [50] F. Carli, N. De Oliveira Rosa, S. Blotas, P. Di Chiaro, L. Bisceglia, M. Morelli, F. Lessi, A. L. Di Stefano, C. M. Mazzanti, G. Natoli, et al. Cellhit: a web server to predict and analyze cancer patients' drug responsiveness. *Nucleic Acids Research*, page gkaf414, 2025.

We will highlight when a chapter or (sub)section of a chapter is based on a published work. The following other publications during the PhD studies do not contribute to the contents of this thesis: Matic et al. [234], Arora et al. [16], and Azzarello et al. [18].

Matic et al. [234] introduces PRECOGx, a machine-learning framework based on protein language model embeddings to predict GPCR interactions with G proteins and β-arrestins, enabling mechanistic insights, variant impact assessment, and visualization of signaling determinants through a publicly available web server

Arora et al. [16] presents a pan-cancer computational analysis of GPCR–ligand and GPCR–biosynthetic enzyme signaling networks, revealing their dysregulation patterns, prognostic value, and potential as drug targets, with experimental validation of selected inhibitors' anti-cancer activity.

Azzarello et al. [18] presents an XGBoost-based method to identify α and β cells in intact, living human pancreatic islets from label-free infrared autofluorescence images, enabling high-precision, non-invasive cell-type recognition for longitudinal studies without immunostaining.

Part II

Preliminaries

Chapter 2

Methodological background

This chapter establishes the methodological foundations of the thesis. We summarize the computational approaches, algorithms, and analytical frameworks employed, providing the technical context needed to interpret results in subsequent chapters. The discussion is selective rather than exhaustive, highlighting core concepts and directing readers to primary references for deeper study.

2.1 Predictive models

2.1.1 XGBoost

XGBoost (eXtreme Gradient Boosting) is a scalable, regularized gradient-boosted decision tree (GBDT) framework, widely regarded as a state-of-the-art method for structured/tabular data [60]. It can address multiple supervised learning tasks, including regression, classification, and ranking, by selecting suitable objective functions. Together with *LightGBM* and *CatBoost*, it constitutes the family of modern tree-based boosting methods that dominate benchmark leaderboards for tabular problems [122, 328].

At iteration t, XGBoost predicts:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(\mathbf{x}_i), \quad f_t \in \mathcal{F},$$

where η is the learning rate and $\mathcal{F}$ is the space of regression trees. The regularized objective is:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l\left(y_i, \hat{y}_i^{(t)}\right) + \sum_{k=1}^{t} \Omega(f_k), \quad \Omega(f) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2,$$

where T is the number of leaves, w_j the leaf weights, and γ, λ regularization terms. Using a second-order Taylor expansion around $\hat{y}_i^{(t-1)}$, we obtain:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} \left[g_i f_t(\mathbf{x}_i) + \frac{1}{2} h_i f_t^2(\mathbf{x}_i) \right] + \Omega(f_t),$$

where g_i and h_i are the first and second derivatives of the loss with respect to $\hat{y}_i^{(t-1)}$. For a tree partitioned into leaves $I_1, \ldots, I_T$, the optimal leaf weight is:

$$w_j^* = -\frac{\sum_{i \in I_j} g_i}{\sum_{i \in I_j} h_i + \lambda},$$

and the corresponding optimal objective reduction is:

$$\mathcal{L}_{\text{opt}}^{(t)} = -\frac{1}{2} \sum_{j=1}^{T} \frac{\left(\sum_{i \in I_j} g_i\right)^2}{\sum_{i \in I_j} h_i + \lambda} + \gamma T.$$

Tree splits are selected by maximizing the gain:

$$\text{Gain} = \frac{1}{2}\left[\frac{(G_L)^2}{H_L + \lambda} + \frac{(G_R)^2}{H_R + \lambda} - \frac{G^2}{H + \lambda}\right] - \gamma.$$

Efficiency is achieved through sparsity-aware split finding, weighted quantile sketching, and cache-aware data layout [60]. Large-scale empirical studies consistently show that tree-based boosting methods like XGBoost outperform most deep-learning approaches on tabular data across regression and classification tasks, while being faster to train and requiring less hyperparameter tuning [122, 328]. Two aspects are particularly important for this thesis. First, XGBoost provides native GPU acceleration[1], which is crucial for scaling to high-dimensional *omics* matrices (often 10^4–10^5 features) without prohibitive runtimes. Second, tree ensembles admit fast, *exact* SHAP value computation via TreeSHAP/TreeExplainer, which leverages dynamic programming along decision paths to produce per-sample attributions in polynomial time while preserving local accuracy and consistency [225]. These capabilities make XGBoost a strong default for scalable and interpretable models in our *omics*-driven analyses.

2.1.2 Multi-layer perceptrons (MLPs)

A *Multi-Layer Perceptron* (MLP) is a class of feed-forward artificial neural networks composed of an input layer, one or more hidden layers, and an output layer [145, 298]. Each layer consists of neurons that apply an affine transformation followed by a non-linear activation, enabling the network to approximate complex, non-linear mappings between inputs and outputs. MLPs are universal function approximators under mild conditions, making them suitable for a wide range of supervised learning tasks including regression, classification, and time series prediction.

For an input vector $\mathbf{x} \in \mathbb{R}^d$, a layer l with weight matrix $W^{(l)}$ and bias vector $\mathbf{b}^{(l)}$ computes:

$$\mathbf{h}^{(l)} = \sigma\left(W^{(l)}\mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}\right),$$

where $\mathbf{h}^{(0)} = \mathbf{x}$ and $\sigma(\cdot)$ is a non-linear activation function (e.g., ReLU, sigmoid, tanh). The output layer applies a task-dependent transformation, such as the softmax function for multi-class classification:

$$\hat{\mathbf{y}} = \text{softmax}\left(W^{(L)}\mathbf{h}^{(L-1)} + \mathbf{b}^{(L)}\right).$$

[1] https://xgboost.readthedocs.io/en/stable/gpu/index.html

MLPs are trained by minimizing a differentiable loss function $l(y, \hat{y})$ over the parameters $\{W^{(l)}, \mathbf{b}^{(l)}\}$ using stochastic gradient descent (SGD) or its variants [299], with gradients computed via backpropagation [302]. Regularization techniques such as weight decay, dropout, and batch normalization improve generalization and training stability.

Although MLPs can model arbitrary functions given sufficient depth and width, empirical evidence shows that they are often outperformed by tree-based boosting methods like XGBoost on tabular data [122, 328], while excelling in high-dimensional, unstructured domains such as images, audio, and natural language.

2.1.3 Graph Neural Networks

Message-passing. Throughout this chapter we will follow [129] to deliver a formal definition of Graph Neural Networks (GNNs), complementing with additional resources where needed.

The key idea is that we want to design a deep learning architecture able to work natively on graph-structured data and generate nodes, edges and whole graph embeddings that depend on graph topology and any feature that might decorate each of these elements.

Let us consider graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ together with nodes' features $\boldsymbol{X_v} \in \mathbb{R}_v^{d \times |\mathcal{V}|}$. We want to use this information to generate new nodes embeddings $\boldsymbol{z_u}$ for all $u \in \mathbb{V}$. In order to do so we will leverage a procedure called *Neural Message Passing* [117] which involves an exchange of vector messages between nodes followed by an update function usually encoded by neural networks. Given a node $u \in \mathcal{V}$, we define its graph neighborhood as

$$N(u) = \{j \mid (i, j) \in \mathcal{E}\}.$$

Then we obtain a hidden embedding $\boldsymbol{h}_u^{(k)}$ as follows:

$$\boldsymbol{h}_u^{(k+1)} = \text{UPDATE}^{(k)}\left(\boldsymbol{h}_u^{(k)},\ \text{AGGREGATE}^{(k)}(\{\boldsymbol{h}_v^{(k)},\ \forall v \in \mathcal{N}(u)\})\right)$$

where UPDATE and AGGREGATE are arbitrary differentiable functions (usually neural networks). One key remark is that the AGGREGATE takes a set as input, therefore, GNNs are *permutation equivariant* by design [**?**]. In other words, the aggregation procedure can be defined as the "neural message" $\boldsymbol{m}_{\mathcal{N}(u)}^{(k)}$ and has to be insensitive with respect to neighbors node order. This is usually achieved by decomposing the AGGREGATE operator as follows:

$$\boldsymbol{m}_{\mathcal{N}(u)}^{(k)} = \text{AGGREGATE}^{(k)}(\{\boldsymbol{h}_v^{(k)},\ \forall v \in \mathcal{N}(u)\}) = \bigoplus_{\forall v \in \mathcal{N}(u)} \psi(\boldsymbol{h}_u^{(k)}, \boldsymbol{h}_v^{(k)})$$

where ψ represents a feed-forward neural network and $\bigoplus$ represents a permutation-invariant aggregator such as sum, max and average.

The forward computational procedure can be therefore summarized as follows:

1. For a given k iteration of the GNN, the AGGREGATE function takes as input the set of embeddings of the nodes in u's neighborhood $\mathcal{N}(u)$ and generates message $\boldsymbol{m}_{\mathcal{N}(u)}^{k}$;

2. The UPDATE function combines message $\boldsymbol{m}_{\mathcal{N}(u)}^{k}$ together with $\boldsymbol{h}_u^{(k-1)}$ to generate the updated embeddings $\boldsymbol{h}_u^{(k)}$.

We remark how for $k = 0$ the initial embeddings are set to be the input features for all nodes:

$$\boldsymbol{h}_u^{(0)} = \boldsymbol{x}_u, \ \forall u \in \mathcal{V}.$$

After running K iterations of the GNN message passing we can use the output of the final layer to define the embeddings for each node as

$$\boldsymbol{z}_u = \boldsymbol{h}_u^{(K)}, \ \forall u \in \mathcal{V}.$$

The basic intuition behind the message-passing framework is that at each iteration every node aggregates both *structural* and *feature-based* information from every node belonging to its neighborhood. At the first iteration $k = 1$ each node will pool information from nodes that can be reached with paths of length 1, whereas with subsequent iterations $k \in \mathbb{N}$ it will be able to reach all nodes at distance k (see figure 2.1).

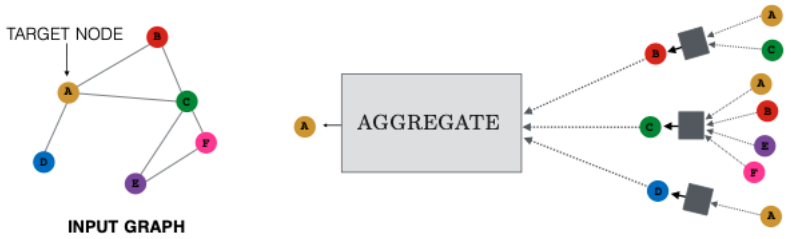

Figure 2.1: Graphical representation of a two-layer message passing model. Source: [129]

Different flavours of GNNs The message-passing framework describes in general terms how a GNN works. However, different implementations of the UPDATE and AGGREGATE operators result in distinct models with distinct expressive power [395]. In more depth, according to the aggregation function, we can identify three classes of graph neural networks [369] (see figure 2.2):

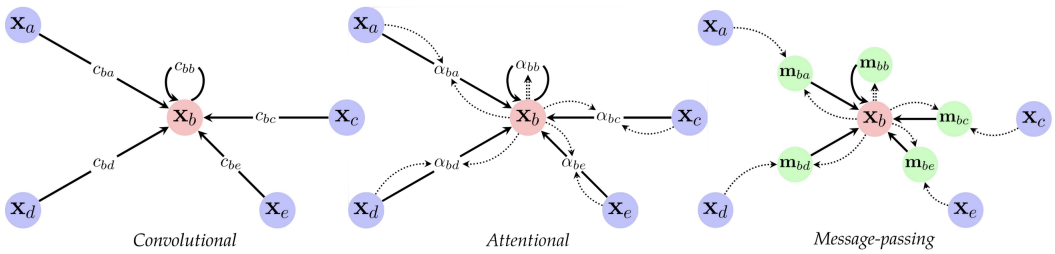

Figure 2.2: Different types of AGGREGATE functions. Source: [369]

Convolutional [186] Usually useful for graphs where edges encode label similarity and computationally efficient. Features of neighbours are aggregated with respect to fixed weights c_{ij}:

$$\boldsymbol{h}_i^{(k)} = \phi\left(\boldsymbol{h}_{\boldsymbol{i}}^{(k-1)}, \bigoplus_{\forall j \in \mathcal{N}(i)} c_{ij}\psi\left(\boldsymbol{h}_{\boldsymbol{j}}^{(k-1)}\right)\right).$$

Attentional [370] Used as middle ground with respect to statistical capacity and computational feasibility. Can be applied also when edges do not encode similarity. Features of neighbors are aggregated with implicit weights (via attention):

$$\boldsymbol{h}_i^{(k)} = \phi\left(\boldsymbol{h}_{\boldsymbol{i}}^{(k-1)}, \bigoplus_{\forall j \in \mathcal{N}(i)} a(\boldsymbol{h}_{\boldsymbol{i}}^{(k-1)}, \boldsymbol{h}_{\boldsymbol{j}}^{(k-1)})\psi\left(\boldsymbol{hf}_{\boldsymbol{j}}^{(k-1)}\right)\right).$$

Message-passing [28, 117] Most generic version of GNN layer. On the one hand these layers are optimal for computational chemistry, reasoning and simulation. On the other hand they are difficult to scale and may incur in learnability issues. Messages are computed with arbitrary functions as introduced above:

$$\boldsymbol{h}_i^{(k)} = \phi\left(\boldsymbol{h}_{\boldsymbol{i}}^{(k-1)}, \bigoplus_{\forall j \in \mathcal{N}(i)} \psi\left(\boldsymbol{h}_{\boldsymbol{i}}^{(k-1)}, \boldsymbol{h}_{\boldsymbol{j}}^{(k-1)}\right)\right).$$

GNNs in action Graphs are pervasive in biology and medicine, from molecular interaction maps to population-scale social and health interactions [203]. With the multitude of bioentities and associations that can be described by networks, they are prevailing representations of biological organization and biomedical knowledge [414]. As a result, the ability to natively model all graph-structured biomedical discoveries to date in a unified framework, matching the inductive biases specified by the input data structure, has vigorously driven the development of the GNNs architectures introduced earlier.

Recently GNNs have been employed successfully at different biological scales [203]:

Molecular level Molecular structure is translated from atoms and bonds into nodes and edges, respectively. Physical interactions or functional relationships between proteins also naturally form a network. This is probably the area in which GNNs have been most successfully applied, pushing the state-of-the-art in tasks such as antibiotics discovery [336], de-novo molecular generation [166], quantum chemistry modelling [117] and molecule property prediction [398].

Genomic level Genetic elements are incorporated into networks by extracting coding genes' co-expression information from transcriptomic data. Single-cell and spatial molecular profiling have further enabled the mapping of genetic interactions at the cellular and tissue level. Here GNNs have been used to predict phenotypes from gene-expression data [233], embed single-cell RNA data [45], impute missing transcriptomic data [377], and combine gene expression data with cell spatial information [403].

Therapeutics level At this scale networks are composed of drugs, proteins, and diseases to allow the modeling of drug-drug interactions, binding of drugs to target proteins, and identification of drug repurposing opportunities. In this category we find powerful applications of GNNs able to predict polypharmacy effects [413], predict drug-target affinity [163, 252] and drug repurposing [275].

2.1.4 CLIP architecture

Contrastive Language–Image Pre-training (CLIP) was introduced as a dual-encoder framework that aligns images and text by learning a shared embedding space with a contrastive objective [161, 283]. Abstracting away the concrete modalities, the same design is *modality-agnostic*: any two data types that can be paired—such as compounds and proteins—can be embedded by separate encoders and brought into correspondence through contrastive learning. Recent work has applied closely related ideas to drug–target interaction (DTI) prediction at scale, demonstrating that contrastive co-embeddings can recover binding relationships and generalize to unseen pairs [330].

Encoders and embeddings. Let $\mathcal{M}_A$ and $\mathcal{M}_B$ denote two modalities. We define encoders $f_\theta : \mathcal{M}_A \to \mathbb{R}^d$ and $g_\phi : \mathcal{M}_B \to \mathbb{R}^d$ that map inputs to a common d-dimensional space:

$$\mathbf{h}_i = f_\theta(a_i), \qquad \mathbf{u}_j = g_\phi(b_j).$$

Optionally, linear projections $W_A, W_B \in \mathbb{R}^{d\times d}$ are applied and the vectors are ℓ_2-normalized:

$$\mathbf{z}_i^A = \frac{W_A\mathbf{h}_i}{\|W_A\mathbf{h}_i\|}, \qquad \mathbf{z}_j^B = \frac{W_B\mathbf{u}_j}{\|W_B\mathbf{u}_j\|}.$$

Normalization makes cosine similarity the natural score and stabilizes the temperature-scaled logits used during training [283].

Contrastive similarity and loss. Given a minibatch of N aligned pairs $\{(a_i, b_i)\}_{i=1}^N$, we form a similarity matrix with temperature $\tau > 0$:

$$s_{ij} = \frac{\langle \mathbf{z}_i^A, \mathbf{z}_j^B \rangle}{\tau}.$$

Training minimizes a symmetric InfoNCE objective that treats in-batch mismatched pairs as negatives:

$$\mathcal{L}_{A\to B} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(s_{ii})}{\sum_{j=1}^{N}\exp(s_{ij})}, \qquad \mathcal{L}_{B\to A} = -\frac{1}{N}\sum_{j=1}^{N} \log \frac{\exp(s_{jj})}{\sum_{i=1}^{N}\exp(s_{ij})},$$

$$\mathcal{L}_{\text{total}} = \tfrac{1}{2}\big(\mathcal{L}_{A\to B} + \mathcal{L}_{B\to A}\big).$$

This objective pulls matched pairs together while pushing mismatched pairs apart, producing a geometry where cross-modal correspondence is captured by proximity [260, 283].

Retrieval and zero-shot inference. At inference, a query from modality A is embedded as $\mathbf{z}^A$ and compared to a candidate set $\{\mathbf{z}_k^B\}$ from modality B; nearest-neighbor search in the shared space yields cross-modal retrieval or matching [181, 283]. In practice, similarities $s_k = \langle \mathbf{z}^A, \mathbf{z}_k^B \rangle / \tau$ act as logits for ranking or downstream probabilistic decisions.

Using CLIP for classification. Beyond retrieval, the same similarities can be turned into a classifier. Suppose we have K labeled classes represented in modality B by *class prototypes* $\{\mathbf{w}_k\}_{k=1}^K$ (each $\mathbf{w}_k$ can be a single labeled example, a textual prompt embedding, or the average of multiple labeled examples). Given a query a from modality A,

$$\text{score}_k(a) \;=\; \frac{\langle \mathbf{z}^A(a), \mathbf{w}_k \rangle}{\tau}, \qquad P(y{=}k \,|\, a) \;=\; \frac{\exp(\text{score}_k(a))}{\sum_{\ell=1}^K \exp(\text{score}_\ell(a))}.$$

This yields a zero-shot or few-shot classifier whose decision rule is a softmax over cross-modal similarities [283]. For binary decisions on a specific pair (a, b), a simple and effective alternative is to treat the similarity $s = \langle \mathbf{z}^A(a), \mathbf{z}^B(b) \rangle$ as a logit and calibrate a sigmoid

$$P(y{=}1 \,|\, a, b) \;=\; \sigma(\alpha\, s + \beta),$$

with (α, β) fitted on a validation set. Both approaches retain the interpretability of the embedding geometry while providing calibrated class probabilities.

2.2 Fundational models

2.2.1 Transformer architecture

The Transformer architecture [368] is a deep neural network model based on *self-attention* mechanisms. Originally introduced for sequence-to-sequence tasks in natural language processing, it has since become the foundation for a wide range of AI systems, including language models, vision transformers, and multimodal architectures. Transformers remove recurrence and convolution in favor of parallelizable attention operations, enabling efficient large-scale training.

Self-attention Introduced to tackle natural language processing tasks, transformer models [82, 282, 368] take as input a collection of n d-dimensional elements $\boldsymbol{X} \in \mathbb{R}^{n \times d_{\text{in}}}$ and transform it into another collection $\boldsymbol{Z} \in \mathbb{R}^{n \times d_{\text{emb}}}$. The original transformer model [368] is composed of two modules: an *encoder* and a *decoder*. Since we will be mainly concerned with BERT-like [82] models, we introduce here only the encoder module.

Transformer models process their inputs in parallel through a series of blocks that alternate the **self-attention** mechanism with feed-forward connections [368].

The self-attention block is the main workhorse of the transformer model and is obtained by combining three elements: keys $\boldsymbol{K} \in \mathbb{R}^{n \times d_{\text{emb}}}$, queries $\boldsymbol{Q} \in \mathbb{R}^{n \times d_{\text{emb}}}$ and values $\boldsymbol{V} \in \mathbb{R}^{n \times d_{\text{emb}}}$. These quantities are computed as

$$\boldsymbol{K} = \boldsymbol{X}\boldsymbol{W}_K, \qquad \boldsymbol{Q} = \boldsymbol{X}\boldsymbol{W}_Q, \qquad \boldsymbol{V} = \boldsymbol{X}\boldsymbol{W}_V$$

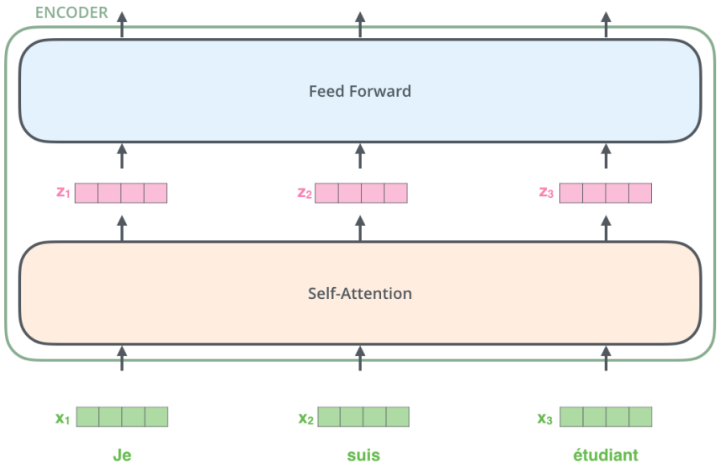

Figure 2.3: The transformer block. Source: [368]

with $\boldsymbol{W}_K, \boldsymbol{W}_Q, \boldsymbol{W}_V \in \mathbb{R}^{d_{\text{in}} \times d_{\text{emb}}}$. Then, to pass from input embeddings $\boldsymbol{X} \in \mathbb{R}^{n \times d_{\text{in}}}$ to output embeddings $\boldsymbol{Z} \in \mathbb{R}^{n \times d_{\text{emb}}}$, we apply

$$\boldsymbol{Z} = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_{\text{emb}}}}\right)\boldsymbol{V}. \tag{2.1}$$

By calling

$$\boldsymbol{A} = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^\top}{\sqrt{d_{\text{emb}}}}\right),$$

we can rewrite (2.1) as

$$\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{V}.$$

Fixing $h \in \{1, \dots, n\}$, the h-th row of $\boldsymbol{Q}\boldsymbol{K}^\top$ is

$$\left[\boldsymbol{q}_h \boldsymbol{k}_1^\top, \dots, \boldsymbol{q}_h \boldsymbol{k}_n^\top\right] \qquad \text{where } \boldsymbol{q}_h \boldsymbol{k}_i^\top = \sum_{j=1}^{d_{\text{emb}}} [\boldsymbol{Q}]_{h,j} [\boldsymbol{K}]_{i,j}.$$

Since $\boldsymbol{Q}$ and $\boldsymbol{K}$ are linear transformations of $\boldsymbol{X}$, each row of $\boldsymbol{Q}\boldsymbol{K}^\top$ collects interaction scores (dot products) between a fixed query and all keys. After scaling by $\sqrt{d_{\text{emb}}}$ and applying a row-wise softmax we obtain $\boldsymbol{A}$. The updated embedding for element h is the h-th row of $\boldsymbol{Z}$:

$$\boldsymbol{z}_h = \sum_{j=1}^{n} [\boldsymbol{A}]_{h,j} \boldsymbol{v}_j,$$

i.e., a convex combination of value vectors $\{\boldsymbol{v}_j\}_{j=1}^n$ with weights from $\boldsymbol{A}$. Thus, self-attention builds context-aware representations by mixing information across positions (see Figure 2.4).

Finally, the transformer block illustrated in Figure 2.3 is a simplified version of [368], which also features skip-connections [135] and norm layers [19] (see Figure 2.5).

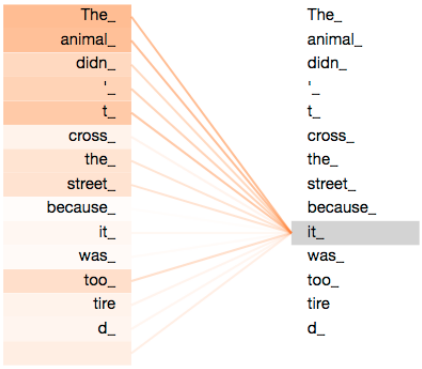

Figure 2.4: Graphical representation of an attention pattern.

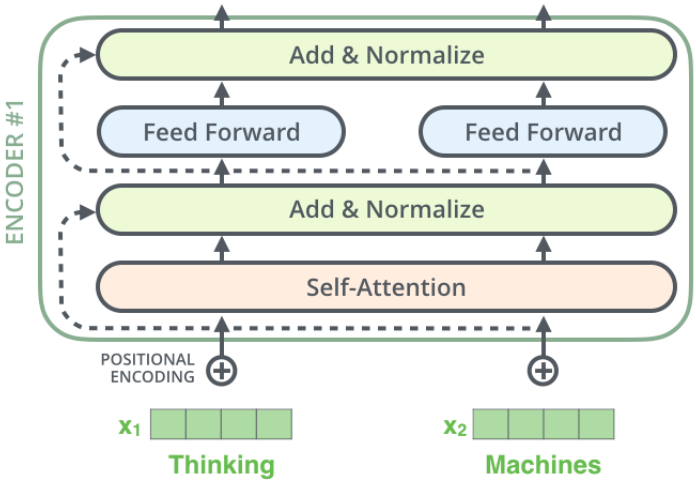

Figure 2.5: Complete version of the transformer encoder block.

Heuristic Interpretation Each query vector $\boldsymbol{q}_t$ (row of $\boldsymbol{Q}$) asks: *"Which tokens are most relevant to me?"* The similarity $\boldsymbol{q}_t \cdot \boldsymbol{k}_s$ is an inner product; if vectors are length-normalized, it approximates cosine similarity. The softmax maps these scores to a distribution over positions, which is used to compute a weighted sum of the corresponding value vectors $\boldsymbol{v}_s$. This implements a differentiable nearest-neighbor retrieval; different heads learn different similarity notions.

Embedding and Positional Encoding Given a length-n tokenized sequence $x_{1:n}$ from a vocabulary $\mathcal{V}$ of size $|\mathcal{V}|$, let

$$\boldsymbol{E} \in \mathbb{R}^{|\mathcal{V}| \times d_{\text{in}}}$$

be the learned embedding matrix. The input embeddings are

$$\boldsymbol{X} = \big[\boldsymbol{E}[x_1]; \ldots; \boldsymbol{E}[x_n]\big] \in \mathbb{R}^{n \times d_{\text{in}}}.$$

Because self-attention is permutation-invariant over positions, positional information must be added. In the original Transformer, *sinusoidal* encodings $\boldsymbol{P} \in \mathbb{R}^{n \times d_{\text{in}}}$ are defined

as

$$P_{t,2i} = \sin\left(\frac{t}{10000^{2i/d_{\text{in}}}}\right), \quad P_{t,2i+1} = \cos\left(\frac{t}{10000^{2i/d_{\text{in}}}}\right),$$

and the block input is

$$\boldsymbol{H}^{(0)} = \boldsymbol{X} + \boldsymbol{P}.$$

Learned positional embeddings are also widely used.

Multi-Head Attention To attend to information from different subspaces, the Transformer uses *multi-head attention* (MHA). Let h be the number of heads and $d_{\text{head}} = d_{\text{in}}/h$. For head i,

$$\text{head}_i \;=\; \text{Attn}\Big(\boldsymbol{H}\boldsymbol{W}_Q^{(i)},\, \boldsymbol{H}\boldsymbol{W}_K^{(i)},\, \boldsymbol{H}\boldsymbol{W}_V^{(i)}\Big),$$

with $\boldsymbol{W}_Q^{(i)}, \boldsymbol{W}_K^{(i)}, \boldsymbol{W}_V^{(i)} \in \mathbb{R}^{d_{\text{in}} \times d_{\text{head}}}$, and

$$\text{MHA}(\boldsymbol{H}) \;=\; \text{Concat}\big(\text{head}_1, \ldots, \text{head}_h\big)\, \boldsymbol{W}_O,$$

where $\boldsymbol{W}_O \in \mathbb{R}^{(h\, d_{\text{head}}) \times d_{\text{in}}}$ and each head uses the same attention as in (2.1) with scaling $\sqrt{d_{\text{head}}}$.

Position-Wise Feed-Forward Networks After MHA, each position is processed independently by the same feed-forward network (FFN):

$$\text{FFN}(\boldsymbol{h}) = \boldsymbol{W}_2\, \sigma(\boldsymbol{W}_1 \boldsymbol{h} + \boldsymbol{b}_1) + \boldsymbol{b}_2,$$

where σ is typically ReLU or GeLU [138], with $\boldsymbol{W}_1 \in \mathbb{R}^{d_{\text{in}} \times d_{\text{ff}}}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{d_{\text{ff}} \times d_{\text{in}}}$.

Residual Connections and Layer Normalization Each sub-layer is wrapped with a residual connection and layer normalization [19]. In the *pre-layer norm* variant:

$$\boldsymbol{H}' = \boldsymbol{H} + \text{MHA}(\text{LN}(\boldsymbol{H})), \qquad \boldsymbol{H}'' = \boldsymbol{H}' + \text{FFN}(\text{LN}(\boldsymbol{H}')).$$

Residuals facilitate gradient flow, while normalization stabilizes training.

Model Variants The original Transformer was designed as an *encoder–decoder* architecture, in which the encoder applies bidirectional self-attention without causal masking, while the decoder combines causal self-attention, implemented through a triangular mask, with cross-attention over the encoder outputs [368]. Later developments adapted this blueprint into two widely used specializations. Encoder-only architectures, such as BERT [83], retain only the encoder block and exploit fully bidirectional attention. These models are trained with the *masked language modeling* (MLM) objective, where a random subset of tokens is replaced with a special `[MASK]` symbol and the model is asked to predict the original values. Decoder-only architectures, exemplified by GPT [41], discard the encoder and operate solely through autoregressive decoding. By enforcing causal masking, each token is conditioned only on its left context, and the model is optimized using the *next token prediction* (NTP) objective.

Language Modeling Objectives In the case of *next token prediction*, the model is trained to maximize the likelihood of the sequence $x_{1:T}$ by factorizing it autoregressively:

$$\log p(x_{1:T}) = \sum_{t=1}^{T} \log p(x_t \mid x_{<t}),$$

where causal masking ensures that $M_{ts} = -\infty$ whenever $s > t$. With output logits $z_t = W_{\text{lm}} h_t$, the loss takes the form

$$\mathcal{L}_{\text{NTP}} = -\sum_{t=1}^{T} \log \frac{\exp(z_{t,x_t})}{\sum_{w \in \mathcal{V}} \exp(z_{t,w})}.$$

This autoregressive formulation makes NTP particularly suited for text generation.

In contrast, *masked language modeling* corrupts the input by replacing a subset $M \subset \{1, \ldots, T\}$ of tokens with a `[MASK]` token. The model then learns to recover only the missing tokens:

$$\mathcal{L}_{\text{MLM}} = -\sum_{t \in M} \log p(x_t \mid x'_{1:T}),$$

where x' denotes the corrupted sequence. Because attention is bidirectional, the model can leverage both left and right context when predicting masked elements. Unlike NTP, which directly targets generative capabilities, MLM is primarily used to produce contextual representations that can be transferred to downstream tasks.

Practical Distinctions In practice, NTP aligns with generative modeling, while MLM produces bidirectional representations optimized for fine-tuning in downstream applications.

Computational Complexity Self-attention is $O(n^2)$ in sequence length n, due to the $QK^\top$ matrix. FFN layers cost $O(n\, d_{\text{model}}\, d_{\text{ff}})$. Training employs AdamW optimization with learning-rate warmup; residual connections and normalization are essential for stability.

2.2.2 Large Language Models (LLMs)

Objective and scaling. Given a tokenized sequence $x_{1:T}$ and parameters θ, autoregressive LLMs minimize

$$\mathcal{L}_{\text{pre}}(\theta) = -\mathbb{E}_{x \sim \mathcal{D}} \sum_{t=1}^{T} \log p_\theta(x_t \mid x_{<t}),$$

with perplexity $\text{PPL} = \exp\big(\frac{1}{T}\sum_t -\log p_\theta(x_t \mid x_{<t})\big)$. Empirically, loss obeys power-law scaling in parameters N, tokens D, and compute C, with compute-optimal training recommending roughly $D \propto N$ under a fixed budget [142, 177].

Instruction tuning and alignment. Post-pretraining, LLMs are adapted to follow instructions via supervised fine-tuning (SFT)

$$\mathcal{L}_{\text{SFT}}(\theta) = -\tfrac{1}{n}\sum_{i=1}^{n} \log p_\theta\big(y^{(i)} \mid x^{(i)}\big),$$

and preference-based alignment. RLHF optimizes a KL-regularized objective with a learned reward model r_ϕ and reference policy π_{ref}, while DPO replaces RL with a direct logistic objective over preference pairs, often matching RLHF quality with simpler training [262, 284, 383].

Dense vs. Mixture-of-Experts (MoE). *Dense* models activate all parameters per token; *MoE* uses conditional computation:

$$y \;=\; \sum_{i \in \text{Top-}k} g_i(x)\, E_i\big(h(x)\big),$$

where a router g selects experts $\{E_i\}$. MoE raises capacity at near-constant per-token FLOPs but requires load-balancing and careful training; both dense and MoE scale effectively at frontier sizes [323].

Retrieval-augmented generation (RAG). RAG conditions generation on retrieved documents $D(x)$ using a retriever $s(q,d) = \phi(q)^\top \psi(d)$ and a generator trained by

$$\mathcal{L}_{\text{RAG}}(\theta) = -\mathbb{E}_{(x,y)} \sum_{t} \log p_\theta\big(y_t \mid y_{<t}, x, D(x)\big),$$

with early/iterative or late fusion variants (e.g., RAG, REALM, FiD) improving factuality and freshness without retraining [125, 156, 180, 202].

2.2.3 Protein Language Models (PLMs)

General Overview. Protein language models (PLMs) employ Transformer-based architectures—originally successful in natural language processing to model amino acid sequences as a "language" [58]. Using self-supervised learning on massive collections of protein sequences (e.g., UniRef, BFD), these models infer structural, functional, and evolutionary patterns without explicit labels. Just as words carry meaning in sentences, PLMs uncover latent semantics in protein "sentences" [384]. They drive progress in tasks like function prediction, structure modeling, and protein design, often outperforming traditional evolutionary profile-based methods [57]. For a thorough introduction to PLMs, see Bepler and Berger [33]. *Figure 2.6* contrasts the three canonical language-modeling paradigms (autoregressive, bidirectional, and masked) that underpin most PLM objectives.

Notation and Common Architecture. Given a protein sequence $s = (a_1, \ldots, a_n)$ from the amino acid set $\mathcal{A}$, we embed tokens via matrix $E \in \mathbb{R}^{|\mathcal{A}| \times d}$, yielding

$$X = [E[a_1]; \ldots; E[a_n]].$$

With positional encoding, a stack of L Transformer blocks produces hidden states:

$$H^{(0)} = X + \text{PosEnc}, \quad H^{(\ell+1)} = \text{Block}(H^{(\ell)}) \quad (\ell = 0, \ldots, L-1).$$

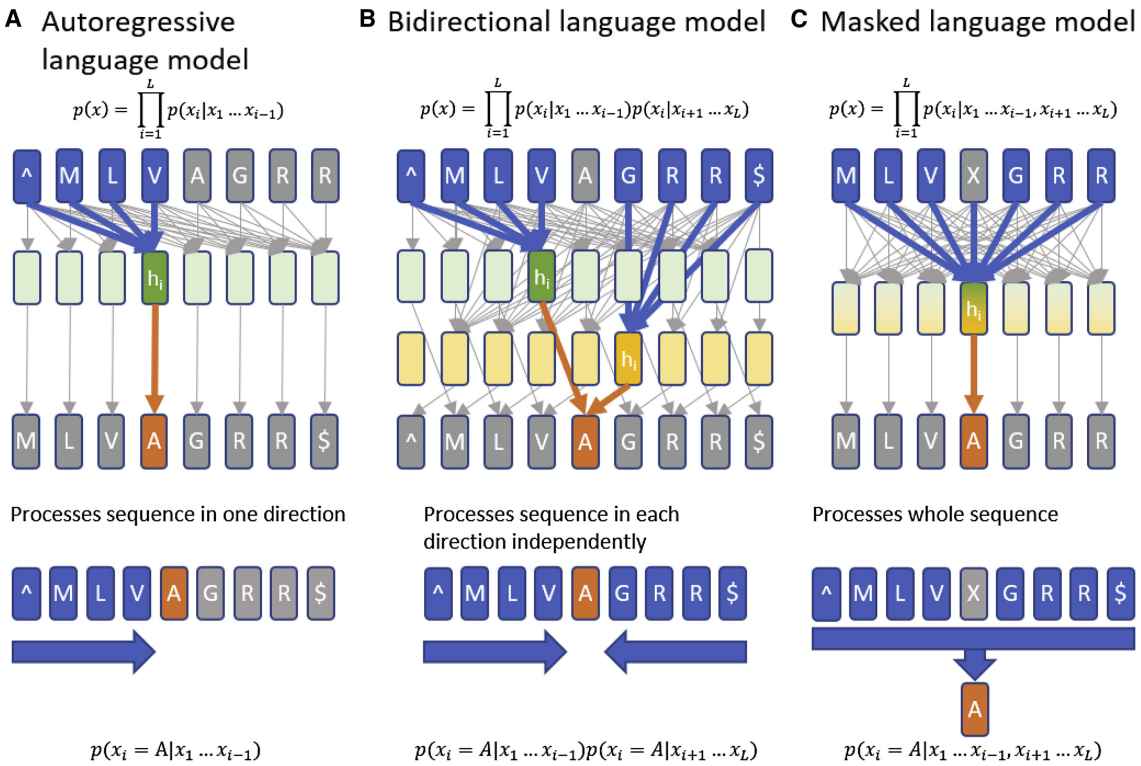

Figure 2.6: **Language-model training paradigms relevant to PLMs.** (A) *Autoregressive* models factorize $p(x) = \prod_{i=1}^{L} p(x_i \mid x_{<i})$; each hidden state h_i depends only on past tokens (start "^" and stop "$" tokens shown). (B) *Bidirectional* models compute independent forward and reverse contexts, estimating the token distribution conditioned on both sides to capture full-sequence context. (C) *Masked language models* replace tokens with a mask ("X") and predict them using all remaining positions, yielding representations well suited for transfer learning.

ESM-2: Bidirectional MLM and Structure-Aware Representations. ESM-2 is an encoder-only Transformer trained with a masked-language modeling (MLM) objective (Fig. 2.6C):

$$\mathcal{L}_{\text{MLM}}(s) = -\sum_{t \in M} \log p_\theta(a_t \mid s'),$$

where M is a set of masked positions and s' is the corrupted sequence. Trained on large, non-redundant protein corpora (e.g., UniRef clusters, UR50), ESM-2 scales up to 15B parameters and supports atomic-resolution structure inference directly from single sequences via ESM-Fold [211, 303]. Medium-sized ESM-2 models (∼650M parameters) also perform robustly, achieving near-par performance with significantly fewer resources [371].

ProtT5: Span-Denoising Seq2Seq for Protein Reconstruction. ProtT5 adapts the encoder–decoder T5 architecture to proteins using a *span corruption* training process closely related to masked objectives but formulated as sequence-to-sequence denoising (cf. Fig. 2.6C). A corrupted input x is constructed by masking multiple contiguous

residues, which the decoder then reconstructs as output spans y. The training goal is:

$$\mathcal{L}_{\text{T5}}(s) = -\log p_\theta(y \mid x).$$

Pretraining uses massive datasets including the BFD (with ∼2.1B metagenomic sequences) and UniRef50, totaling hundreds of billions of tokens [93, 137]. ProtT5 embeddings are highly effective for downstream tasks such as localization and function prediction, often outperforming older PLMs like ESM-1b and ProGen2 in binding- and function-related benchmarks [137, 146].

PLMs in Action: Comparing ESM-2 and ProtT5. While both models leverage attention to contextualize amino-acid positions—embodying the same similarity-driven reweighting mechanism—their architectures and learning objectives differ. ESM-2's bidirectional MLM focuses on reconstructing masked residues, encouraging deep bidirectional context (Fig. 2.6C). In contrast, ProtT5's span-denoising seq2seq formulation allows generation of multi-residue sequences conditioned on context. Both approaches yield embeddings that generalize well across tasks like function prediction and fitness estimation [316].

ESM-3: Towards Multimodal Protein Design. ESM-3 represents the next frontier in PLMs: a unified model that integrates sequence, structural, and functional modeling in a generative framework. It aims not only at representation but at active design (e.g., generating novel functional proteins like fluorescent variants "esmGFP") [133]. This positions ESM-3 as a multimodal, design-oriented successor to purely sequence-based models.

2.2.4 TabPfn

The Tabular Prior-data Fitted Network (TabPFN) [144] is a transformer-based foundation model for small- to medium-sized tabular datasets (up to 10,000 samples and 500 features) that applies *in-context learning* (ICL) to tabular prediction tasks in a single forward pass. Instead of being trained on a specific dataset, TabPFN is pre-trained once on millions of synthetically generated datasets derived from structural causal models designed to capture the heterogeneity of real-world tabular data, including categorical variables, missing values, uninformative features, and outliers.

The architecture treats each table cell as a separate token and alternates *feature-wise* and *sample-wise* attention, ensuring invariance to row and column permutations and enabling scalability beyond training sizes. During inference, the model jointly receives labeled training samples and unlabeled test samples, producing predictions without gradient updates.

Formally, TabPFN approximates the Bayesian posterior predictive distribution

$$\hat{p}(y_{\text{test}} \mid X_{\text{test}}, X_{\text{train}}, y_{\text{train}})$$

for the synthetic-data prior defined during pre-training, effectively encoding a learned algorithm that generalizes to new datasets. This approach consistently outperforms strong tree-based baselines (e.g., CatBoost, XGBoost) under strict time constraints, while also supporting uncertainty estimation, density modeling, data generation, and fine-tuning for related tasks.

2.3 XAI techniques

Explainable Artificial Intelligence (XAI) is a research field aimed at making decision-making processes of AI systems transparent, interpretable, and accountable. Addresses the opacity of modern high-performing models, especially deep neural networks, by providing human-understandable explanations of how predictions are generated, what information is used, and what actions are taken. This capability is essential for validating knowledge, challenging assumptions, and enabling human oversight.

2.3.1 Feature importance

Feature importance is a fundamental concept in XAI, quantifying the degree to which each input feature influences the target variable or the model's predicted outcome. It serves as an indispensable tool for understanding which features an AI model prioritizes and relies upon for its predictions. Beyond mere interpretability, feature importance can also serve practical purposes, such as identifying potential issues within the dataset (e.g., data leakage) or guiding further feature engineering efforts to improve model performance.

2.3.2 Local vs. Global Explainability

XAI methods are often classified by the *scope* of the explanation: *global* methods characterize model behaviour over the entire input domain, while *local* methods explain an individual prediction [124].

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a predictive model with input vector $x = (x_1, \ldots, x_d)$. A *local explanation* aims to approximate the prediction $f(x)$ as:

$$f(x) \approx \phi_0 + \sum_{j=1}^{d} \phi_j(x),$$

where $\phi_j(x)$ quantifies the contribution of feature j to the specific prediction for x [222, 289]. Examples include LIME and SHAP, where $\phi_j(x)$ are derived from perturbation sampling or Shapley values.

In contrast, a *global explanation* estimates the average influence of each feature over the data distribution $\mathcal{D}$:

$$G_j = \mathbb{E}_{x\sim\mathcal{D}}[\,\phi_j(x)\,],$$

or through functionals such as partial dependence functions:

$$\mathrm{PDP}_j(z) = \mathbb{E}_{x_{\setminus j}\sim\mathcal{D}}[\,f(z, x_{\setminus j})\,],$$

and their unbiased alternative, accumulated local effects (ALE) [13, 105]. Permutation-based measures also assess the change in model error upon shuffling feature j [101].

The two perspectives are *complementary*. Local attributions support case-level auditing and debugging, while global summaries aid bias detection, model monitoring, and hypothesis generation. Moreover, faithful local explanations can be aggregated to recover consistent global importance measures—demonstrated for tree ensembles in [227]:

$$G_j \approx \frac{1}{n} \sum_{i=1}^{n} \phi_j(x^{(i)}).$$

In high-stakes applications, one should also weigh the option of inherently interpretable models over post hoc explainability [300].

2.3.3 Permutation Importance

Permutation importance quantifies the dependence of model performance on a given feature j by measuring the drop in predictive accuracy when its values are randomly permuted, breaking the association between x_j and the target [101]. Formally, for a loss function L and test set $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$:

$$\mathrm{PI}_j = \frac{1}{n}\sum_{i=1}^{n} L\big(y^{(i)}, f(x^{(i)})\big) \; - \; \frac{1}{n}\sum_{i=1}^{n} L\Big(y^{(i)}, f(\tilde{x}_j^{(i)})\Big)\,,$$

where $\tilde{x}_j^{(i)}$ is $x^{(i)}$ with feature j permuted across samples.

2.3.4 SHAP Importance with Specialized Explainers

SHAP (SHapley Additive exPlanations) attributes feature contributions $\phi_j(x)$ to predictions using Shapley values, grounded in a unified additive framework [222]. A global SHAP feature importance is often defined as:

$$\mathrm{SHAP}_j = \mathbb{E}_{x\sim\mathcal{D}}[\,|\phi_j(x)|\,]\,.$$

Two widely used SHAP explainer implementations include:

- *KernelExplainer*: a model-agnostic explainer based on a weighted linear regression over perturbed samples, offering general applicability at the cost of high computation and sensitivity to background data choice [222].
- *TreeExplainer*: a specialized algorithm for tree-based models (e.g., XGBoost, LightGBM) that computes exact Shapley values in polynomial time by exploiting the model's structure. It is deterministic, highly efficient, and supports aggregation of local explanations into global summaries [224, 226].

2.4 Classic Bionformatics methods

2.4.1 Over-representation analysis (ORA)

Over-representation analysis asks whether a predefined category (e.g., a Gene Ontology term or pathway) appears among an "interesting" gene list more often than expected by chance, given a chosen background universe. Let U be the universe of $N = |U|$ genes, $A \subseteq U$ an annotated set with $K = |A|$, $L \subseteq U$ the input list with $n = |L|$, and $x = |A \cap L|$ the observed overlap. Under the null of random sampling without replacement,

$$X = |A \cap L| \sim \mathrm{Hypergeometric}(N, K, n), \qquad \Pr(X = x) = \frac{\binom{K}{x}\binom{N-K}{n-x}}{\binom{N}{n}}.$$

A one-sided enrichment p-value is

$$p_{\text{ORA}} = \sum_{i=x}^{\min(K,n)} \Pr(X = i),$$

and for depletion one sums $i \leq x$. This test is algebraically equivalent to a one-sided Fisher's exact test on the 2×2 table induced by A and L [291].

Because ORA tests many categories, false discovery rate (FDR) control is standard. With m hypotheses and ordered p-values $p_{(1)} \leq \cdots \leq p_{(m)}$, the Benjamini–Hochberg (BH) rule declares the largest k with $p_{(k)} \leq \alpha\, k/m$ significant; equivalently one uses adjusted q-values $\tilde{p}_{(i)} = \min_{j \geq i} \frac{m}{j} p_{(j)}$ [31].

Competitive null and modeling choices. ORA is a *competitive* test: it asks whether genes in A are more frequent in L than genes outside A, relative to the universe U [184]. Results therefore depend on U (e.g., all measured genes vs. all genes) and on the annotation database/version. Gene sets are not independent (e.g., GO's DAG); topology-aware procedures (e.g., `elim`/`weight` in `topGO`) reduce local dependencies and redundancy [11]. In RNA-seq, differential detection favors long/highly expressed genes; Wallenius' noncentral hypergeometric models as implemented in `GOseq` adjust for such bias [401].

Limitations and alternatives. ORA binarizes evidence (in/out of L) and ignores gene–gene correlation within sets, which can reduce power and inflate false positives. Rank-based, self-contained methods such as GSEA operate on a genome-wide ranked statistic and test whether signals are systematically shifted within the set rather than over-represented in a truncated list [184, 340].

Reporting. A transparent ORA report should state (N, K, n, x) and the exact test (hypergeometric/Fisher), the multiple-testing procedure (e.g., BH-FDR at level α), the universe definition, database and version, and any adjustments for category topology or selection bias.

2.4.2 Morgan/ECFP molecular fingerprints

Morgan (circular) fingerprints [243, 294] encode a molecule as a fixed-length binary or count vector summarizing circular substructures up to radius r (e.g., ECFP_4 uses $r{=}2$). Let the molecule be a hydrogen-suppressed graph $G = (V, E)$. Atom invariants $x_i^{(0)}$ (atomic number, valence, ring flags, charge) are iteratively updated by hashing each atom with its neighbors and bond types,

$$c_i^{(t)} = \mathrm{H}\Big(x_i^{(t)},\, \text{multiset}\{(x_j^{(t)}, b_{ij}) : (i,j) \in E\}\Big), \quad x_i^{(t+1)} {:=} c_i^{(t)}, \quad t = 0, \ldots, r,$$

and the identifiers $\{c_i^{(t)}\}$ gathered over $t \leq r$ are folded via a secondary hash into a length-d vector. Similarity is typically the Tanimoto coefficient [356]:

$$T(\mathbf{a}, \mathbf{b}) = \frac{\sum_j (a_j \wedge b_j)}{\sum_j a_j + \sum_j b_j - \sum_j (a_j \wedge b_j)} \quad \text{(binary)}$$

$$T(\mathbf{x},\mathbf{y}) = \frac{\mathbf{x}^\top \mathbf{y}}{\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 - \mathbf{x}^\top \mathbf{y}} \quad \text{(counts)}$$

Morgan/ECFP features are strong, interpretable baselines for ligand-based virtual screening and QSAR, and are widely used as inputs to classical and modern ML models.

2.4.3 Uniform Manifold Approximation and Projection (UMAP)

Uniform Manifold Approximation and Projection (UMAP) is a non-linear dimensionality reduction method grounded in a manifold-and-topology view of data [235]. It assumes observations $x_i \in \mathbb{R}^D$ are sampled from a locally connected Riemannian manifold M. UMAP first constructs a *fuzzy* neighborhood graph encoding local geometry, then finds a low-dimensional embedding $\{y_i \in \mathbb{R}^d\}$ whose fuzzy graph best matches the high-dimensional one under a cross-entropy objective.

High-dimensional fuzzy graph. For each point x_i, let ρ_i be the distance to its nearest neighbor and $\sigma_i > 0$ a local scale chosen so that the (soft) k-nearest neighbor set has approximately constant information content:

$$\sum_{j \in \mathcal{N}_k(i)} \exp\left(-\frac{d(x_i,x_j)-\rho_i}{\sigma_i}\right) = \log_2 k.$$

Directed membership strengths are defined as:

$$\mu_{i|j} = \begin{cases} \exp\left(-\dfrac{d(x_i,x_j)-\rho_i}{\sigma_i}\right), & d(x_i,x_j) > \rho_i, \\ 1, & \text{otherwise}, \end{cases}$$

and symmetrized by a fuzzy union:

$$w_{ij} = \mu_{i|j} + \mu_{j|i} - \mu_{i|j}\mu_{j|i}.$$

This construction formalizes the *fuzzy simplicial set* that UMAP seeks to preserve. In practice, the k-NN graph is built efficiently via NN-descent [**?**], enabling scalability to large datasets.

Low-dimensional graph and optimization. In the embedding space, edge probabilities are modeled by a heavy-tailed kernel:

$$p_{ij} = \left(1 + a\|y_i - y_j\|^{2b}\right)^{-1},$$

where $a, b > 0$ are fitted from the `min_dist` hyperparameter, controlling cluster compactness. UMAP minimizes the fuzzy-set cross-entropy between high- and low-dimensional graphs:

$$\mathcal{L}(Y) = -\sum_{i<j} \left[w_{ij} \log p_{ij} + (1-w_{ij}) \log(1-p_{ij})\right],$$

using stochastic optimization with negative sampling; spectral (Laplacian-eigenmaps) initialization improves global structure. Hyperparameters such as $n_{\text{neighbors}}$, `min_dist`, and the metric control the trade-off between local and global structure preservation.

Applications and comparison. Compared with t-SNE [231], UMAP's cross-entropy formulation, heavier-tailed kernel, and graph-based initialization often preserve more global relationships while maintaining competitive local fidelity. UMAP is widely adopted in single-cell transcriptomics, where it offers fast runtimes and stable neighborhood preservation [29].

Chapter 3

Background on Drug-Target Binding methods

3.1 The funnel problem

High-throughput screening bottleneck

Determining whether a chemical compound binds to a given biological target is a core challenge in drug discovery. The search space of possible drug–target pairs is astronomically large, and conventional trial-and-error screening approaches remain time-consuming and expensive [206, 350]. Computational models for DTI prediction can dramatically reduce experimental burden by prioritizing promising candidates. In addition to lowering development costs, accurate predictions of binding affinity or activity enable virtual screening campaigns and drug repurposing efforts [206, 350].

Vast yet sparse chemical space

The difficulty of the DTI problem stems first from the sheer size of the *chemical space* of drug-like molecules, which has been estimated to range from 10^{33} to 10^{60} possible small molecules depending on structural and physicochemical constraints [278]. Even commercially available libraries, such as ZINC20 and ZINC-22 [155, 354], have recently expanded to contain billions of compounds. As a result, the initial screening pool is extremely large, and we can only experimentally validate a tiny fraction of these candidates. Therefore, any model developed must be fast, scalable and able to generalize well beyond the observed chemical types. [288, 304].

Data sparsity & bias

A second major challenge is the extreme sparsity and bias of experimental ligand–target measurements. For instance, a chemogenomic interaction matrix derived from ChEMBL (probably the biggest resource around) contained data for fewer than 0.05% of all possible drug–target combinations [154], heavily skewed toward a subset of targets and assay types [53]. This sparsity is exacerbated by assay noise, batch effects, and systematic artefacts, all of which risk misleading computational models if not carefully addressed [21, 102, 103]. Effective funneling therefore requires rigorous filtering of both chemical libraries and data quality.

Activity cliffs

Finally, small structural modifications to a molecule can lead to dramatic potency shifts, a phenomenon known as an *activity cliff* [337, 338]. These cliffs represent sharp discontinuities in structure–activity relationships (SAR), where highly similar compounds exhibit large differences in biological activity. From a modeling perspective, activity cliffs are particularly challenging because they violate the smoothness assumption underlying many machine learning approaches, which generally presume that structural

similarity implies activity similarity [338]. Figure 3.1 illustrates an example where minor substitutions in ligand scaffolds lead to over a hundred-fold variation in binding affinity, textitasizing the sensitivity of molecular recognition to subtle chemical changes.

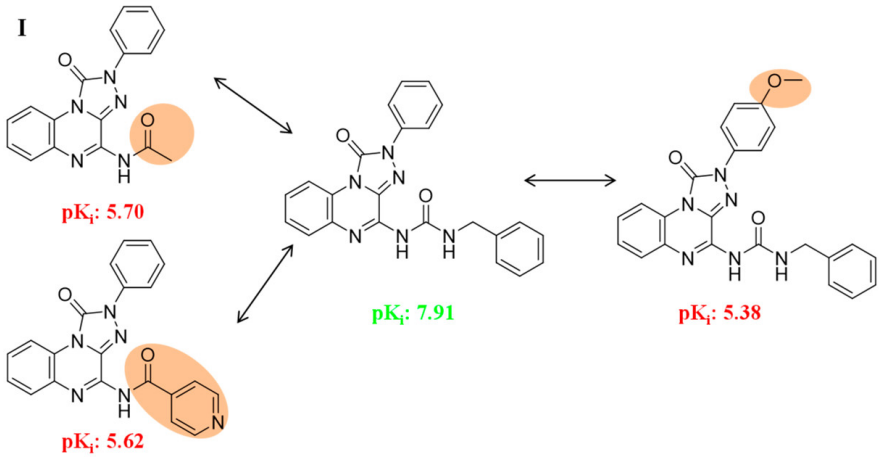

Figure 3.1: **Example of an activity cliff in drug–target interactions.** Small structural modifications (highlighted in orange) in these closely related ligands result in large differences in binding affinity, as reflected by changes in pK_i values (7.91 vs. 5.38–5.70). Such dramatic potency shifts challenge the smooth structure–activity relationship assumption often exploited by machine learning models, making activity cliffs a key obstacle in predictive modeling. Adapted from Stumpfe et al. [338].

In summary, the funnel problem reflects an adverse combination of (i) an astronomically large, sparsely sampled search space; (ii) noisy and biased measurements; and (iii) complex, context-dependent molecular binding phenomena. State-of-the-art DTI strategies aim to progressively narrow this funnel by integrating intelligent library design, robust data curation, and modeling approaches that incorporate uncertainty and biophysical realism [304].

3.2 Large-scale datasets for drug-target interactions

Primary bioactivity repositories

Progress in DTI prediction has been enabled by large, curated pharmacological databases that catalog measured interactions between small molecules and protein targets. Chief among these are *BindingDB* [215, 216] and *ChEMBL* [112, 405], which serve as primary sources of quantitative bioactivity data (K_i, K_d, IC_{50}, and related assay readouts). BindingDB focuses on experimentally determined protein–ligand affinities extracted from the literature and patents, with quality control and cross-links to external resources. Only a minority of entries have associated co-crystal structures; structure coordinates are typically obtained from the Protein Data Bank (PDB) [34]

Benchmark subsets for affinity prediction

Several widely used DTI benchmarks derive from or are linked to these repositories. The *Davis* [80] kinome panel provides K_d values for kinase–inhibitor pairs measured across a broad swath of the human kinome. The *KIBA* [347] collection integrates heterogeneous kinase bioactivities into a unified score to improve comparability across assays. These focused datasets were included in the *Therapeutics Data Commons* [148], a collec-

tion of machine learning-friendly resources that enabled many of the regression-style evaluations in sequence- and graph-based DTI modeling efforts.

Structure-resolved complexes (SBDD)

For structure-based approaches, the PDB archives biomolecular complex structures [34], and *PDBbind* assembles protein–ligand complexes with matched binding affinities into standardized training and test sets [379]. These resources support pose prediction, scoring, and affinity estimation when 3D information is available.

Decoys for virtual screening

Decoy-based benchmarks such as the *Directory of Useful Decoys, Enhanced* (DUD-E) also exist. DUD-E offers, for each target, a curated set of experimentally validated actives (sourced from ChEMBL) and approximately 50 decoys per active drawn from ZINC. These decoys are chosen to match key physicochemical properties while being topologically dissimilar to reduce the chance that decoys are latent actives. Moreover, ligands are clustered by Bemis–Murcko scaffolds to minimize chemotype bias. This setup enables more realistic enrichment testing in virtual screening workflows by requiring methods to distinguish true binders from property-matched but structurally distinct decoys.

Knowledgebases and networks

Other public resources aggregate domain knowledge at different abstraction levels. *DrugBank* [188, 386] consolidates approved and investigational drugs with their mechanisms, targets, and interactions. *STITCH* integrates evidence for protein–chemical associations into networks with confidence scoring and tissue context [195, 345]. These knowledgebases complement primary activity tables by providing curated annotations that can regularize learning and facilitate downstream interpretation.

Target development levels (IDG/Pharos)

Another resource is *Pharos* [183], the public portal to the *Target Central Resource Database* (TCRD) developed by NIH's *Illuminating the Druggable Genome* (IDG) program. It aggregates and harmonizes evidence from a broad set of sources (including DrugCentral, ChEMBL, BindingDB, UniProt, expression atlases, phenotype and disease resources) to provide a target-centric view of the human proteome and to prioritize understudied proteins [183, 251, 325]. A central construct is the *Target Development Level* (TDL), which stratifies proteins by available clinical and chemical knowledge (*Tclin*, *Tchem*, *Tbio*, *Tdark*). For DTI studies, Pharos offers practical utilities: quantifying target coverage bias across families (kinases, GPCRs, ion channels), selecting target strata for fair evaluation, and linking targets to high-quality ligand and disease annotations via programmatic access and bulk downloads [183, 325].

Next-generation PLI benchmark

Recently, *PLINDER* [89] introduced a large, richly annotated protein–ligand interactions resource designed to reduce information leakage, provide principled train–test splits across sequence, structure, pocket, and ligand similarity, and support rigorous evaluation of docking and co-folding methods. While still a preprint, PLINDER is rapidly becoming a useful companion to PDB/PDBbind for assessing modern structure-aware models.

Proteome-wide fragment screening

On a different note, Offensperger et al. [258] report a systematic, proteome-wide screen of 407 structurally diverse small-molecule fragments in HEK293T cells using a chemoproteomics strategy. This resource maps over 47,600 fragment–protein interactions across 2,667 proteins and notably found that approximately 86% of the interacting proteins previously lacked any annotated ligands. These fragment hits provide valuable starting points for medicinal chemistry, enabling the rational design of higher-affinity compounds based on weak-binding cores. Crucially, this resource offers extensive coverage of proteins (rather than just a few selected targets) and allows for the analysis of targetability across protein families, as well as insights into fragment promiscuity and

selectivity landscapes.

3.3 Molecular Representations and Features

3.3.1 Ligand features

Molecular featurization

Representing molecules in a drug–target prediction task can be achieved through diverse strategies, each balancing *interpretability*, *computational efficiency*, and *representation power*.

Fingerprints

A widely adopted approach is the use of circular fingerprints such as *Morgan* or *ECFP*, derived directly from SMILES strings. These yield sparse binary vectors that are extremely fast to compute and interpretable at the substructure level, although they neglect three-dimensional information and may suffer from bit collisions [294]. More details on fingerprints are given in chapter 2.4.2.

Physicochemical descriptors

Alternatively, low-dimensional *physicochemical descriptors* [212] (e.g. molecular weight, logP) can be computed efficiently using RDKit [32] or `datamol` [78]. They provide strong intuition about ADME-related properties and are easy to scale across datasets, but their expressive power is limited and correlations among descriptors can reduce performance.

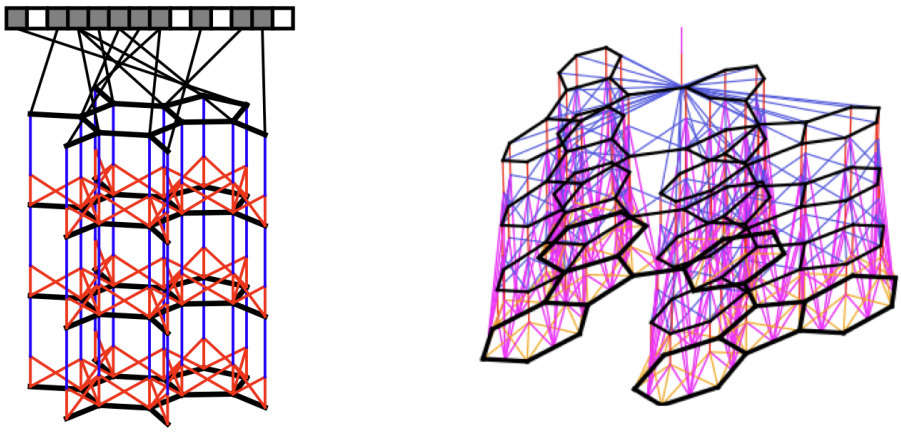

Figure 3.2: **Graph-based molecular fingerprints.** Left: schematic of neural graph fingerprints, where atoms (nodes) and bonds (edges) iteratively exchange information to construct learned representations. Right: detailed view highlighting how bond information is incorporated into the message-passing process. From Duvenaud et al. [91].

Graph neural networks

A complementary line of work represents each molecule as a *graph* (atoms as nodes, bonds as edges) and learns features *end-to-end* with *graph neural networks*. Early graph convolutions directly generalize circular fingerprints, enabling task-specific feature learning on molecular graphs [91] (see figure 3.2). Message Passing Neural Networks (MPNNs) formalize this idea by iteratively exchanging information along bonds to build permutation-invariant graph embeddings with strong expressivity for structure–property relationships [117]. In practice, GNNs capture substructure context and long-range dependencies better than fixed fingerprints, but they are *data-hungry* and are often trained from scratch for each endpoint, which raises computational cost and overfitting risks on

small datasets. Recent *pretraining* strategies (node/edge/graph-level self-supervision) and large-scale molecular pretraining (e.g., GROVER [296]) mitigate data requirements and improve transfer, though at the expense of substantial pretraining compute and added complexity [147].

Fragment autoencoders

More sophisticated representations can be obtained using autoencoders, which embed molecules into latent spaces that capture both topological and physicochemical properties. Offensperger et al. [258] introduce an autoencoder that takes as input the ECFP fingerprint of a chemical fragment, projects it into a learned low-dimensional latent space, and jointly predicts a combination of structural properties—represented as a graph embedding—and physicochemical descriptors. (see figure 3.3)

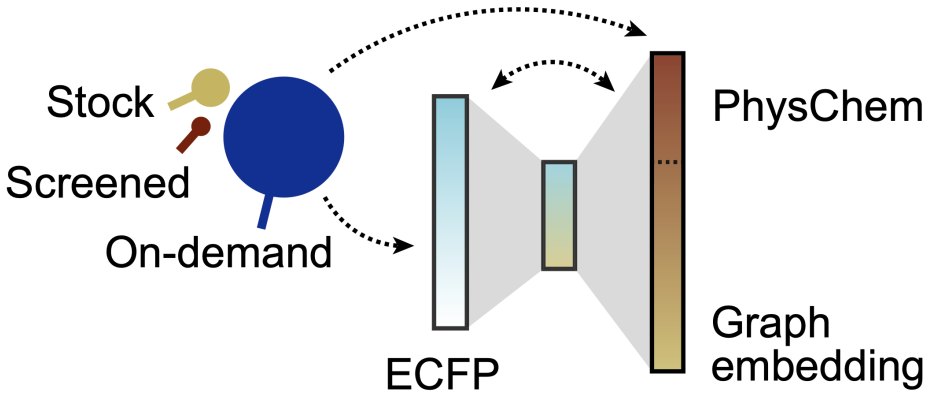

Figure 3.3: **Fragment-based autoencoder architecture for molecular representation.** The model takes molecular fragments (e.g., from stock, screened, or on-demand libraries) as input and encodes them into extended connectivity fingerprints (ECFP). These are compressed into a latent representation, from which the autoencoder jointly reconstructs two complementary views: physicochemical descriptors (*PhysChem*) and graph-based embeddings. This strategy enables the latent space to capture both topological and physicochemical properties of molecules, providing richer features for downstream drug–target prediction tasks. Adapted from Offensperger et al. [258].

SMILES language models

Recent advances in *SMILES language models*, including ChemBERTa [64] and ChemGPT [104], leverage transformer-based architectures to learn contextual embeddings of molecular strings. Implementations are available in the `molfeat` library [79], making them broadly accessible. Such models capture subtle structural regularities and generalize well across tasks, but they are computationally heavier and less transparent than handcrafted features. This models leverage the transformer architecture introduced in chapter 2.2.1.

Foundation models

Finally, *large foundation models* such as MolGPS push representation learning to scale, achieving strong general-purpose molecular embeddings and few-shot transfer capabilities [344]. Their main drawbacks are the heavy pretraining costs and limited interpretability, though they represent the current frontier of molecular representation learning.

3.3.2 Target features

Protein featurization

As with ligands, proteins must be represented in a form suitable for predictive modeling. A broad spectrum of approaches exists, each balancing *biological interpretability*, *computational efficiency*, and *ability to capture structural and functional complexity*.

Sequence descriptors

Simple *sequence-derived* descriptors summarize amino-acid composition, k-mers, or physicochemical properties (e.g. charge, hydropathy) using curated resources such as *AAindex* [182]. Turn-key toolkits like `iFeature` and `iFeatureOmega` compute hundreds of such features with built-in pipelines for normalization and selection [61, 62]. These descriptors are lightweight and interpretable, but often redundant and unable to capture higher-order context.

Evolutionary profiles

Evolutionary information provides a long-standing baseline. *PSSMs* from PSI-BLAST [12] and *HMM* profiles from HHblits [286] encode conservation and co-variation signals that correlate with structure and function. Such profiles are highly informative, though their computation requires large database searches, making them costly for large-scale screening.

Protein LMs

Transformer-based *protein language models* (pLMs) trained on massive sequence corpora generate contextual embeddings with strong transfer across tasks: ESM-1b [210], ESM-2 [210], and ProtTrans/ProtT5 [92]. These embeddings frequently outperform handcrafted features, but at the expense of higher computational demands and reduced transparency. Their design leverages principles introduced in chapter 2.2.1.

Structure features

Three-dimensional structures enable rich residue- and site-level features. AlphaFold2 achieves near-experimental accuracy [173], and AlphaFold DB offers large-scale coverage for downstream feature extraction [365, 366]. From experimental or predicted structures, one can derive secondary structure, solvent accessibility, disorder, and confidence metrics (e.g. pLDDT). When structure is unavailable, sequence-based proxies such as PSIPRED and IUPred remain standard [168, 239]. Binding-site features can be obtained using geometry-based pocket detectors (fpocket [199], P2Rank [191]) and interaction profilers (PLIP), with BioLiP2 offering curated complexes for benchmarking [7, 308, 408].

Domains & ontology

Domain and family annotations (Pfam) and *functional ontologies* (GO) offer interpretable, expert-curated descriptors that provide valuable priors for mechanism-aware modeling [99, 352].

Geometric GNNs

When structures are known, *equivariant* GNNs can directly learn from 3D protein graphs defined by backbone and side chains. SE(3)-equivariant attention and geometric vector perceptrons are popular architectures [106, 167]. These models effectively capture spatial constraints and support residue-level or site-specific predictions, but require heavy regularization and substantial compute.

Frontier models

Foundation-style *generative* pLMs are beginning to unify sequence, structure, and function. ESM-3 [133] exemplifies this trend, demonstrating controlled functional protein generation. While powerful, such models demand extensive pretraining resources and raise questions of interpretability and scalability.

3.4 Quantifying Binding and Defining Interaction

Binding affinity metrics

Experimental assays produce quantitative measures of how strongly a compound interacts with a target. Common metrics include the inhibitory constant (K_i) or dissociation constant (K_d), typically reported in molar units, which reflect the concentration at which the drug binds to or inhibits the target by 50%. The IC is a related concept: the concentra-

tion of compound that achieves 50% inhibition of enzyme activity or cellular growth in an assay. Lower K_d/IC values indicate stronger binding (higher affinity) [216]. Because these values span many orders of magnitude, it is routine to take negative logarithms and use the pK or pIC scale (e.g., a K_d of 1 μM corresponds to pK = 6).

Censored and thresholded data

Not every data point comes as a precise numeric affinity. High-throughput screens often yield qualitative results. For instance, a compound might be reported as "active" at the highest screening concentration or "inactive" if no effect was seen up to that limit. In BindingDB [216], many entries include exact K_d or IC values, but others might only indicate "greater than X μM" (meaning affinity is weaker than the tested range). Such censored data require careful handling. When constructing binary classification datasets from continuous data, researchers must choose a threshold to dichotomize the outcomes. A common practice is to label a compound as active if, for example, $K_i < 1–10$ μM (potent binder), and inactive if $K_i > 10–100$ μM. The choice of threshold (and how to treat intermediate cases) can strongly affect the balance and bias of the training data. For example, the Davis dataset capped all undetectable activities at 10 μM (pK = 5) by default, and some models train on a filtered subset excluding those ambiguous points.

Binary vs continuous prediction

These considerations directly influence how DTI prediction is formulated. Models can either treat the problem as a regression task, aiming to predict continuous affinity values (e.g., K_d, IC), or as a classification task, predicting whether a compound is "active" below a certain concentration threshold [206]. The classification approach presents the additional challenge of selecting an appropriate cutoff, which can be somewhat arbitrary and context-dependent. For this work, we adopt the thresholds defined by the Pharos platform [324]: for Kinases, 30 nM; for GPCRs, 100 nM; for Nuclear Receptors, 100 nM; for Ion Channels, 10 M; and for other categories, 1 M. While convenient, such binarization inevitably discards information and can introduce biases. Future evaluations may benefit from more nuanced strategies, such as the probabilistic thresholding approach introduced by Mervin et al. [238], which maps activities into continuous values between 0 and 1 using a Gaussian cumulative distribution function.

Activity vs inactivity labels

It is also crucial to consider what constitutes a "negative" interaction. Often, an unmeasured drug–target pair is (implicitly) treated as non-interacting, but this assumption is fragile and introduces label noise; positive–unlabeled strategies have therefore been proposed for DTI to avoid contaminating the negative class with unknowns [273]. Some benchmark construction efforts explicitly gather *confirmed inactive* examples by retaining only measurements shown inactive at high concentration (e.g., $\geq$20 μM) and by enriching underrepresented targets with additional inactives to achieve class balance [290]. At the same time, literature-driven resources such as BindingDB tend to contain an unusually high fraction of positives, since compounds are often tested in analogue series around an initial hit and successful actives are preferentially reported, a well-documented *analogue bias* [373]. Synthetic negatives can further regularize training: *network-derived negatives* select protein–ligand pairs at large shortest-path distance in the interaction graph (e.g., $\geq$7 hops), which correlates with very weak binding, and combine these with absolute non-binders from BindingDB to balance positives and negatives per node [53]. Alternatively, *sphere-exclusion* in chemical space oversamples presumed inactives far from known actives, reducing false negatives while covering a broad chemistry [237]. In summary, whether using regression or classification, one must be mindful of how ground truth is defined.

3.5 Spectrum of Computational Methods

A diverse range of computational strategies has been developed for DTI prediction, extending from physics-based simulations grounded in structural biology to fully data-driven machine learning paradigms. This broad *spectrum of methods* illustrates how the field seeks to reconcile different priorities: on one side, detailed mechanistic interpretability and physical realism, and on the other, computational scalability and the capacity to generalize across the vast chemical and biological spaces relevant to drug discovery. Together, these complementary approaches highlight the multifaceted nature of DTI modeling and the need to balance accuracy with practicality in real-world applications.

Physics-based foundations

At the mechanistic end of the spectrum lie *physics-based approaches*. Classical molecular docking uses protein and ligand 3D structures to propose binding poses and assign empirical scores. Docking can screen ultra-large libraries at scale [229], yet its scoring functions often produce false positives, limiting precision. More rigorous alchemical free-energy methods such as free energy perturbation (FEP) or MM/GBSA explicitly model conformational ensembles and solvent effects, delivering accurate relative binding free energies [378]. However, these methods remain too computationally expensive for high-throughput exploration and are therefore mainly applied at the lead-optimization stage.

Classical machine learning

Moving toward data-driven methods, *classical machine learning* models exploit similarity patterns between drugs and targets. For instance, *SimBoost* [136] uses gradient boosting on engineered features derived from drug/protein similarities and network topology, providing a strong and efficient baseline for affinity prediction without requiring deep architectures. Such methods are interpretable and sample-efficient but rely heavily on careful feature design.

Kernelized matrix methods

Another influential line of work frames DTI prediction as a *link-prediction problem* on the bipartite drug-target graph. *KronRLS* [246], based on kernelized regularized least squares, leverages Kronecker product kernels to combine drug and target similarities, enabling information sharing across related proteins and compounds [266]. These approaches are theoretically elegant and effective when similarity information is available, but they face difficulties in *cold-start* scenarios where no prior data exist for a new drug or target.

Sequence-based deep learning

The advent of deep learning introduced *end-to-end sequence-based models*. *DeepDTA* [264] demonstrated that convolutional neural networks can directly process SMILES strings and protein amino-acid sequences, achieving superior performance compared with classical methods. These approaches minimize manual feature engineering and generalize broadly to novel molecules and proteins. Nevertheless, their reliance on sequence alone neglects explicit structural information, limiting their ability to capture spatial complementarity in binding.

Graph neural networks

To incorporate structural priors, models increasingly employ GNNs for molecular representation. *GraphDTA* encodes ligands as atom-bond graphs and applies message passing to capture local and long-range substructure dependencies, which are then fused with protein sequence encoders to predict affinity [252]. Building on this, *AI-Bind* introduced strategies to enhance *out-of-distribution generalization*, combining network-derived synthetic negatives with large-scale pretraining to improve robustness on un-

seen targets and ligands [53].

Transformer models

The success of attention-based architectures has motivated *transformer-based DTI models*. *MolTrans* [149] extracts sub-structural motifs from drugs and contextual embeddings from proteins, applying cross-attention to learn rich interaction features. Similarly, *TransDTI* [174] leverages bidirectional transformer encoders for both molecules and proteins, improving performance on classification and regression benchmarks. Transformers excel at capturing long-range dependencies and benefit from large-scale pretraining, but they demand extensive data and computational resources.

Foundation models

At the frontier, *foundation models* aim to unify structural and sequence information in generalist architectures. The recent *Boltz-2* framework exemplifies this trend, proposing accurate and efficient binding affinity prediction while simultaneously modeling 3D complex structures [268]. Though still a preprint, Boltz-2 highlights a paradigm shift toward large, multi-modal models capable of transfer across diverse bioactivity tasks.

Balancing trade-offs

In practice, each methodological class occupies a niche. Physics-based docking and free-energy methods provide mechanistic interpretability but remain limited in throughput. Classical ML models offer speed and transparency, while deep learning architectures capture complex cross-modal relationships at the expense of interpretability and compute. Hybrid pipelines that combine docking (for pose generation), machine learning (for large-scale prioritization), and physics-based refinement (for affinity estimation) currently offer the most pragmatic balance between accuracy, scalability, and mechanistic insight.

3.6 Ongoing Challenges and Future Directions

3.6.1 Methods scalability

As discussed in Chapter 3.1, the chemical space of drug-like molecules is extraordinarily vast, making it infeasible to conduct exhaustive explorations of libraries like *ZINC* [155]. This highlights the necessity for approaches that optimize both *predictive power* and *scalability*. Recent advancements in deep learning architectures, particularly transformer-based models [149], graph neural networks [53, 252], and, more recently, foundation models such as Boltz-2 [268], have demonstrated enhanced accuracy in predicting drug–target interactions. However, these models come with high computational demands, which could create bottlenecks in ultra-large-scale screening. A promising and practical approach is the development of *tiered pipelines*, where efficient pre-screening models quickly narrow down the search space. This allows more resource-intensive architectures, like Boltz-2, to be applied selectively to the most promising candidates.

3.6.2 Data quality, bias, and generalization

Heterogeneity in assays

Despite rapid advances in machine learning, the performance of DTI models remains fundamentally constrained by the quality and structure of the underlying data. Public repositories such as *ChEMBL* [405] and *BindingDB* [216] aggregate bioactivity measurements generated under diverse experimental conditions, which introduces substantial heterogeneity. Common affinity metrics, including IC_{50}, K_i, and K_d, are often intermixed, despite their interpretation depending strongly on assay design and substrate

concentration [404]. Cross-dataset analyses reveal striking inconsistencies: even when assays nominally target the same protein, naive aggregation of IC_{50} values shows that two-thirds of data pairs differ by more than 0.3 log units, while maximal curation improves concordance (Kendall's $\tau \approx 0.71$) at the expense of coverage [175, 198]. This illustrates a persistent trade-off between dataset size and measurement reliability.

Noise floor in affinity data

Experimental variability further compounds these issues. Repeated measurements for identical ligand–target pairs deviate on average by 0.68 log units (a $\sim$4.8-fold difference in affinity), even after extensive filtering [175]. Such irreducible variability effectively defines a noise floor and sets an upper bound on achievable model accuracy, regardless of algorithmic sophistication.

Bias and leakage

Beyond measurement noise, biases in data collection and reporting introduce structural artifacts. Literature-derived datasets are enriched for positive results, since weak or null findings are less likely to be published, skewing models toward overpredicting binding affinity [53, 242]. Many benchmarks also suffer from scaffold or analogue leakage between training and test sets, which inflates performance estimates by rewarding memorization rather than genuine inference [373]. These problems have motivated countermeasures such as positive–unlabeled learning [273], explicit collection of inactive examples, and leakage-resistant evaluation strategies, including scaffold- or time-split validation [326].

Generalization to novel cases

The consequences of these biases are evident in model generalization. A central goal of DTI prediction is to extrapolate to *novel chemistry and biology*, such as compounds with unseen scaffolds or orphan targets lacking annotated ligands. Yet, models that perform well under random cross-validation often degrade substantially when evaluated under scaffold or family-based splits, exposing reliance on low-level similarity rather than robust binding rules [373]. Even under controlled splits, generalization tends to be limited to proteins evolutionarily related to the training set, while performance deteriorates for divergent or sparsely annotated targets. Richer biological representations such as evolutionary profiles, structural embeddings, or transfer learning from large pretrained models may offer a path to bridge these gaps.

Robust evaluation metrics

Finally, robust evaluation requires careful choice of metrics. Standard AUROC can be misleading in highly imbalanced datasets, where negatives vastly outnumber positives. Precision–recall analysis (AUPRC) offers a more faithful measure by directly weighting performance on the positive class, which is the primary focus in drug discovery [307]. Stratified evaluations across underrepresented targets further help distinguish whether improvements reflect genuine generalization or overfitting to well-studied proteins. Such fine-grained benchmarking is critical to ensure that algorithmic progress translates into practical discovery benefits.

Curation as a bottleneck

Taken together, these observations underscore that advances in model architecture cannot circumvent the intrinsic noise, bias, and sparsity of current datasets. Rigorous curation pipelines, assay-specific metadata harmonization, and robust evaluation protocols are indispensable prerequisites for reliable DTI prediction. Modeling efforts should therefore be viewed as inseparable from data quality, with even state-of-the-art architectures bounded by the limitations of the measurements they are trained on.

3.6.3 Interpretability and mechanistic insight

The most accurate DTI architectures increasingly rely on deep neural networks whose internal computations are opaque, yet *actionable* interpretability is essential for hypothesis generation and medicinal chemistry iteration. Beyond post hoc plausibility, useful explanations should localize which ligand substructures and protein residues drive a prediction, quantify their contributions, and remain stable under small input perturbations [165]. Established attribution techniques such as Integrated Gradients and SHAP provide principled routes to feature attribution with desirable axioms (e.g., sensitivity, implementation invariance, and additivity), and are now commonly applied to molecular graphs, sequences, and structures [223, 342]. For graph-based DTI models, methods like *GNNExplainer* highlight predictive subgraphs and edge features, offering chemically meaningful rationales (e.g., pharmacophores) when faithful [399]. However, attention heatmaps alone should not be equated with explanations, as faithfulness does not generally follow from attention weights [158]; rigorous evaluation of explanations (faithfulness, stability, human utility) therefore remains necessary [248].

Chapter 4

Background on response prediction in cancer cell lines

4.1 The Challenge of Personalised Cancer Therapy

Tumour heterogeneity and precision medicine

The central challenge in modern oncology is the profound heterogeneity of cancer [35, 77]. Tumours with similar clinical and histological phenotypes can exhibit vastly different molecular characteristics, leading to a wide spectrum of responses to a given therapy [35, 77]. This reality has fostered a paradigm shift away from "one-size-fits-all" treatments and towards the goal of personalised, or precision, medicine. The core ambition of this approach is to tailor therapeutic strategies to the unique molecular profile of an individual patient's tumour, thereby maximising efficacy while minimising toxicity [362].

Clinical data scarcity/sparsity

A key component of precision oncology is the ability to predict, in advance, how a patient's cancer will respond to a specific drug. This objective is known as Drug Response Prediction (DRP) [6]. Despite significant methodological advances, the development of robust computational models for DRP remains constrained by the limited availability of large-scale, patient-level clinical response data. Generating such datasets would ideally require the systematic testing of diverse pharmacological agents across broad and representative patient populations, yet these types of studies are ethically challenging and practically unfeasible due to cost, recruitment, and safety considerations [267, 390]. As a result, the field has predominantly relied on preclinical models as substitutes for human cancers [56].

Cell lines as proxies

Among these models, cancer cell lines have emerged as the primary workhorse for high-throughput pharmacogenomic research [70, 153]. They are relatively inexpensive and straightforward to culture, which permits large-scale, systematic screening experiments that are essential for training data-driven models. While cell lines do not fully recapitulate the complexity of *in vivo* tumours, particularly the tumour microenvironment, they harbour clinically relevant genomic alterations and provide a powerful system for interrogating the molecular underpinnings of drug sensitivity [70, 153]. The problem, therefore, is computationally framed as how to leverage the vast molecular and pharmacological data from cancer cell lines to build robust, generalisable models that can predict cellular sensitivity to anticancer compounds [56, 100, 394].

4.2 Large-Scale Pharmacogenomic Datasets

Foundational resources

The progress in DRP has been propelled by several ambitious, large-scale projects that have generated publicly available pharmacogenomic datasets. These resources pair comprehensive molecular characterisations of hundreds of cancer cell lines with their corresponding sensitivity profiles to a wide array of chemical compounds. The foundational datasets that have shaped the field include:

- *National Cancer Institute 60 (NCI-60) [327]:* A canonical panel of 60 human cancer cell lines spanning nine tissue types. NCI-60 has been screened against tens of thousands of compounds using standardised five-dose GI_{50} assays, and extensively characterised at the genomic, transcriptomic, and proteomic levels. It provided one of the earliest systematic resources for mechanism-of-action studies and benchmarking.
- *Cancer Cell Line Encyclopedia (CCLE) [25]:* A large collaborative effort led by the Broad Institute, initially profiling 947 cell lines and later expanding to include deep exome sequencing, transcriptomics, and proteomics. CCLE serves as a reference atlas of cancer cell line genomics and underpins numerous drug sensitivity analyses.
- *Cancer Therapeutics Response Portal (CTRP) [27, 320]:* Provides systematic pharmacological profiles for hundreds of small molecules across ∼860 CCLE-characterised cell lines. By linking drug response metrics (e.g. AUC, EC_{50}) to molecular features, CTRP has been instrumental in uncovering drug mechanisms and predictive biomarkers.
- *Genomics of Drug Sensitivity in Cancer (GDSC) [153]:* A landmark project from the Wellcome Sanger Institute, screening hundreds of anti-cancer agents across nearly 1,000 molecularly annotated human cancer cell lines (GDSC1/2). It provides a rich benchmark for drug response prediction and for identifying genomic markers of therapeutic sensitivity.
- *The PRISM Repurposing Dataset [70]:* A recent high-throughput resource from the Broad Institute, leveraging DNA barcoding to screen thousands of existing and experimental compounds in pooled assays across hundreds of cell lines. This approach dramatically increases throughput while preserving quantitative comparability across conditions, enabling systematic exploration of drug repurposing opportunities.

4.2.1 Available Data Modalities

Key omics features

The predictive power of DRP models is derived from the rich, multi-layered molecular data, or 'omics', provided by these screening projects. The primary data modalities that serve as input features for the models include genomics, epigenomics, transcriptomics, and proteomics. Transcriptomic data, specifically *gene expression (GE)* profiles from microarray or RNA-sequencing, are the most widely used and have consistently been shown to be the most predictive single data type [54, 73, 100, 153]. Genomic features are

also critical, including *mutational status (MUT)* from exome or whole-genome sequencing, and *copy number variations (CNV)* or *aberrations (CNA)*, which quantify the amplification or deletion of genes [25, 110, 115]. Epigenomic data, such as *DNA methylation (DM)* profiles, provide another layer of information on gene regulation [153]. Proteomic measurements (e.g., mass-spectrometry or RPPA) further capture post-transcriptional regulation where available [254]. Finally, information about the drugs themselves, such as their chemical structures represented by *molecular descriptors and fingerprints*, is often incorporated to enable predictions for novel compounds and to help models generalise across different chemical classes [56].

High dimensionality and sparsity of modalities

A key challenge in working with these datasets is their inherently *high dimensionality*: even within a single modality, the number of measured features (e.g., tens of thousands of genes) can vastly exceed the number of samples. When multiple modalities are integrated, dimensionality grows further, exacerbating the risk of overfitting and making it essential to employ strategies that *filter signal from noise*, such as feature selection, dimensionality reduction, or regularisation [49, 179]. In addition, the *incomplete and uneven coverage* of multi-omic profiles across cell lines or patient samples limits the large-scale deployment of integrative prediction models [348]. While emerging methods for *multi-view learning* can partially address this issue by leveraging overlapping subsets of data [15, 285, 374], the lack of systematic, fully matched multi-omic datasets remains a bottleneck for building robust models that fully exploit the complementary information from all molecular layers [411].

4.3 Quantifying Drug Response

Dose–response to scalar

To train and evaluate supervised machine learning models, a quantitative and continuous measure of a cell line's response to a drug is required. This value is typically derived from dose–response experiments, where cell viability is measured across a range of drug concentrations, yielding a characteristic sigmoidal curve [126]. Several metrics are used to summarise this curve into a single value representing the degree of sensitivity.

Half-maximal inhibitory concentration

The most common metric for quantifying drug sensitivity is the *IC50*, or half-maximal inhibitory concentration. This value represents the molar concentration of a drug that is required to inhibit cell growth or viability by 50% relative to untreated controls. A lower IC50 value signifies that less drug is needed to achieve a potent effect, indicating that the cell line is more sensitive to that compound. Conceptually similar is the *GI50* (half-maximal growth inhibition), which was used in the NCI-60 screen [327]. These point-estimate metrics are standard outputs for datasets like CCLE,GDSC and PRISM[26, 70, 110].

Activity area

Alternative metrics aim to capture the overall effect of the dose–response relationship. The *Area Under the Curve (AUC)* or *Activity Area (AAC)* integrates the response across the entire tested concentration range, providing a more holistic view of drug efficacy. A lower AUC value (or a higher AAC value) typically corresponds to greater sensitivity. Such integrated measures are widely used in large pharmacogenomic resources (e.g., CCLE's activity area; CTRP and GDSC AUC) [25, 153, 320]. While these integrated measures capture overall potency and efficacy, they lack the direct biological interpretability of point estimates such as IC_{50}, which can be understood in terms of a specific dose.

Log fold-change

More recently, the PRISM project has adopted reporting just the *log fold-change (LFC)* in cell abundance as a drug sensitivity readout. LFC quantifies the change in representation of a barcoded cell line in a pooled culture after drug treatment relative to a control, with more negative values indicating stronger growth-inhibitory or cytotoxic effects and thus greater sensitivity. While LFC is not a new concept (dose–response curves and IC_{50} estimates are often derived from multiple LFC measurements across concentrations), PRISM applies it at a single dose for a large number of compounds (∼6,300 drugs), enabling broad coverage at substantially lower cost. This single-dose approach is inherently noisier than full dose–response profiling, but it allows high-throughput screening at a scale not feasible with traditional IC_{50} assays, which PRISM reports for only a smaller subset of compounds after secondary multi-dose retesting [70].

4.4 Open Challenges and Future Directions

Despite significant advances in model development, several fundamental challenges persist that limit the clinical translation of DRP. These challenges represent key areas for future research and innovation.

4.4.1 Experimental Data Inconsistency

Inter-dataset inconsistency

A well-documented issue is the inconsistency in drug response measurements across the major pharmacogenomic datasets [394]. This debate gained prominence when Haibe-Kains et al. [128] demonstrated that, despite strong concordance in genomic data, drug response measurements between the CCLE and GDSC datasets were highly discordant, raising concerns about the reliability of pharmacogenomic studies for identifying gene-drug associations. A subsequent reanalysis by a consortium including members of both original teams showed that, when methodological differences such as the choice of sensitivity metric and the drug concentration range tested are carefully accounted for, substantial agreement can be recovered [1]. Nevertheless, concordance remains imperfect: on a set of shared drugs, GDSC and PRISM only reach a concordance level (in terms of Pearson correlation) of 0.60 [70]. For the same cell line-drug pair, the reported sensitivity values can differ significantly between datasets. This variability arises from differences in experimental protocols, such as the choice of cell viability assay (e.g., CellTiter-Glo versus Syto60) and data normalisation procedures. This inter-dataset noise establishes a practical upper bound on the performance that can be expected from any predictive model and severely complicates efforts to integrate data from multiple sources.

4.4.2 Clinical Translation and Model Interpretability

Translationl gaps in starting data

The ultimate goal of DRP is to inform clinical decision-making. However, a significant translational gap exists between preclinical cell line models and human patients. Cancer cell lines are grown in monolayer cultures that lack the complex three-dimensional architecture, cellular heterogeneity, and crucial interactions with the tumour microenvironment and the host immune system that profoundly influence therapeutic outcomes *in vivo* [35, 77]. While more sophisticated models like patient-derived organoids and

xenografts are being developed to bridge this gap [139, 372], they are not yet available at the scale required for training large-scale predictive models.

Computational methods to bridge translational gaps

To bridge this gap, computational models have emerged to better align cell-line data with patient tumor profiles. For example, the unsupervised alignment method Celligner [380] enables direct comparison of cell-line and tumor expression patterns by removing systematic differences, thereby improving the selection of cell-line models that more closely resemble patient tumors. Additionally, a deep learning framework has been proposed to transform data from cancer cell lines and patient-derived xenografts into a latent space shared with human tumors [84]. Despite their potential, these approaches may be challenging for non-technical users to deploy effectively. A complementary approach is provided by CELLector [245], which guides cell line selection based on genomic signatures rather than transcriptomic alignment. CELLector identifies recurrent molecular subtypes in patient cohorts and prioritizes cell lines that recapitulate these genomic contexts. The two strategies address different aspects of the translational challenge: Celligner ensures global transcriptomic comparability, capturing expression-level similarities that may reflect functional states beyond what is encoded in the genome, but assumes that batch correction adequately removes technical confounders. CELLector, by contrast, relies on discrete genomic features that are more robust to technical variation and directly actionable for genotype-driven therapies, but may miss transcriptionally defined phenotypes not captured by mutation or copy-number profiles. In practice, the two approaches could be used in combination: CELLector to select genomically appropriate models, and Celligner to verify transcriptomic concordance.

Lack of interpretability of DRP models

Furthermore, many of the most powerful and accurate DRP models, particularly those based on deep learning, function as "black boxes" [56, 100, 394]. It is often difficult to discern the biological reasoning behind their predictions. For a model to be trusted and adopted in a clinical setting, it must be interpretable. There is a pressing need for models that not only predict response accurately but also identify the key molecular features and biological pathways driving that response, thus providing actionable insights for clinicians. Balancing the trade-off between predictive power and biological interpretability remains a key challenge for the field [116, 301].

4.4.3 Axes of Distribution Shift and Why They Matter

In DRP, *out-of-distribution* (OOD) generalisation describes performance when either the compounds or the biological systems (cell lines/tumours) at test time differ systematically from those seen during training. Three primary axes are routinely studied:

Held-out drug (compound OOD) Models are asked to predict responses for *previously unseen* drugs. This scenario captures prospective use-cases such as screening novel chemicals or repurposed drugs, where neither the exact structure nor close analogues are present in training. Because random splits over-represent close analogues of test molecules, chemically aware splits are essential: *Bemis–Murcko scaffold* splits reduce analogue leakage by ensuring that test molecules do not share core scaffolds with training molecules [393]. In medicinal chemistry settings, *time-split* validation further approximates prospective performance by training on earlier-registered compounds and testing on compounds registered later in time [326]. These practices collectively

guard against over-optimistic estimates that arise when near-duplicates are inadvertently shared across folds.

Held-out cell lines (biology OOD). Here the compounds remain in-distribution, but models must generalise to *new cellular contexts* (previously unseen cell lines, tissues, or molecular subtypes). This axis approximates the clinical goal of transferring preclinical knowledge to new patients and disease contexts and asses whether the model is able to caputre underlying biological drivers rather than memorizing cell-line identities. Tissue-aware and cluster-aware splits, which prevent placing highly similar lines in training and testing, provide more accurate estimates by reducing identity leakage from nearly identical molecular profiles. Cross-study validation (e.g., train on GDSC and test on CCLE or PRISM) compounds this difficulty by adding differences in assay protocols and preprocessing; performance typically drops substantially under such shifts [394].

Cross-study validation. A particularly rigorous form of out-of-distribution evaluation involves training and testing on datasets derived from different experimental studies. For instance, this can include training on GDSC dataset and testing on the PRISM dataset, both of which consist of cell lines, or training on one of these datasets and testing on TCGA, which profiles patient tumor samples. This transition from preclinical cell line models to human tumors introduces a significant domain shift due to the added complexity of the tumor microenvironment and the variability between patients. As a result, this domain shift often leads to notable performance degradation, even for models that perform well within a single research study [394]. Such evaluations are essential for assessing the robustness of models in real-world applications, where they will inevitably encounter data generated under diverse laboratory and clinical conditions.

4.4.4 Few-shot Adaptation Across Contexts

A promising approach to address out-of-distribution gaps is few-shot learning. This method involves pretraining a model on large pharmacogenomic panels and then adapting it to a new domain using only a small number of labeled examples. Research by Ma et al. [230] demonstrated that few-shot fine-tuning can effectively bridge the transitions from cell lines to patient-derived tumor cultures (PDTC) and patient-derived xenografts (PDX), allowing for adaptation across different tissues. This process enhances the translation of results from high-throughput screening to patient-derived models [230]. This paradigm is particularly relevant in clinical contexts where data scarcity is an issue, and it complements the technique of robust splitting. While robust splitting assesses the generalization capacity of the base model, few-shot adaptation offers a systematic approach to further specialize the model for the target domain.

4.4.5 From IC_{50}/LFC to Patient Stratification

In vitro metrics like IC_{50} or log fold-change (LFC) are often treated as proxies for drug efficacy, but excessive reliance on potency alone can be deeply misleading. Compounds exhibiting uniformly low IC_{50} across a wide range of cell lines may simply be broadly cytotoxic rather than truly effective against cancer-specific signaling pathways. This

indiscriminate cytotoxicity raises significant translational red flags, since it tends to correlate with poor safety margins in vivo. To mitigate this pitfall, it is critical to complement potency with a measure of *selectivity*. The *selectivity index* (SI) operationalizes this concept as a ratio that contrasts compound activity in non-malignant versus malignant cells:

$$\mathrm{SI} = \frac{\mathrm{IC}_{50}^{\text{normal}}}{\mathrm{IC}_{50}^{\text{tumour}}}.$$

Potency vs. selectivity

An $\mathrm{SI} \geq 2$ is often considered evidence of preferential cytotoxicity toward tumour cells, which suggests a potentially safer therapeutic window [385]. The key insight is that while drug potency is essential for effectiveness, it must also be paired with selectivity to have clinical value. A low IC_{50} in cancer cells is important, but it alone cannot ensure therapeutic success. Without a tumour-specific therapeutic window, a drug's clinical utility may be limited. Therefore, potency metrics should always be assessed alongside selectivity indices for a comprehensive understanding of a drug's potential.

Limitations of SI

However, measures such as the SI require that each compound be tested in both cancer and matched normal cell lines, effectively doubling the experimental burden. This requirement is often impractical in large-scale drug screening efforts, where profiling extensive panels of normal cells is resource-intensive and not always feasible. In future work, it would be valuable to develop *distribution-based specificity indices* that leverage the statistical properties of drug response across diverse cancer cell lines alone, without the need for parallel normal cell assays. Such approaches could provide a scalable proxy for selectivity estimation, complementing potency measures and reducing experimental overhead.

Integration with PK/PD

Furthermore, pharmacokinetic and pharmacodynamic (PK/PD) considerations remain essential. A compound with favourable *in vitro* selectivity but poor absorption, rapid metabolism, or an unfavourable toxicity profile can still fail clinical translation [207]. Ultimately, the path from DRP model output to patient stratification requires integrating potency, selectivity, and exposure feasibility into a unified decision framework.

Part III

Contributions

Chapter 5

Ligand–Target Interaction Prediction

This chapter is based on the BindSight framework, an un-published and on-ongoing line of work. This framework was developed to address the challenge of predicting interactions between small molecules and protein targets from large-scale chemogenomic data.

5.1 A visual characterization of the dataset biases

Before introducing the *BindSight* framework, we devote this section to a systematic visual characterization of biases that are inherently present in DTI datasets. This analysis motivates the methodological design choices underlying our framework. The central idea is to study how different data splitting strategies influence the distribution of compound–protein pairs in the feature space, highlighting potential pitfalls for machine learning models trained under naïve assumptions.

Procedure. Molecules were embedded using *Mol2Vec* [157], while proteins were embedded using the transformer-based *ESM-2* [210]. Both embeddings were L_2-normalized to ensure comparable contribution to the joint representation. We then built compound–protein pair embeddings by concatenating the normalized vectors and projected them into two dimensions using t-SNE. Importantly, no label information (binding vs. non-binding) was used at any point in this procedure. The resulting two-dimensional projections therefore reflect purely statistical properties of the datasets. We considered one training split and three test splits, designed to mimic distinct scenarios in drug discovery:

- **Known proteins – unknown ligands (lead optimization):** new ligands are evaluated for proteins already represented in training.
- **Unknown proteins – known ligands (drug repurposing):** new targets are evaluated for ligands already represented in training.
- **Unknown proteins – unknown ligands (virtual screening):** both the protein and the ligand have never been observed in training, the most extreme OOD scenario.

Training vs. known protein–unknown ligand split. Figure 5.1 compares the training set (known protein–known ligand pairs) with the first test split (known protein–unknown

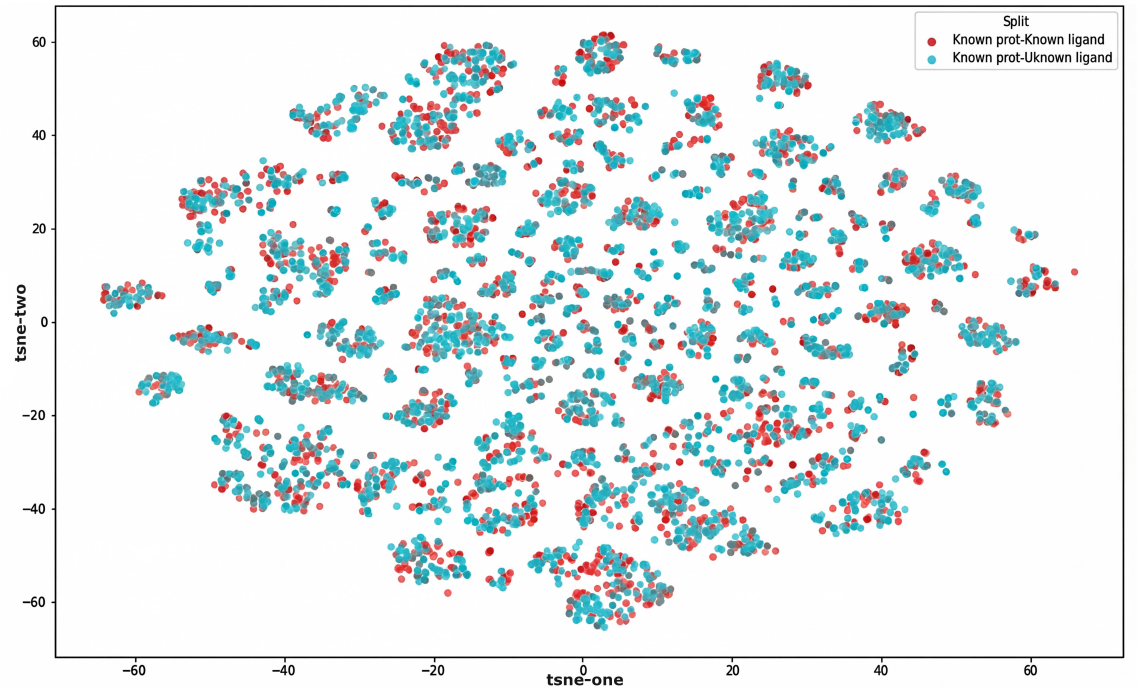

Figure 5.1: **Training vs. known protein–unknown ligand split.** Points from the two splits are largely intermixed, highlighting a low degree of distributional shift.

ligand pairs). In this case, the two distributions are strongly overlapping, with points from both splits intermixing extensively. This indicates that the test pairs are not substantially out of distribution with respect to training. Consequently, one may expect relatively high predictive performance when evaluating models under this split.

Training vs. unknown protein–unknown ligand split. A completely different picture emerges when comparing the training set with the third split, corresponding to the unknown protein–unknown ligand scenario (Figure 5.2). Here the two distributions are almost disjoint, revealing that this setting constitutes a genuine OOD challenge. This visualization foreshadows the expected poor performance of models that are trained without explicit precautions, regularizations, or augmentation strategies. It also highlights the importance of carefully defining benchmarks that reflect the intended generalization regime. For brevity, we omit the analogous comparison for the unknown protein–known ligand split, which exhibits a similar pattern; an explanation of this behavior is provided in the following paragraphs.

Protein-specific clustering of pairs. In order to better understand the source of these shifts, we colored the embeddings by protein identity rather than by split (Figure 5.3). Strikingly, we observed that compound–protein pairs cluster primarily at the level of the protein, even though the protein embedding accounts for only half of the pair representation. This strongly suggests that compounds tested against the same protein are themselves correlated, typically forming chemical series derived from the same scaffold.

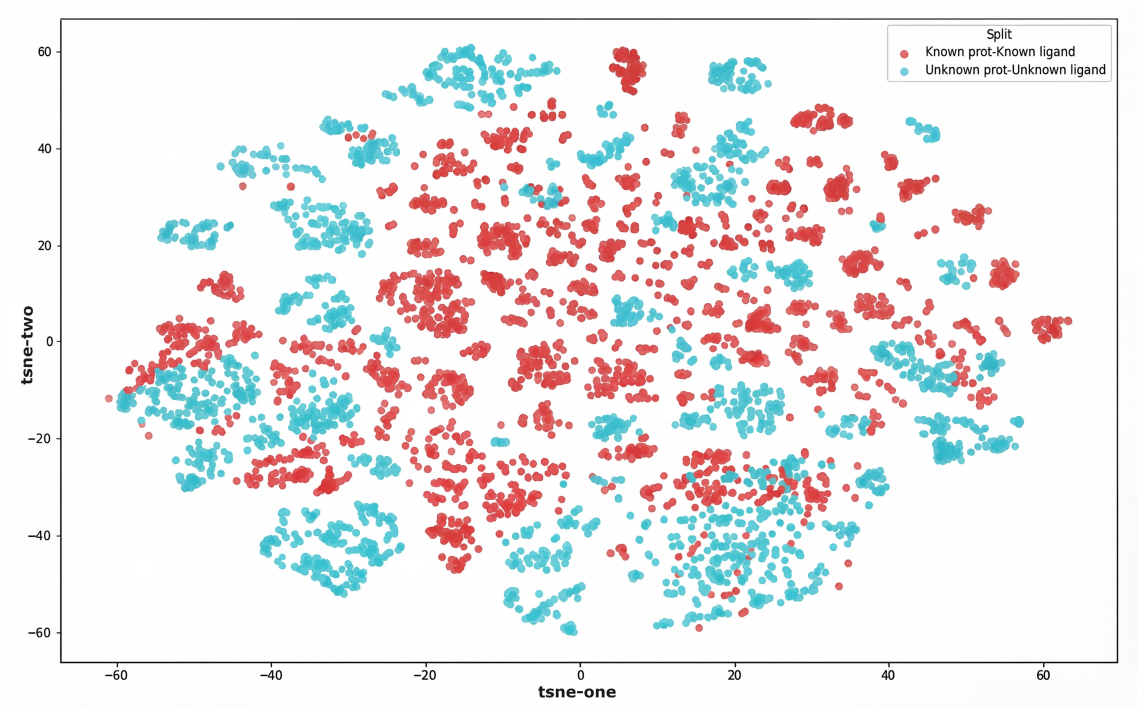

Figure 5.2: **Training vs. unknown protein–unknown ligand split.** The two distributions are largely separated, illustrating a hard out-of-distribution generalization problem.

Molecule-only embeddings colored by protein. Finally, we repeated the same procedure using only molecule embeddings, without including the protein component (Figure 5.4). When coloring molecules by the proteins with which they have been assayed, the same clustering behavior persisted. This reveals that the correlation originates not from the protein features but rather from the way compounds are organized in chemical series, reflecting dataset curation practices.

Interpretation. Taken together, this analysis demonstrates that DTI datasets are affected by strong statistical biases, often hidden under naïve splitting strategies. These biases are sufficient to explain why models evaluated on the known protein–unknown ligand split can reach deceptively high accuracy, while failing catastrophically in more realistic OOD scenarios. The observed clustering patterns, in particular the auto-correlation of ligands within protein-specific chemical series, underline the importance of carefully designing evaluation protocols. These insights laid the conceptual foundation for the development of *BindSight*, a framework explicitly designed to address these challenges.

5.2 A General Framework for Drug–Target Interaction Prediction

Framework design. BindSight was conceived not as a monolithic model, but as a flexible framework integrating all the essential stages of ligand-target interaction prediction. The guiding principle was to build a system that is modular, efficient, and deployable in

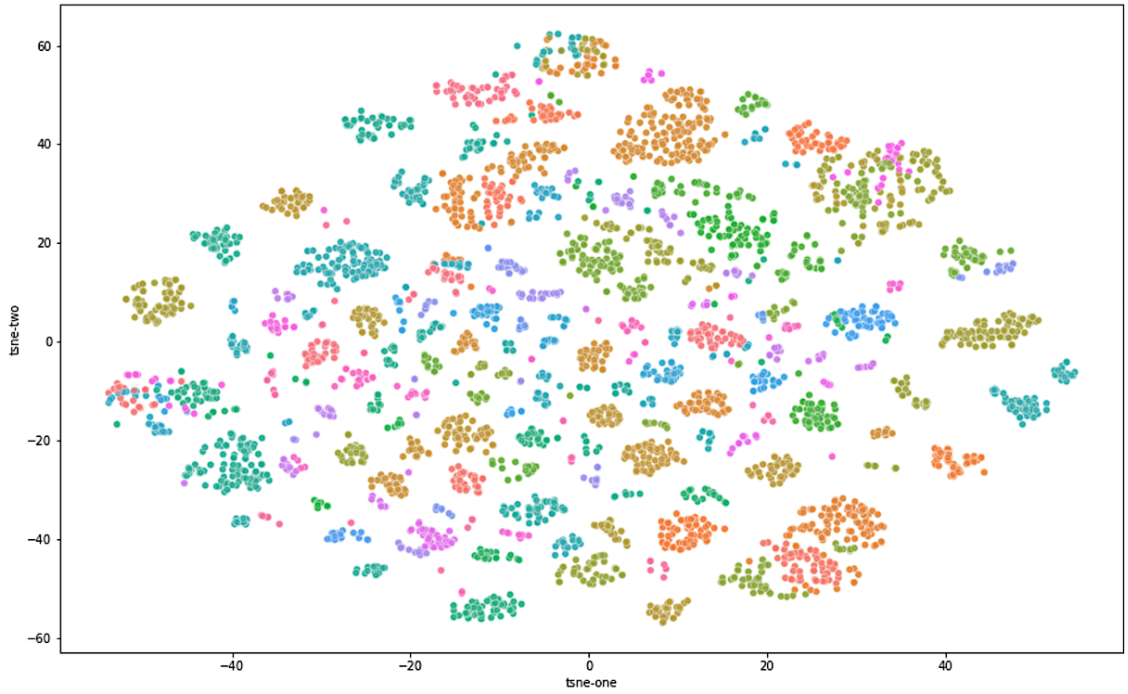

Figure 5.3: **Pair embeddings colored by protein.** Pairs form protein-specific clusters, indicating that compounds tested against the same target are themselves highly correlated. Colors indicate UniProt ID (some IDs share colors due to palette limitations).

lightweight environments such as Google Colab, thereby lowering the entry barrier for both academic and applied users. The framework supports the entire pipeline: data curation and preprocessing, feature extraction, model training and evaluation, and scalable deployment. A distinctive feature of the design is the adoption of a two-phase prediction strategy. The first phase is a rapid, large-scale virtual screening step that can quickly reduce massive compound libraries to a tractable subset of candidates. The second phase applies computationally heavier re-scoring models, enabling more refined prioritisation. This separation ensures both scalability and accuracy, and makes the framework suitable for real-world screening campaigns.

CLIP-style architecture. The modelling component adopts a contrastive learning paradigm akin to CLIP (see methods), projecting proteins and molecules into a shared latent space where interaction likelihood is scored by cosine similarity. This architecture is highly modular, allowing interchangeable feature extractors for both proteins and ligands. Importantly, embeddings can be precomputed and stored, enabling fast similarity searches at inference time. The framework is designed so that new models can be incorporated seamlessly through interchangeable embeddings, where protein and drug representations can be swapped independently via the unified `prepare_dataset()` interface, consistent model interfaces where all neural networks inherit from `torch.nn.Module` with standardized forward pass signatures, flexible data handling with custom PyTorch [269] datasets that support complex multi-modal inputs while maintaining compatibility, and experiment tracking through Weights & Biases integration that enables systematic hyperparameter optimization and model comparison.

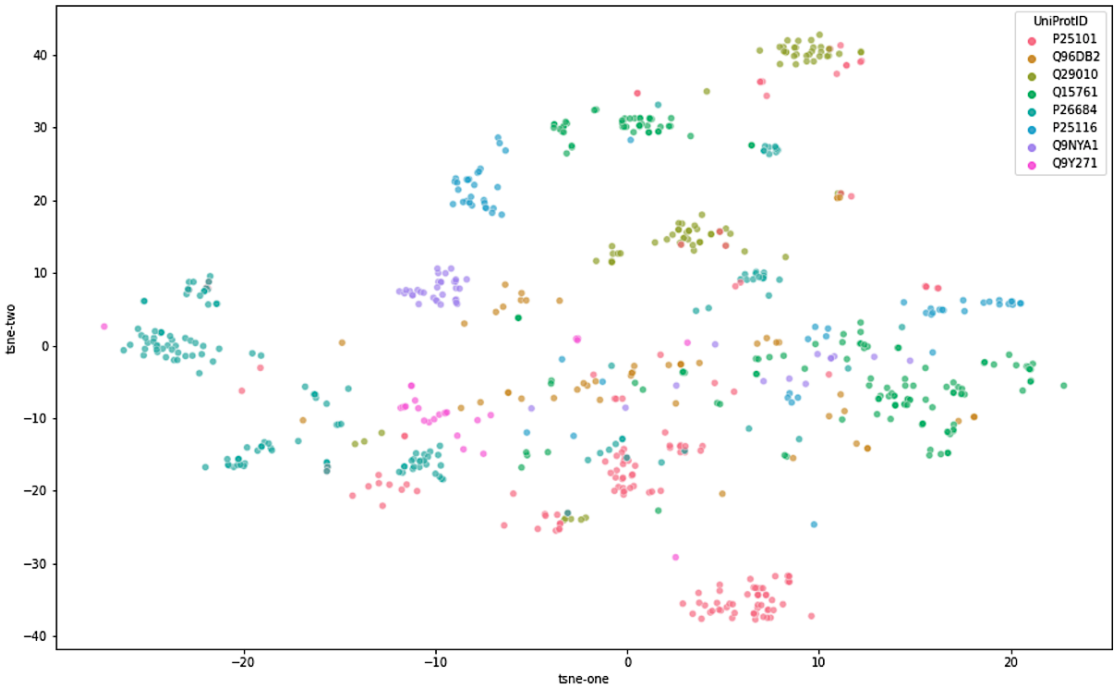

Figure 5.4: **Molecule embeddings colored by associated protein.** Even in the absence of protein features, molecules cluster according to the protein they have been tested against, revealing strong chemical-series biases.

Implemented models. The framework supports diverse protein representations such as UniProt [69] embeddings for pre-trained protein features, ESM-3 [133] embeddings from state-of-the-art protein language models, ESM-C embeddings from Evolutionary Scale Modeling, GraphDTA-style [252] encodings using one-dimensional convolutional networks on protein sequences, and ProtTrans [93] embeddings from transformer-based protein models. For drug representations, the framework accommodates Morgan fingerprints as circular molecular fingerprints, fragment embeddings for substructure-based representations, and graph neural networks using Graph Isomorphism Network encoders for molecular graphs [395]. Supporting components include custom loss functions such as FocalLoss [208] for handling class imbalance in datasets and negative sampling strategies for managing imbalanced protein-drug interaction datasets. This modular architecture enables researchers to rapidly prototype new combinations of encoders, architectures, and training strategies while maintaining reproducibility and scalability across different computational environments.

5.3 Data Curation and Enhancement

Data. Data from BindingDB [216] and Offensperger et al. [258] were used as sources. Activity classification thresholding was performed according to the thresholds established in chapter 3.4.

Quality control

Cleaning and standardisation. Reliable protein–drug interaction modeling begins with systematic filtering of raw biochemical records. The pipeline removes structurally complex entries by restricting to single-chain proteins and validates the presence of essential identifiers such as SMILES strings, UniProt IDs, and protein sequences. Integration with *Pharos* ensures that each record maps to standardized gene symbols, development levels, and protein family assignments, guaranteeing consistent biological annotation. Records lacking these features are excluded to avoid downstream inconsistencies.

Chemical standardization

Ligands undergo comprehensive structure validation using RDKit/DataMol parsing, followed by a repair and sanitization process that corrects valence issues, bond orders, and common artifacts. Molecules are normalized for tautomers, ionization states, and stereochemistry while preserving biologically relevant features. Multiple molecular formats are then generated in parallel: canonical SMILES, SELFIES, InChI/InChIKey, and Murcko scaffolds. This provides robust and complementary representations for similarity calculations and scaffold-based analysis. Exception handling ensures that failed conversions do not compromise the dataset, maximizing retention without sacrificing quality.

Binding affinities

Affinity measurements (Ki, Kd, IC_{50}) are standardized through systematic cleaning. Comparison operators are removed, values are converted to numeric form, and units are harmonized. Measurements are then transformed into the $-\log_{10}$(M) scale (pKi/pKd/pIC_{50}) to improve statistical properties. Non-positive entries are discarded, and extreme outliers are removed via quantile thresholds at the 0.1% and 99.9% levels. When multiple measurements exist for a protein–ligand pair, values are aggregated by averaging, thereby reducing assay-specific noise while retaining biological signal.

Protein annotation

Proteins are enriched with UniRef50 and UniRef90 homology clusters, enabling evolutionary grouping, cross-species comparisons, and leakage-aware data splitting. Pharos metadata further augments targets with gene nomenclature, development levels (Tdark, Tbio, Tchem, Tclin), and family assignments (e.g. kinases, GPCRs, nuclear receptors). These annotations provide biological context while ensuring that datasets remain comparable across studies.

Implementation & validation

The pipeline is optimized for scalability through parallel processing (pandarallel for ligands, ThreadPoolExecutor for UniRef mapping), memory-efficient deduplication of molecules, and robust error handling. Data integrity is preserved through automated documentation of each preprocessing step, retention of raw measurements, and standardized output formats with consistent column naming. Validation layers confirm structural correctness, measurement plausibility, and annotation completeness, ensuring that the final dataset is both comprehensive and reproducible.

Summary

Overall, this pipeline maximizes usable coverage while enforcing rigorous quality control. By integrating molecular standardization, affinity normalization, and biological annotation, it provides a reproducible and biologically consistent foundation for machine learning in computational drug discovery. A comprehensive graphical representation of the data cleaning pipeline is given in figure 5.8.

Intelligent negative sampling

Synthetic negatives. Protein–drug interaction datasets are highly imbalanced, with positive binders vastly outnumbered by potential negatives. To address this, we developed an intelligent sampling framework that avoids naïve random selection, which often introduces false negatives and biases. The method builds an *exclusion sphere* around each

protein, removing from the negative pool any ligand that is a known binder, a binder of orthologous proteins, a binder of structurally similar proteins, or a chemically similar compound to positives. This multi-layered exclusion minimizes the risk of mislabeling true interactions as negatives. To further ensure coverage of chemical diversity, the algorithm optionally tessellates chemical space using optimized k-means clustering, with stratified and grid-based sampling maintaining balanced representation across clusters. Similarities in both protein and chemical embedding spaces are computed via FAISS-accelerated [87] nearest-neighbor search, enabling scalable application to millions of compounds. Crucially, this approach yields negative sets that are both diverse and biologically plausible, providing robust training signals, reducing label noise, and improving model generalization in large-scale virtual screening tasks.

5.4 A new experiment/evaluation design

A comprehensive graphical representation of the pipeline presented in this chapter is given figure 5.9.

Protein promiscuity-based grouping. To account for the strong heterogeneity in data availability across targets, proteins were stratified into quantile-based categories according to their number of known binding interactions. Specifically, promiscuity was quantified as the count of confirmed ligand partners per protein (e.g., interactions with affinity above a defined threshold), which serves as a proxy for both biological binding breadth and experimental study depth. Using tertile partitioning, proteins were grouped into three categories: Q1, representing low-promiscuity proteins with few or no known ligands (often corresponding to novel, difficult, or rare disease targets); Q2, capturing proteins with moderate numbers of interactions and emerging translational relevance; and Q3, encompassing highly promiscuous, well-characterized proteins such as kinases and GPCRs that dominate drug discovery pipelines. This stratification enables balanced evaluation and optimization of models across targets with very different levels of characterization.

Greedy stratified splitter. We developed a scaffold-aware stratified splitting algorithm to obtain unbiased train/validation/test partitions for protein–drug interaction prediction. Instead of random splitting, which risks structural leakage, compounds are grouped by Bemis–Murcko scaffold so that all analogues fall into the same split. Stratification is performed simultaneously across three dimensions: (i) proportional coverage of protein families (enzymes, GPCRs, others), (ii) consistent ratios of binding vs. non-binding pairs, and (iii) balanced representation of protein promiscuity quartiles. To allocate scaffolds, the algorithm computes a composite cost function combining a *distribution cost*, the absolute deviation of each split's label frequencies from the global targets, and a *capacity penalty* discouraging overfilling. Scaffolds are greedily assigned to the split minimizing this combined cost, iteratively preserving both target distributions and split sizes. The procedure begins with a global distribution analysis, aggregates label counts per scaffold, and proceeds until all scaffolds are placed, after which final splits are validated for consistency. By maintaining structural integrity and multi-dimensional balance, this approach yields evaluation sets that reflect the biological heterogeneity of

the dataset, ensuring fair model comparison, robust hyperparameter selection, and realistic estimates of generalization.

Focal Loss Originally proposed for object detection tasks in computer vision [208], Focal Loss has since been widely adopted across domains characterized by severe class imbalance. In the BindSight framework, it is employed to address the skewed distribution of protein–drug interaction datasets, where positive binding interactions are significantly outnumbered by negatives. This loss function extends the standard binary cross-entropy by introducing a focusing parameter γ that down-weights the contribution of easy-to-classify examples, thereby encouraging the model to concentrate on harder, more informative cases. In addition, an α weighting factor balances the importance between positive and negative classes, with a default value of 0.75 favoring the minority positive class. The focal loss is computed as $\alpha(1 - p_t)^\gamma \log(p_t)$, where p_t represents the model's estimated probability for the true class, effectively reducing the relative loss for well-classified examples while maintaining the loss for misclassified ones.

Multi-objective search. We optimised hyperparameters using an Optuna [10] framework with a Tree-structured Parzen Estimator (TPE) [381] sampler across 500 trials, each evaluated by 3-fold cross-validation under the stratified scaffold splitting scheme described above. This ensured that performance estimates reflected true generalization while maintaining balanced distributions across relevant biological dimensions. The optimisation pursued three objectives: maximising AUPRC on Q1 proteins, maximising AUPRC on Q2 proteins, and favouring smooth performance across folds to limit overfitting—thereby addressing the dual challenge of achieving accuracy on both well-characterised and poorly studied targets. Hyperparameter spaces were specified via YAML templates with fixed core settings (e.g. embedding types, loss categories) and conditional sampling for architecture-specific parameters, spanning categorical, integer, and log-scaled continuous domains. For MLPs, the search explored layer depth (1–4), hidden dimensions (128–4096), and shared versus separate protein/drug encoders, whereas GraphDTA trials focused primarily on training dynamics due to its fixed GNN architecture. Loss functions were tuned through focal loss parameters ($\alpha = 0.5$–1.0, $\gamma = 1$–7, binary vs. continuous), along with batch size (64–512), learning rate (10^{-6}–10^{-4}), weight decay (10^{-4}–10^{-1}), and early stopping (2–5 epochs). Scalability was achieved through distributed trial execution, database-backed study management and Optuna pruning for early termination of underperforming runs. All experiments were tracked with Weights & Biases, ensuring reproducibility and interpretability through detailed logs of configurations, training curves, and optimal epochs. This framework enables computationally efficient and biologically robust hyperparameter selection, supporting fair model comparison and reliable generalization in large-scale drug discovery pipelines.

5.5 Preliminary results

All results are reported under a strict *scaffold-out* evaluation: test compounds (or fragments) have Bemis–Murcko scaffolds never observed during training, and proteins are assessed within promiscuity strata (Q1–Q3) as defined previously. Because of the extreme class imbalance typical of DTI data, we focus on the area under the precision–recall

curve (AUPRC), which is more informative than ROC-AUC in this regime [307]. We compare Binary Cross-Entropy (BCE) with Focal Loss [208] while keeping architectures and the stratified scaffold splitting protocol fixed.

Effect on Understudied Targets (Q1) Across both datasets, Focal Loss consistently improves the AUPRC for Q1 proteins compared to BCE (Fig. 5.10). The improvements are robust, as indicated by a higher median and a tighter interquartile range across folds. This aligns with the design of Focal Loss, which re-weights training towards hard, minority-class examples [208].

Early Enrichment and PR-Curve Morphology Beyond summary metrics, precision–recall curves highlight qualitative improvements (Fig. 5.5). Under Focal Loss, the curves exhibit a sharp rise at very low recall followed by a gradual decline, indicating strong *early enrichment*. This means that the top-ranked predictions are highly enriched in true positives, a desirable property in screening pipelines where only the top predictions can be experimentally validated. By contrast, BCE curves show a rapid drop in precision, reflecting weaker separation of positives at the top of the ranking.

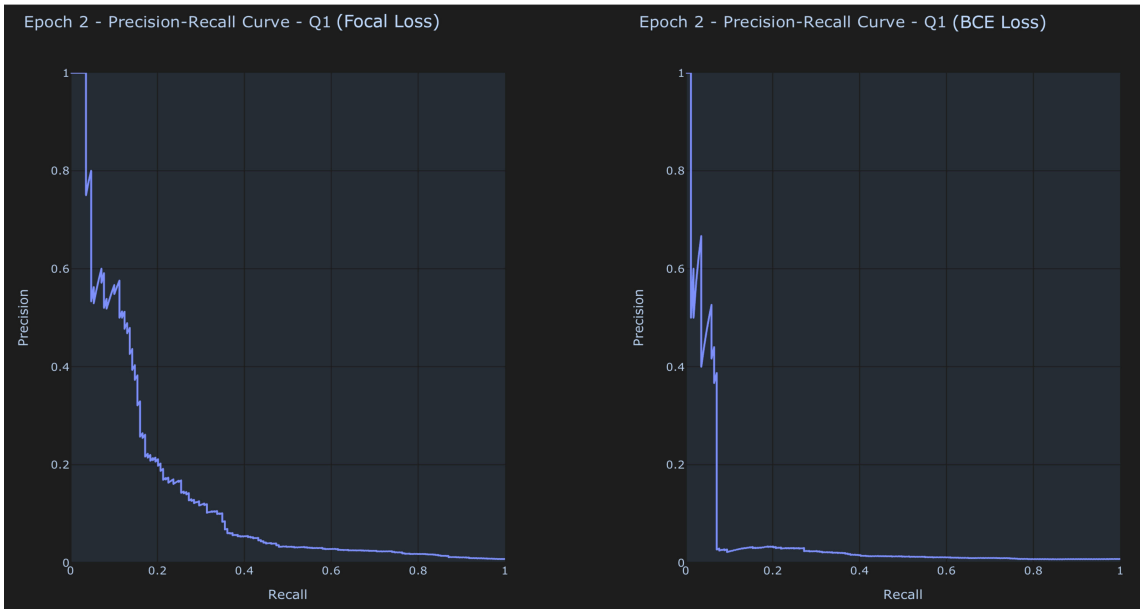

Figure 5.5: **Precision–recall curves for Q1 proteins.** Focal Loss (on the left) produces curves with high precision at low recall and a slower decline thereafter, consistent with superior early enrichment. BCE (on the right) shows rapid precision loss.

Score Distribution and Thresholding A complementary perspective is provided by the distribution of predicted probabilities (Fig. 5.6). With BCE, nearly all scores fall below 0.5, meaning that no Q1 examples cross a standard classification threshold. In contrast, Focal Loss generates a well-separated high-confidence tail, with several predictions surpassing 0.5. This not only improves ranking (as captured by AUPRC) but also yields predictions that are directly usable in binary decision-making without requiring dataset-specific threshold calibration.

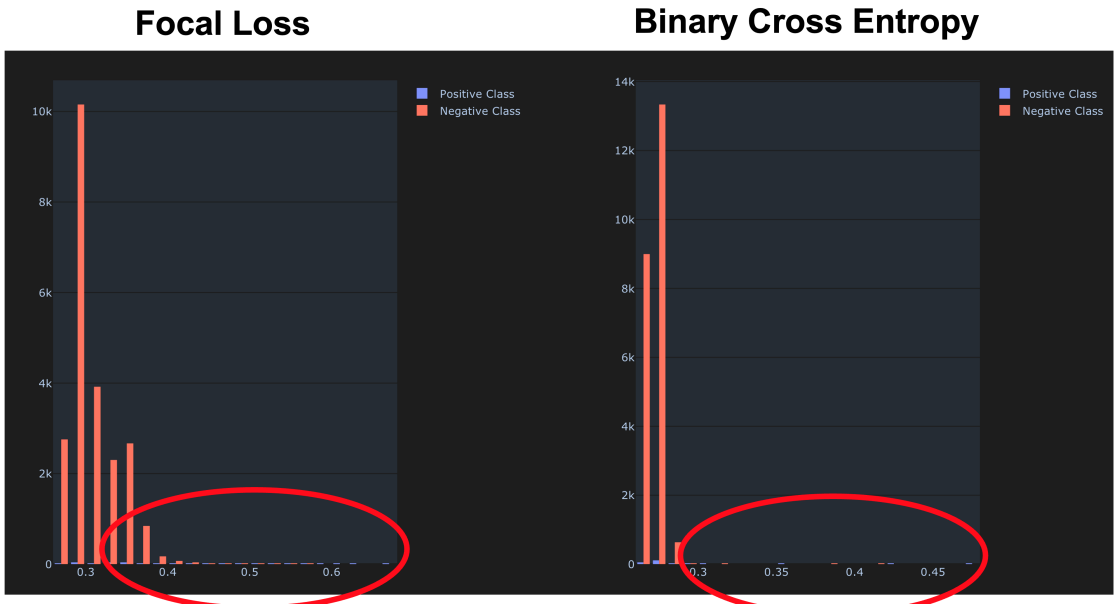

Figure 5.6: **Distribution of predicted scores.** BCE predictions (right) collapse below 0.5, limiting threshold-based decisions. Focal Loss (left) generates a distinct tail of high-confidence scores (red circles), enabling practical triage.

Evaluating representation learning strategies

Trying different featurizations. The modular design of the *BindSight* framework allowed us to systematically evaluate how alternative protein representations influence model performance. The choice of protein featurization was incorporated as a tunable hyperparameter within the optimization procedure described in Section 5.4. We compared three distinct protein language model (PLM) embeddings: ProtTrans, ESM-3, and ESM-C. ProtTrans embeddings [93] (model `prottrans_t5_xl_u50`) were retrieved directly from UniProt, where they are available as pre-computed representations, while ESM-3 [133] and ESM-C [94] embeddings were generated through the evolutionaryscale.ai API.

Multimodal PLMs do not always generalize better

The optimization ran for one thousand trials. Among the top one hundred configurations, ranked by the harmonic mean of AUPRC on Q1 and Q2 proteins from the Offensperger et al. [258] dataset, ProtTrans-based models accounted for 56%, ESM-3 for 29%, and ESM-C for 15% of the best-performing trials (see Figure 5.7A). This outcome indicates that ProtTrans achieved superior performance for this task, despite being an older and smaller model than the other two PLMs. Interestingly, ProtTrans also outperformed ESM-3, even though ESM-3 is a generative model that jointly learns from sequence, structural, and functional signals, which in principle could yield richer and more transferable embeddings. We speculate that ESM-3's richer, multimodal representations, while capturing more biological information, may also introduce greater susceptibility to overfitting in the presence of noisy labels (a concern we have highlighted throughout the chapter). However, this remains a hypothesis that would require further investigation to confirm.The distribution of the top fifty trials for each embedding type, shown in Figure 5.7B, further confirms this trend and suggests that model size and multimodal objectives do not necessarily guarantee improved transferability across biochemical prediction tasks.

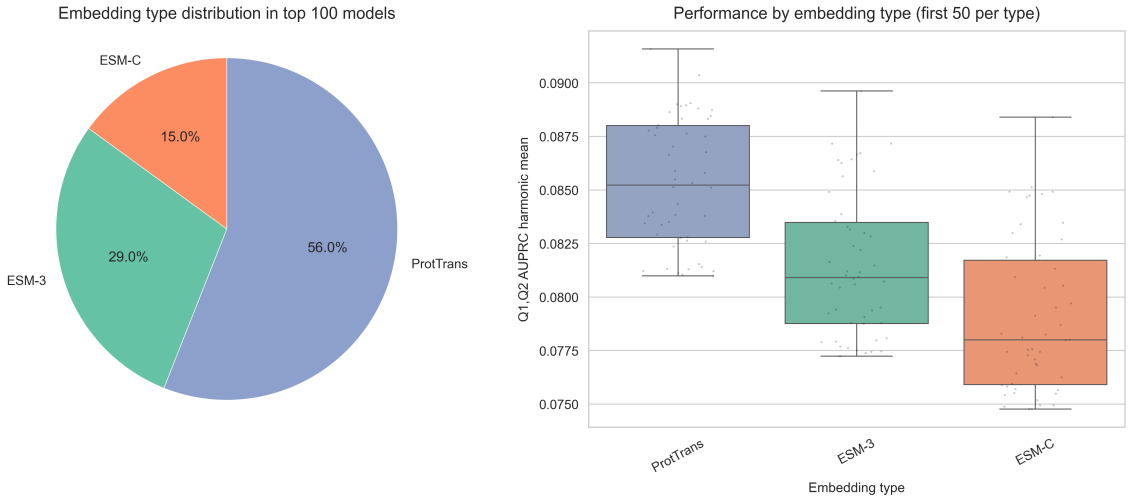

Figure 5.7: **Impact of protein embedding choice on *BindSight* performance.** (A), Distribution of embedding types among the top one hundred Optuna trials shows that ProtTrans dominates with 56% of the best configurations, followed by ESM-3 with 29% and ESM-C with 15%. (B), Performance distribution of the first fifty trials per embedding type measured as the harmonic mean of AUPRC on Q1 and Q2 proteins on Offensperger et al. [258]. Points indicate individual trials and boxes indicate the interquartile range with whiskers extending to 1.5 times the interquartile range. Together, these panels show that ProtTrans yields higher median performance in this setting, indicating that larger or multimodal PLMs do not necessarily confer better transfer across these biochemical prediction tasks.

Summary. Taken together, these experiments highlight both the benefit of Focal Loss and the importance of the proposed evaluation framework. By enforcing scaffold-out splits and stratifying proteins by promiscuity levels, we are able to disentangle performance across well- and poorly studied targets. This design reveals that Focal Loss is particularly effective on Q1 proteins, where data scarcity and imbalance are most severe, and that improvements are not artefacts of scaffold leakage or family imbalance. Moreover, the combination of stratified splitting, loss function comparison, and multi-objective hyperparameter search provides a comprehensive and biologically meaningful assessment protocol. Rather than reporting aggregate scores dominated by Q3 proteins, our framework surfaces nuanced differences in model behaviour, such as early enrichment and thresholdable predictions. These insights would remain hidden under conventional random splits or single-metric evaluation.

5.5.1 Two-stage prediction strategy.

Retrieval and re-scoring pipeline

In the first stage, candidate interactions are retrieved through a CLIP-like encoder, which aligns protein and ligand representations within a shared latent space to efficiently narrow down the search domain. This stage produces a ranked shortlist of protein–ligand pairs that are then passed to a more computationally intensive second stage based on TabPFN re-scoring.

Drug-specific TabPFN models

In preliminary experiments, we focused on proteins with fewer than one hundred annotated interactions in BindingDB, representing a particularly challenging subset. Each TabPFN model was trained in a drug-specific manner, relying solely on molecular fingerprints and compound-specific bioactivity data. This design allowed the model to learn local structure–activity patterns unique to each drug, while still benefiting from the strong inductive bias encoded in TabPFN's prior-fitting procedure. Because the current TabPFN implementation is limited to roughly five hundred input features, the 2048-dimensional Morgan fingerprints were reduced via principal component analysis (PCA) before training. Although imposed by model constraints, this dimensionality reduction also helped mitigate overfitting in the low-data regime. Preliminary results on this low-sample BindingDB subset achieved AUROC $\approx 85\%$ and AUPRC $\approx 76\%$, supporting the effectiveness of the two-stage design in identifying relevant binders even under severe data scarcity.

5.6 Software Engineering Contributions

Parallel ensembles. A single MLP does not saturate modern GPUs. To better utilise available resources, we developed parallel ensemble architectures combining multiple models with parallel batch normalisation and coordinated batch management. This enables simultaneous training of several learners on distinct data subsets, maximising throughput and improving ensemble performance. The parallel implementation maintains independent parameter sets for each model in the ensemble, allowing efficient parallel matrix operations across multiple neural networks. Each model processes the full dataset independently with optional per-model batch normalization that maintains separate running statistics, ensuring proper gradient flow and parameter updates. Coordinated batch management synchronizes data loading across multiple loaders, handling varying batch sizes and ensuring balanced training across ensemble members. This architecture supports flexible ensemble sizes, typically ranging from 4 to 16 models, leading to improved generalization through model diversity and significantly higher computational resource utilization compared to sequential training approaches.

HPC-friendly infrastructure. The framework integrates seamlessly with HPC environments. Hyperparameter optimisation can be distributed across multiple processes, coordinated by custom utilities to ensure reproducibility. Experiment tracking is handled through Weights & Biases, and most configuration is externalised to YAML files, allowing users to modify settings without changing the codebase. Bayesian optimization techniques are employed with distributed sampling coordination that maintains reproducible random seeds across multiple processes using persistent storage mechanisms. Comprehensive experiment tracking provides automatic logging of performance metrics, visualization plots, and training statistics. Configuration management uses templated files with conditional logic supporting different model architectures and dataset-specific parameters, enabling environment-independent deployment. The configuration system supports parameter overrides for custom mappings, data specifications, and training settings, while maintaining reproducibility through seeded random number generation and deterministic data processing across different computational environments.

Future-proof modularity. The codebase is designed for extensibility. Hooks are included to accommodate new datasets, feature extractors, and model architectures, making the framework adaptable to future methodological developments. The modular architecture centers around unified interfaces for data preparation and feature extraction, supporting multiple established datasets while providing extension points for custom implementations. Dataset preparation functions handle diverse data sources and formats, while interchangeable embedding modules provide various protein and molecular representations including advanced language models, traditional fingerprints, and graph-based approaches. The model implementations follow consistent interfaces with standardized neural network inheritance patterns, supporting both conventional architectures and parallel variants for ensemble training. Advanced loss functions handle class imbalance and continuous optimization objectives, while specialized layers enable sophisticated training techniques. This design philosophy allows researchers to rapidly prototype combinations of encoders, architectures, and training strategies without modifying core framework components, ensuring long-term adaptability to emerging methodologies in drug-target interaction prediction.

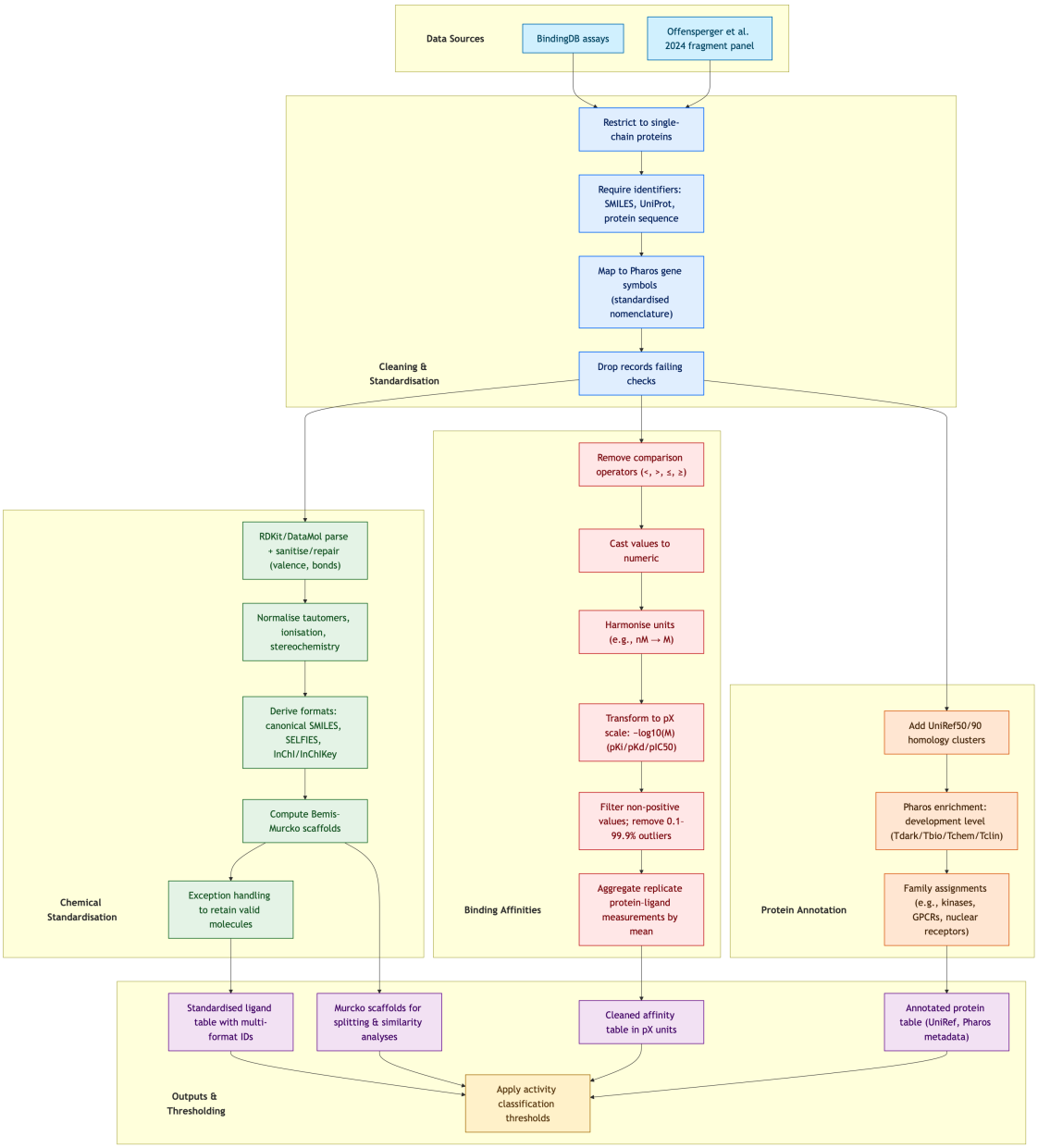

Figure 5.8: **Curation and standardisation pipeline for protein–ligand data.** Data originate from BindingDB and the Offensperger et al. fragment panel [216, 258]. Records undergo quality control filters (single-chain proteins; required SMILES, UniProt, and sequence identifiers), followed by chemical standardisation with RDKit/DataMol (sanitisation, normalisation of tautomers/ionisation/stereochemistry, derivation of canonical SMILES/SELFIES/InChI, and Bemis–Murcko scaffolds). Affinity data are cleaned in discrete steps—removal of comparison operators, numeric casting, and unit harmonisation—then transformed to pX scale (pKi/pKd/pIC_{50}), trimmed for non-positive values and 0.1–99.9% outliers, and aggregated across replicates. Proteins are annotated with UniRef50/90 clusters and Pharos metadata (development level and family). The outputs are standardised ligand identifiers, a cleaned pX affinity table, an annotated protein table, and scaffold sets; activity classification thresholds are applied as defined in Chapter3.4.

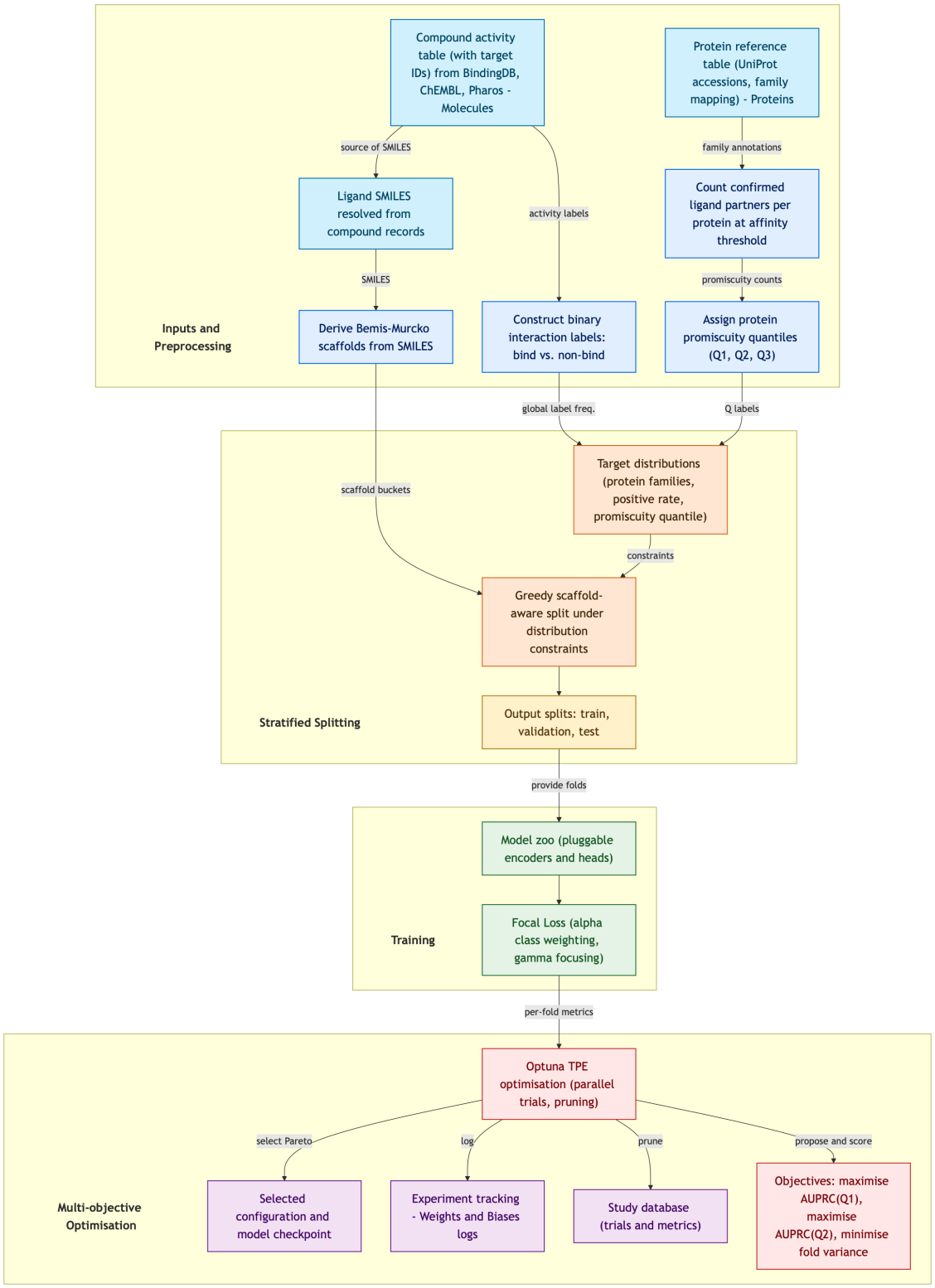

Figure 5.9: **BindSight experiment–evaluation workflow.** Inputs and preprocessing derive SMILES, scaffolds, and protein promiscuity quantiles from curated molecule–protein records; a scaffold-aware greedy splitter enforces balanced distributions across families, label prevalence, and promiscuity. Models are trained with focal loss to address class imbalance, and hyperparameters are tuned via Optuna TPE in a multi-objective scheme that maximises AUPRC on low- and mid-promiscuity targets (Q1, Q2) while minimising cross-fold variance.

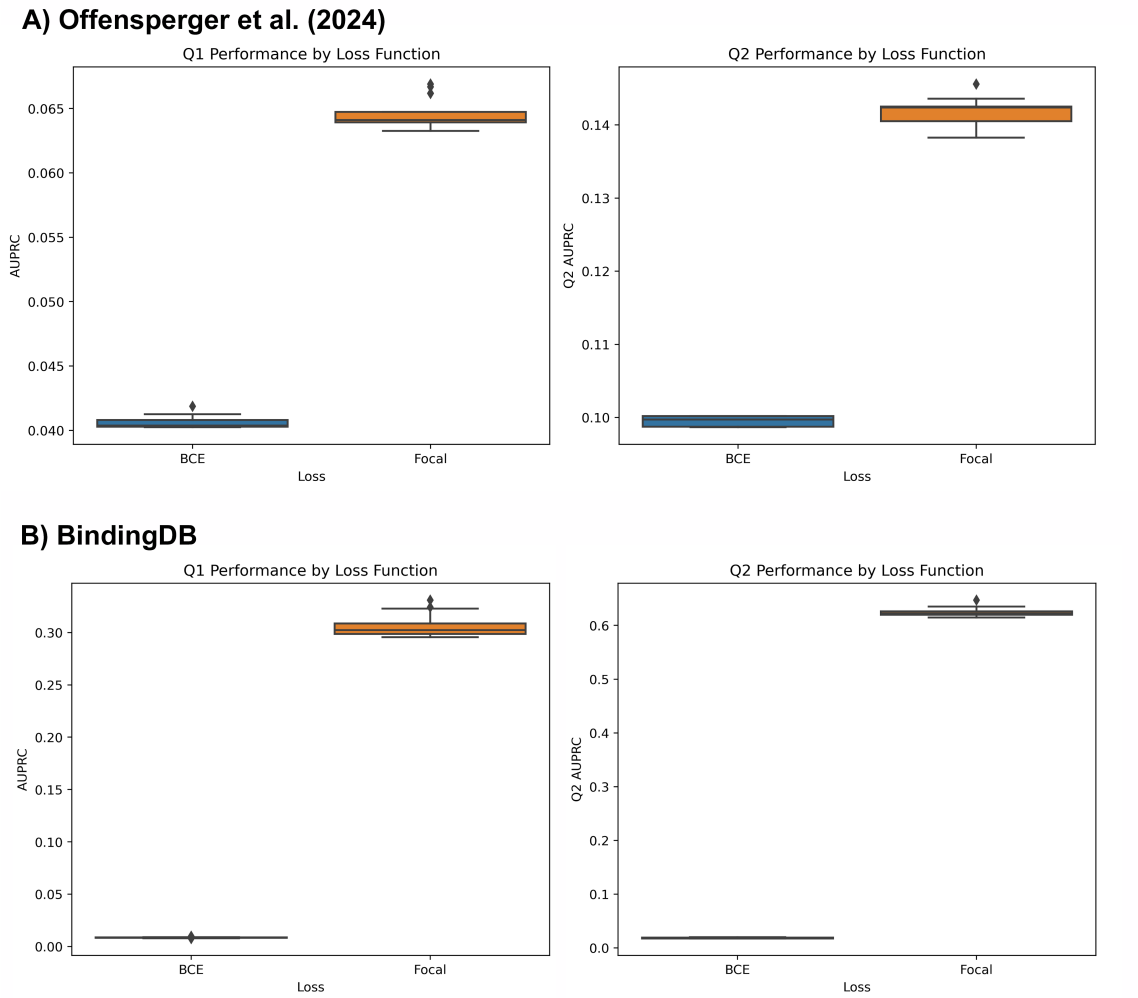

Figure 5.10: **AUPRC comparison across datasets on Q1 and Q2 proteins.** Boxplots show performance for BCE (blue) and Focal Loss (orange) under scaffold-out evaluation and Q1 (on the left) and Q2 (on the right) stratification. Picture showcase good performance of FocalLoss on both Offensperger et al. [258] (A) and BindingDB (B) datasets.

Chapter 6

Learning and actioning general principles of cancer cell drug sensitivity

This chapter is based on the journal paper F. Carli, P. Di Chiaro, M. Morelli, C. Arora, L. Bisceglia, N. De Oliveira Rosa, A. Cortesi, S. Franceschi, F. Lessi, A. L. Di Stefano, et al. Learning and actioning general principles of cancer cell drug sensitivity. *Nature Communications*, 16(1):1654, 2025

6.1 CellHit: a Scalable and Interpretable Drug Sensitivity Prediction model

We developed *CellHit*, a machine learning framework designed to predict cancer cell line drug sensitivity from transcriptomic data and to deploy these predictions on patient bulk RNA-seq profiles following alignment via *Celligner* [380]. The framework was designed to achieve high predictive accuracy while maintaining model interpretability, allowing for the extraction of biological insights from the predictive process.

Design rationale and model selection. Two principal strategies where benchamarked for representing the DRP problem:

1. *Joint drug-cell models*, where a unique models is trained on a combination of both molecular descriptors and gene expression profiles of cell lines;
2. *Per-drug models*, where a separate predictive model is trained for each compound using only the transcriptomic features of the cell lines.

For the joint models, we explored different ways of representing drugs, including extended-connectivity fingerprints [295], embeddings from ChemBERTa [9], and simple one-hot encodings. Each representation was combined with full transcriptomic profiles of the cell lines and evaluated using two machine learning algorithms, namely XGBoost [60] and multi-layer perceptrons, with performance compared against published benchmarks [56]. As shown in Figure 6.2, the simplest representation (one-hot encoding) of

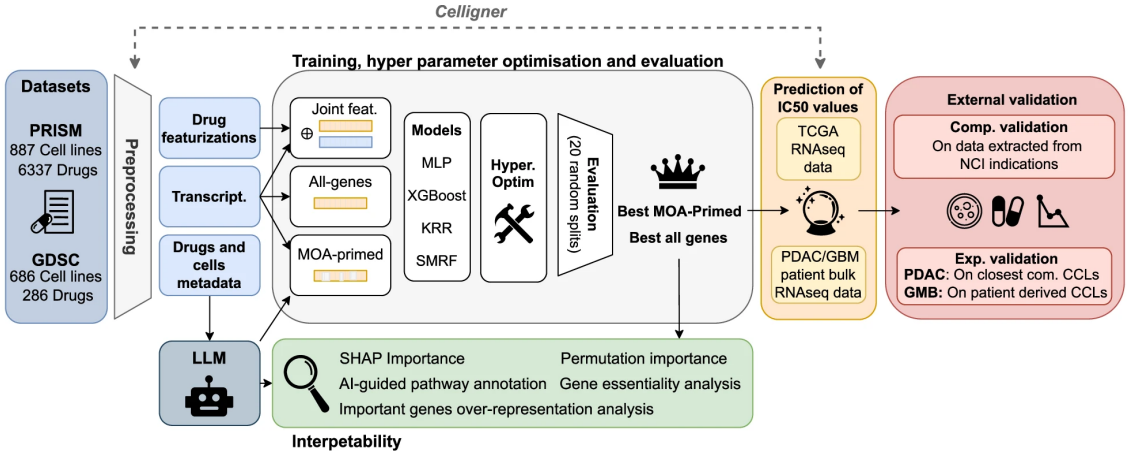

Figure 6.1: **CellHit framework.** Overview of the pipeline including data sources (GDSC, PRISM), preprocessing, Celligner alignment, model families, evaluation strategy, and interpretability components.

the drug identity achieved the highest predictive accuracy, with a Pearson correlation of $\rho = 0.89$ and a mean squared error (MSE) of 1.55 on the GDSC dataset, outperforming the more complex fingerprint and embedding approaches. This result suggested that most of the predictive signal was contained within the transcriptomic profiles, while the chemical features contributed relatively little. In other words, the models were primarily capturing drug-specific response signatures rather than relying on molecular similarity across compounds. Motivated by this finding, and aiming to maximise both interpretability and computational scalability, we adopted a *per-drug modelling* strategy. Instead of learning from drug features, each model was trained exclusively on the transcriptomic profiles of all available cell lines for a single compound. We selected XGBoost as the core learning algorithm because of its strong performance on high-dimensional tabular data, its competitiveness in previous benchmarks, and the straightforward interpretability of its tree-based feature importance measures. To ensure a balanced trade-off between predictive accuracy and robustness, hyperparameters were tuned through a multi-objective optimisation procedure that jointly considered correlation and mean squared error.

Data and preprocessing. For model training, we relied on two major pharmacogenomic resources. First, the *GDSC2* dataset [153], which contains measurements of half-maximal inhibitory concentrations (IC_{50}) for 286 drugs tested across 686 cancer cell lines, yielding 169,208 drug–cell line pairs. Second, the *PRISM Repurposing* dataset [70], a large-scale screen comprising 6,337 drugs evaluated against 887 cell lines, resulting in approximately 3.81 million drug–cell line associations, with responses quantified as single-dose log-fold change (LFC) values. For both resources, we integrated transcriptomic profiles obtained from the Cancer Cell Line Encyclopedia (CCLE) RNA-seq data [26]. Gene expression values were processed through $\log_2(\text{TPM} + 1)$ transformation and harmonised across the GDSC and PRISM cell line identifiers to ensure consistency. To facilitate translation toward clinical applications, we further aligned the cell line transcriptomes with bulk RNA-seq data from patient tumours using the *Celligner* [380] method, enabling downstream deployment of the models in a clinically relevant context.

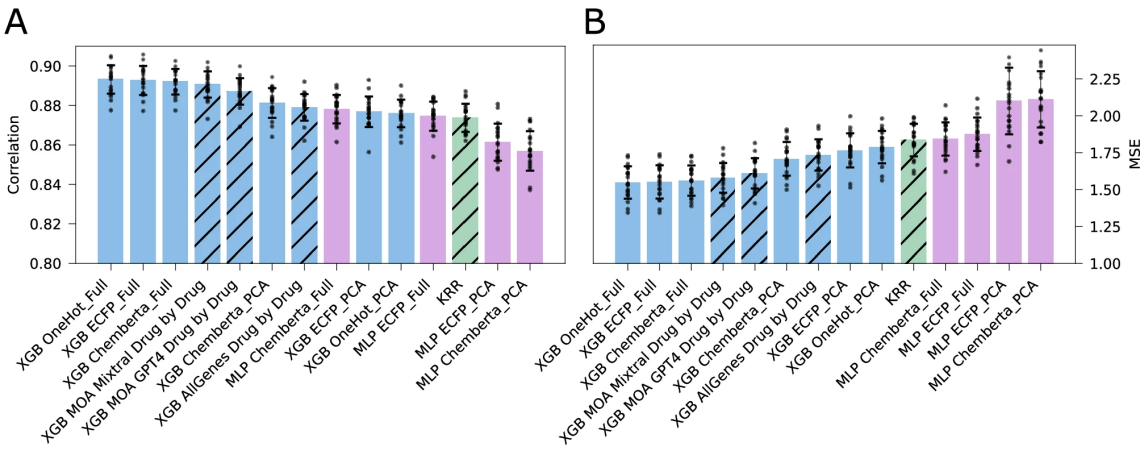

Figure 6.2: **Performance of all trained models.** (A) Bar plot comparing the performance of different model architectures (MLP, XGBoost, and literature baselines) and input feature representations (cell features and drug features) in terms of Pearson correlation with observed drug sensitivities. Different colors denote learning algorithms (e.g., light blue XGBoost and purple MLP). Etched bars highlight models using only transcriptomic data. Results are averages over 20 distinct test splits; error bars show SD. (B) Bar plot of Mean Squared Error (MSE) for the same models as in (A), averaged over 20 test splits; error bars show SD.

Evaluation strategy. To prevent data leakage and ensure that predictive performance was not inflated by tissue-specific similarities, we generated train, validation, and test partitions by stratifying cell lines according to their tissue of origin, as defined by the OncoTree classification system [196]. This stratification guaranteed that models were evaluated on cell types distinct from those used during training, thereby providing a more realistic assessment of generalisation. To obtain robust and statistically reliable estimates, all experiments were repeated across 20 independent random splits of the data, and the results were aggregated. Model performance was quantified using Pearson correlation (ρ), which measures the strength of linear association between predicted and observed responses, and MSE, which captures the average magnitude of prediction error. These metrics were reported both as global summary scores across the entire dataset and at the per-drug level, enabling a fine-grained evaluation of predictive accuracy.

Performance on GDSC. When applied to the GDSC dataset, the per-drug models reached an overall predictive accuracy that was nearly identical to that of the joint models, with an average Pearson correlation of $\rho = 0.88$ and a MSE of 1.73 (see Figure 6.2).At the individual compound level, performance varied markedly (see Figure 6.3). The distribution of per-drug correlations centred around a median of $\rho = 0.40$ (mean $= 0.41$, s.d. $= 0.12$), but with a clear right tail: for example, the Venetoclax model reached $\rho = 0.72$, indicating that some compounds are particularly amenable to prediction from transcriptomic features alone. Notably, one quarter of drugs achieved $\rho > 0.5$ across the test splits, showing that gene-expression profiles by themselves capture substantial predictive signal for a large subset of compounds.

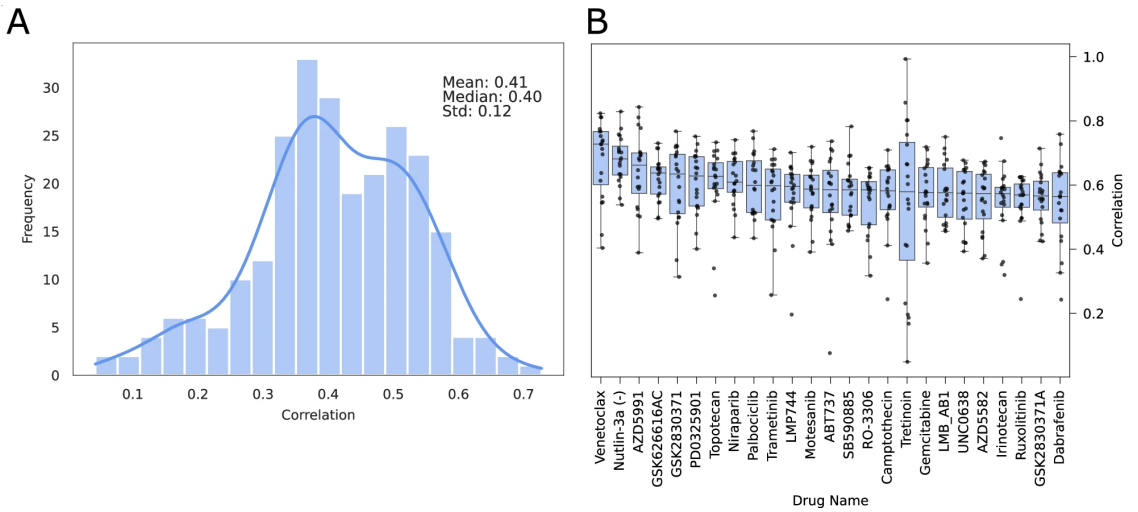

Figure 6.3: **Per-drug performance of *CellHit*** on GDSC. (A) Histogram of Pearson correlation coefficients for drug-specific models using all genes, showing median, mean, and standard deviation. (B) Box plots illustrating variability across 20 random training/testing splits. Each box shows the median (central line), interquartile range (box edges), and whiskers for variability.

Interpretability. A defining strength of *CellHit* lies in its emphasis on interpretability. By training separate XGBoost models for each drug, the framework makes it possible to directly estimate the contribution of individual genes to drug response. These contributions are quantified using two complementary strategies: SHAP values, which provide a consistent measure of each gene's marginal effect on model predictions [223], and permutation importance, which evaluates how prediction accuracy changes when the values of a gene are randomly shuffled [39]. A gene is only considered relevant if both methods independently highlight it, a requirement that reduces noise and increases confidence in the selected features. The resulting gene sets not only improve transparency in model decision-making but also serve as a starting point for generating biological hypotheses about the mechanisms underlying drug sensitivity.

6.2 LLM-Guided Curation of Mechanism-of-Action Pathways

Overview of the pipeline. We built a systematic and reproducible pipeline to annotate the biological pathways that are most likely involved in the MOA of specific drugs. The approach combines two complementary resources: LLMs, which provide flexible text-based reasoning capabilities, and curated pathway databases, which ensure biological accuracy and grounding. In practice, the pipeline leverages either the proprietary GPT-4 model [5] or the open-source Mixtral Instruct 8×7B MOE [162] model and the Reactome [159] pathway knowledge base. This pipeline introduces techniques to minimise typical LLM failure modes such as hallucinated links or unsupported claims.

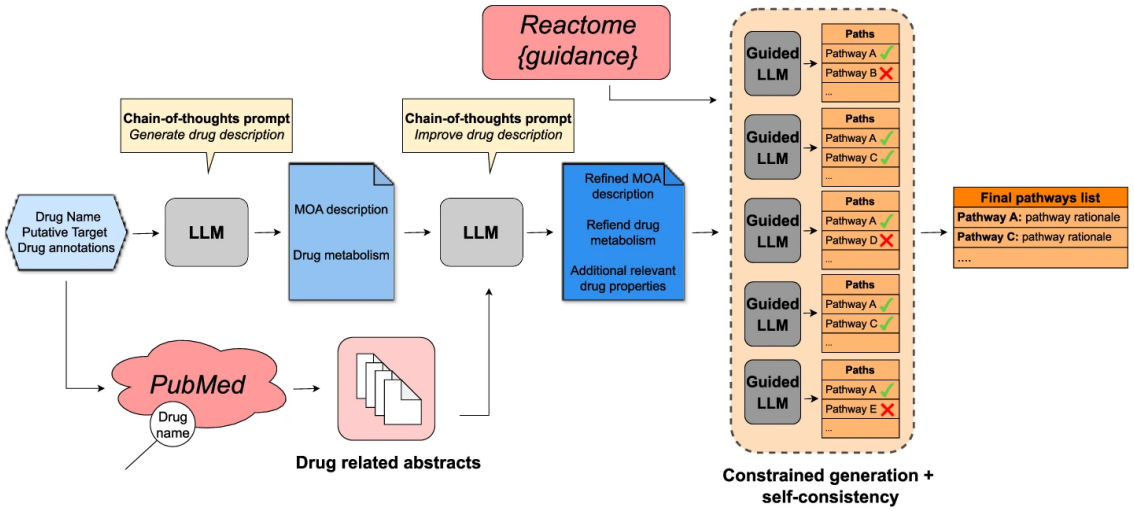

Figure 6.4: **LLM-guided annotation of drug–pathway associations.** Workflow depicting the use of a LLM for generating drug MOAs and identifying semantically relevant pathways. Starting from the drug's available metadata, an LLM is repeatedly tasked with specialized prompts to generate a drug textual description. In parallel, PubMed is queried programmatically with the drug name to retrieve abstracts related to the drug. The information is integrated in a final textual description. The obtained drug description is used by a "Guided" LLM to choose which are the Reactome pathways which are most likely to modulate drug efficacy. This last procedure is repeated 5 different times and only pathways selected at least two times are retained.

Pre-processing of Reactome pathways. To use Reactome effectively within our pipeline, we first converted its hierarchical pathway structure into a directed acyclic graph (DAG). In this graph, each node corresponds to a pathway, while edges indicate parent–child relationships defined in the curated hierarchy. By applying a topological sorting algorithm to the DAG, we were able to arrange the pathways into successive levels: pathways at level 0 have no parents and represent broad biological processes, pathways at level 1 are direct descendants of those top-level categories, and deeper levels correspond to increasingly specific biological functions. This hierarchical organisation captures the granularity of Reactome while also allowing us to control how much detail is passed to the language model. Without such filtering, the model could be burdened with highly redundant or overly fine-grained pathway annotations, which might obscure rather than clarify the drug–pathway associations. To strike a balance between interpretability and biological breadth, we focused our analyses on level 1 pathways, resulting in a curated set of 171 pathways that provide both manageable complexity and sufficient coverage of major biological processes.

Two-stage prompting and retrieval augmentation. For each drug in the GDSC and PRISM datasets, the pipeline starts by gathering metadata such as the compound name, synonyms, known or predicted targets, and MOA descriptions. This information is then provided to the LLM in two consecutive steps. In the first step, the model produces a detailed free-text description of the drug, expanding on its pharmacological and biochemical context. In the second step, it narrows this description down to a set of

relevant Reactome *Level 1* pathways that are most likely to modulate the drug's efficacy. As illustrated in Figure 6.4, the pipeline optionally integrates retrieval-augmented generation (RAG), where the LLM processes PubMed abstracts linked to each compound to enrich drug descriptions with literature-based evidence. To improve robustness, a refinement stage further elaborates on MOA and drug metabolism, generating additional properties when available. Pathway assignment is then performed through a constrained generation process using the `guidance` library, which frames the task as structured selection among predefined Reactome pathways. To guard against spurious associations, this selection is repeated independently across five runs, and only pathways consistently recovered in at least two runs are retained. This self-consistency mechanism improves reproducibility, reduces false positives, and ensures that the final curated pathway list reflects both biological plausibility and LLM stability.

Expansion of MOA coverage. By combining large language models with pathway knowledge bases, we were able to broaden the coverage of drug mechanisms of action well beyond what was available from existing curated resources. In the GDSC dataset, the pipeline successfully recovered pathway associations for 253 out of 287 drugs, corresponding to 88% of the compounds and resulting in 138 distinct Reactome pathways linked through 5,662 curated drug–pathway associations. In the much larger PRISM dataset, we annotated mechanisms of action for 6,305 drugs, demonstrating the scalability of the approach. When evaluated in downstream predictive tasks, these LLM-derived annotations consistently provided stronger performance than mappings based only on drug target names or manual Reactome assignments, underscoring their added value for capturing biologically meaningful drug–pathway relationships.

Validation via gene-importance enrichment. To assess whether the curated pathways were biologically meaningful, we examined the genes that contributed most strongly to prediction in the drug-specific "all-genes" XGBoost models, which were trained exclusively on baseline transcriptomic profiles and all genes. For each drug, we first identified the most influential genes by selecting those consistently ranked as important by both SHAP values and permutation importance, ensuring robustness against method-specific biases [39, 223]. These gene sets were then tested for enrichment within the curated pathways linked to the corresponding drug's mechanism of action. In the GDSC dataset, 114 drug-specific models showed at least one pathway with significant enrichment (FDR < 0.1), and in 65 of these cases (57%) the enriched pathways overlapped with the curated MOA assignments (Figure 6.5). In the larger PRISM dataset, the enrichment analysis repeatedly highlighted pathways such as "Cell Cycle, Mitotic," "Apoptosis," and "MAPK family signalling cascades," which are well established as central targets of many cancer drugs. Importantly, the fraction of recovered MOA-pathways was higher for the LLM-based annotations compared to target-name or ligand-based mappings, particularly when considering the subset of drugs with strong predictive performance ($\rho > 0.5$; Figure 6.6). These results demonstrate that the predictive features identified by the models are not only statistically significant but also align with known biological mechanisms, providing an independent validation of the LLM-based pathway annotations.

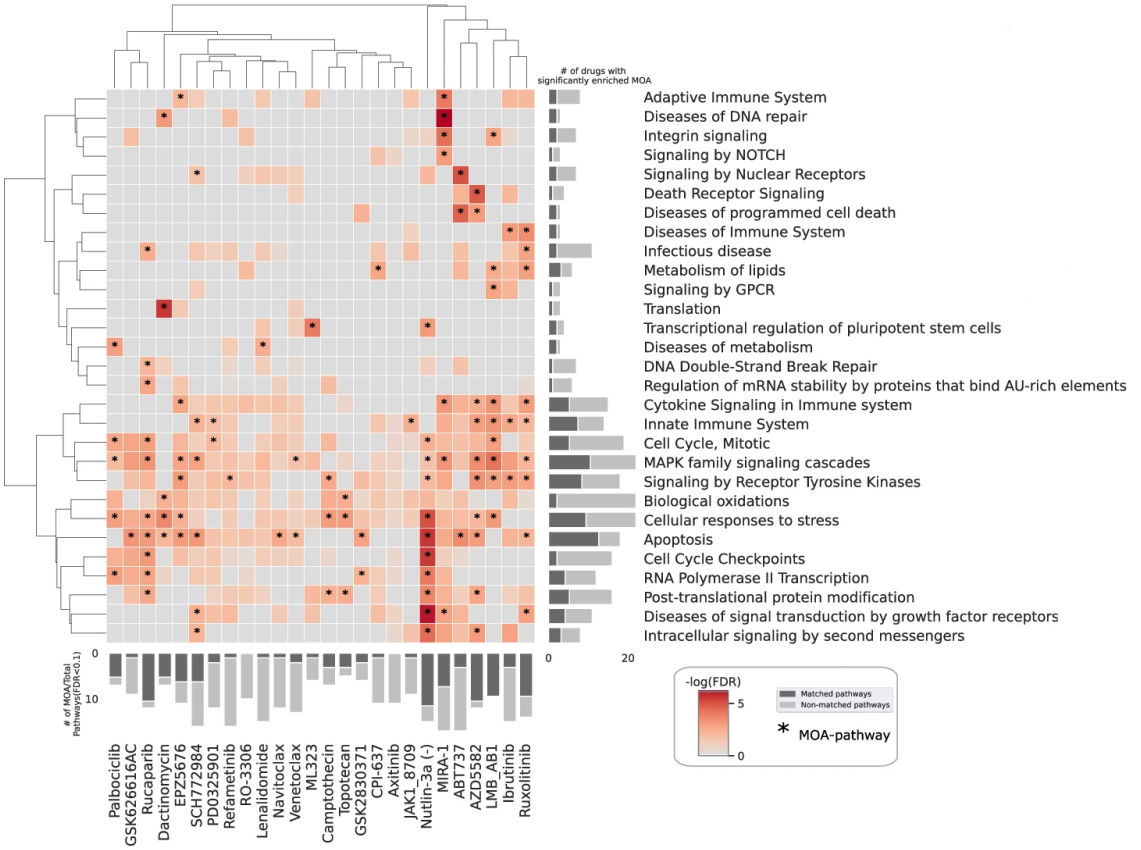

Figure 6.5: **Significant MOA-pathway enrichments across drug models.** Heatmap of significant MOA-pathways for various drug models, filtered by a correlation threshold $\rho > 0.5$. Pathways and drugs are shown along the y- and x-axes, respectively. Color intensity reflects enrichment significance ($-\log_{10}$(FDR)), with starred entries marking pathways linked to drugs via at least one annotation criterion. Adjacent bar plots indicate the number of significantly enriched pathways per drug (bottom) or per pathway (right), with dark gray segments highlighting curated MOA annotations.

6.3 Learning General Principles of Drug Sensitivity from Model Interpretations

A central aim of our analysis was to determine whether drug–cell line sensitivity models, trained solely on basal transcriptomics, could capture general biological principles beyond individual drug–target associations. A defining strength of *CellHit* lies in its emphasis on interpretability: by training separate XGBoost models for each drug, the framework makes it possible to directly assess the contribution of individual genes to predicted drug response. To quantify these contributions, we employed two complementary approaches. SHAP values provide a consistent estimate of the signed marginal effect of each gene on model predictions [223], while permutation importance measures the loss in predictive accuracy when the expression values of a gene are randomly shuffled, thereby breaking its association with response [39]. A gene was considered "important" only if it was identified as influential by both methods across multiple independent train/test splits, a criterion designed to minimise method-specific biases and reduce

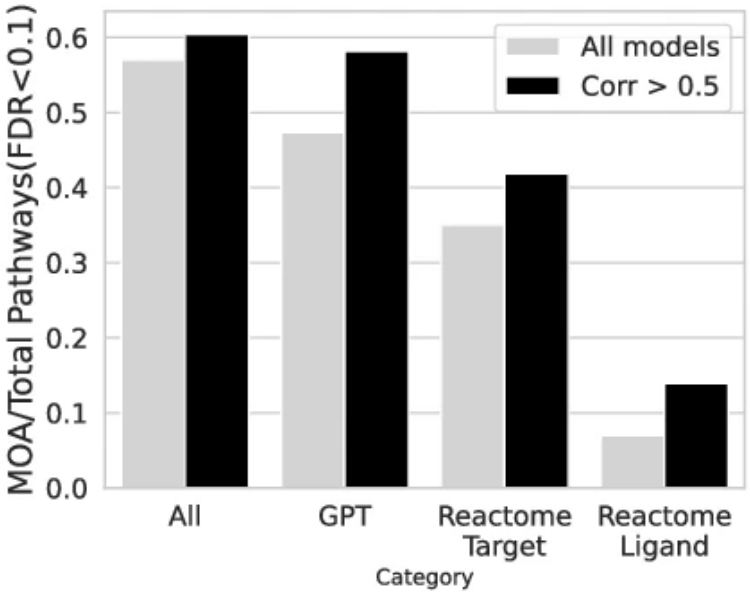

Figure 6.6: **Recovery of curated MOA-pathways under different annotation strategies.** Fraction of significantly enriched pathways (FDR < 0.1) matching drug MOAs under alternative annotation schemes: LLM-derived annotations (GPT), Reactome target-based, and Reactome ligand-based mappings. Bars show results for all models (light gray) and for the subset with predictive correlation $\rho > 0.5$ (black). LLM-derived annotations consistently outperform target- and ligand-based mappings.

noise. The resulting high-confidence gene sets not only clarified which features drove model performance but also provided biologically interpretable signals that could guide downstream analyses of drug mechanisms.

Mechanistic explainability Focusing on Venetoclax (the best-performing GDSC model) we observe a tight, mechanistically coherent linkage between signed local attributions, perturbation-based importance, and measured drug response. BCL2, Venetoclax's target, carries one of the strongest negative SHAP contributions to the predicted IC50 (teal), and shuffling its expression induces a marked deterioration in test correlation (orange "correlation-delta"), indicating high permutation sensitivity (Fig. 6.7). At the per–cell-line level, sorting samples by experimental IC50 reveals that higher BCL2 expression co-occurs with lower IC50 and more negative SHAP values, showing that the model explicitly exploits BCL2 dependence to predict sensitivity (Fig. 6.8). These consistent patterns across attribution signs and observed responses create a clear mechanistic narrative, reflecting the expected behavior of a BCL2 inhibitor. This illustrates that the model's explanations are based on target-driven pharmacology rather than misleading correlations.

Recovery of drug targets and mechanism-of-action signals. To evaluate how faithfully the models reflected underlying biology, we examined whether the features identified as most predictive corresponded to the drugs' nominal targets or to pathways implicated in their MOA (Fig. 6.9). In the GDSC dataset, 39% of the drug-specific models

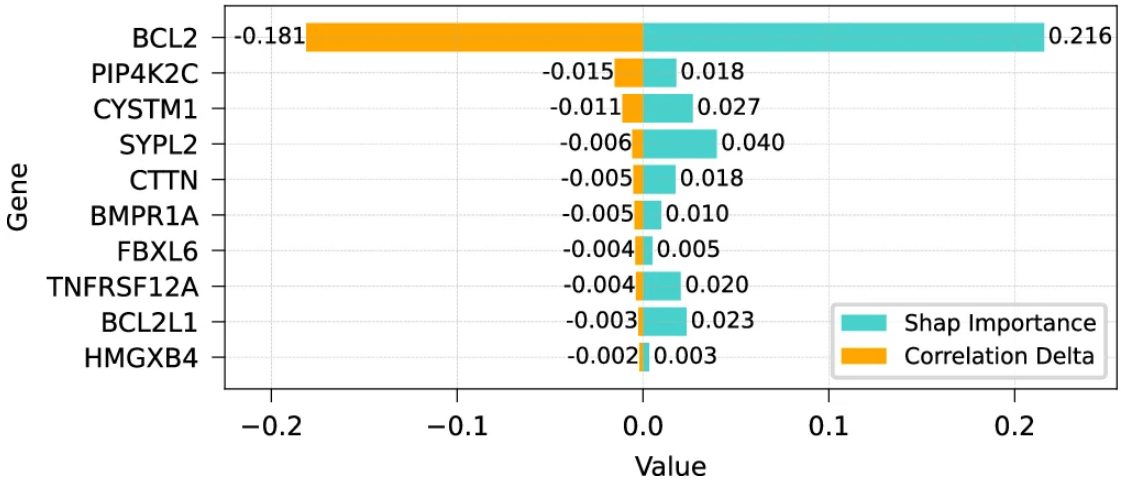

Figure 6.7: **Feature importance analysis for Venetoclax.** SHAP (teal) and correlation delta (orange) importances for the Venetoclax drug. Permutation importance reflects the decrease in the model's prediction accuracy when a feature's values are shuffled, indicating its importance (greater drops signify higher importance). SHAP importance represents a feature's contribution to the model's prediction, with larger absolute values indicating greater importance.

highlighted the annotated target gene as important in at least one of 20 independent train/test splits, while 70% of targets were ranked at or above the 90th percentile relative to a random gene background distribution. Similarly, in the PRISM dataset, 62% of models with annotated targets successfully recovered them in at least one split, and 73.7% of targets reached the 90th percentile threshold. At the level of individual compounds, several cases showed consistent recovery of the expected target across all splits. For instance, the NAD^+ biosynthesis inhibitor STF-31 invariably identified its target *NAMPT*, while the MDM2 inhibitor CGM097 consistently recovered *MDM2*. Comparable patterns were observed in drug families: BCL2 inhibitors such as Venetoclax, Navitoclax, and ABT-737 robustly highlighted *BCL2* as a top-ranked negative SHAP contributor, in line with their known pro-apoptotic mechanism. These findings indicate that model-derived gene importance profiles not only capture statistically enriched pathway signals but also reliably point back to the molecular targets and biological processes underlying drug sensitivity.

Computation of the Baseline Recovery Distribution. To estimate how often a drug's target gene could be detected by chance, we computed a separate baseline for each drug. For every model and training–testing split, we converted SHAP and permutation importance scores into binary indicators, assigning a value of 1 when a gene's score was greater than zero and 0 otherwise. We then took the intersection of these two sets to retain only genes consistently deemed important by both methods. For each gene, we calculated how frequently it was recovered across 20 independent splits, and used these values to build an empirical background distribution of "random recovery" rates. The 90th and 95th quantiles of this distribution were used as significance thresholds: a nominal target was considered robustly recovered for a given drug if its observed

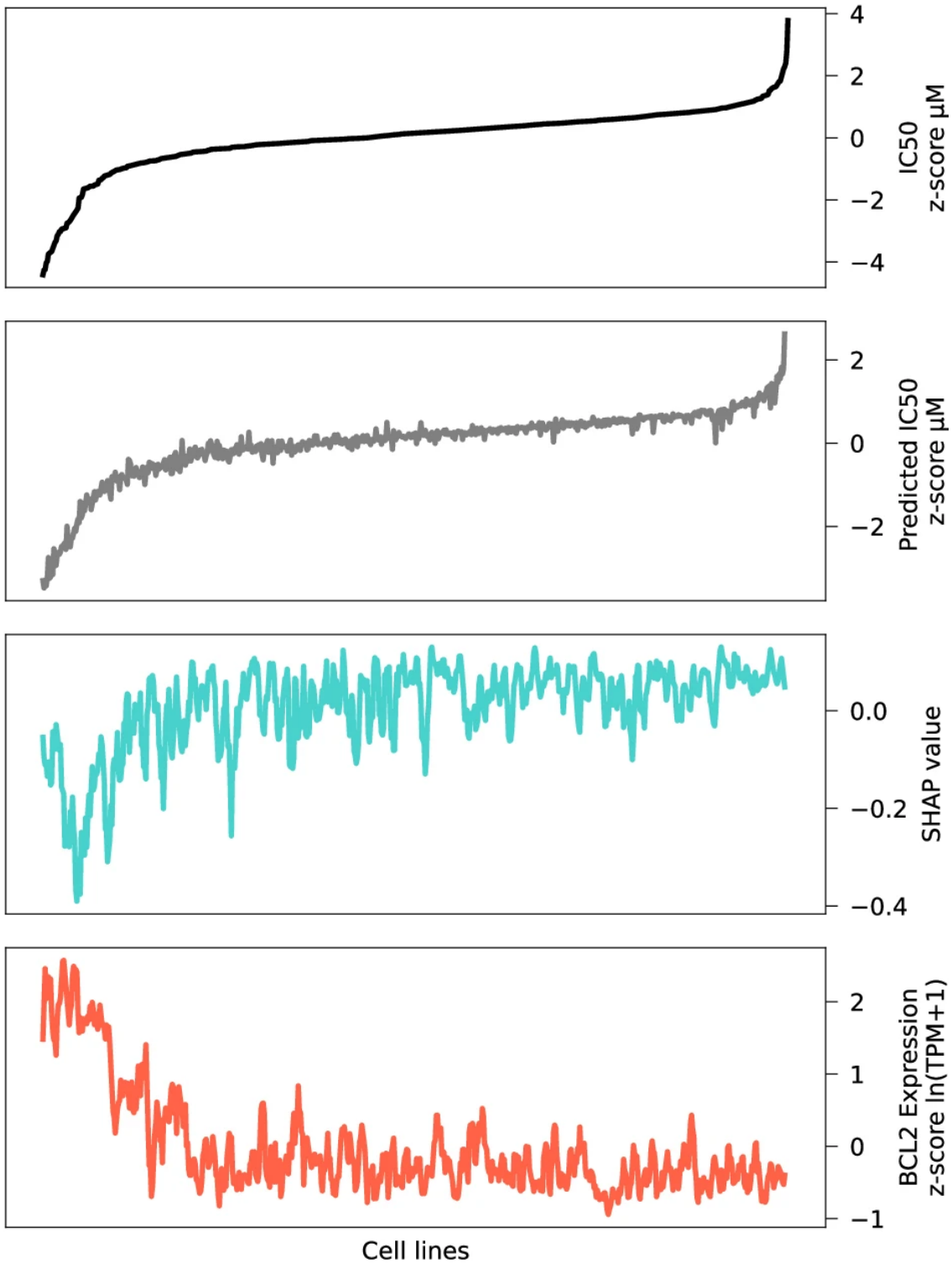

Figure 6.8: **Per–cell-line assessment of the Venetoclax model.** The top plot (black) shows experimental IC50 z-scores, while the second plot (gray) depicts predicted IC50 values. The third plot (teal) shows SHAP values for BCL2, and the fourth plot (red) displays BCL2 expression levels. Together, the plots demonstrate that lower IC50 values (greater sensitivity) are associated with higher BCL2 expression and more negative SHAP values, consistent with the expected mechanism of action of Venetoclax.

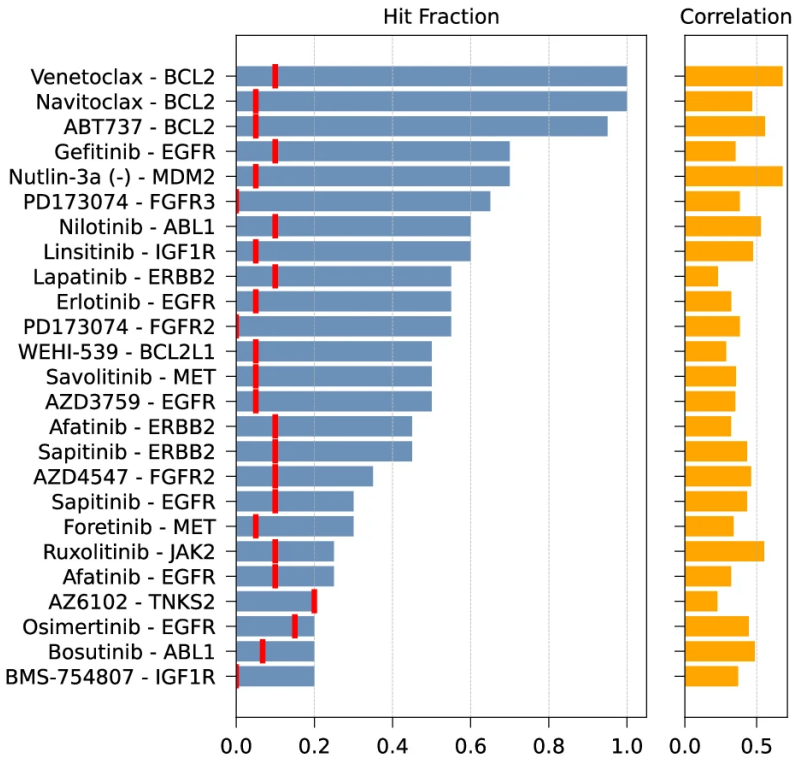

Figure 6.9: **Target recovery for the top 25 ligand–target pairs.** Left: for each drug, the bar length shows the *Hit Fraction*, i.e., the fraction of 20 tissue-stratified train/test splits in which the drug-specific model identified the putative target gene as important. Red tick marks indicate the 95th percentile of the Hit Fraction distribution across all genes for that drug (null threshold). Right: the bar length shows the median Pearson correlation between predicted and observed responses across the same splits.

recovery frequency exceeded these thresholds.

Tissue-Specific Essentiality as a Learned Principle. To assess whether the models also learned general cellular dependencies, we evaluated whether important genes also recovered information about gene dependencies of cancer cell lines from different tissues. To this end, we retained drug-cell line instances yielding the most significant predictions, ranked the top k most important genes based on SHAP values, and pooled them on the basis of the tissue of origin of the cell lines. We then evaluated the recall of the top k important genes to identify core essential genes from an updated dependency map across 27 cancer tissues [265]. Remarkably, when aggregating the top 100 genes by SHAP importance, we identified core essential genes with a recall greater than 0.9 in several tissues (Fig. 6.10).

Network Connectivity of Essential Genes. To further characterise the biological relevance of the predictive features, we examined their organisation within protein–protein interaction networks. Essential genes identified as prediction-important tended to occupy central positions, forming highly connected hubs. In lung cancer models, for instance, genes such as *BCL2L1* (Bcl2-like 1) and *YAP1* (Yes1 Associated Transcriptional

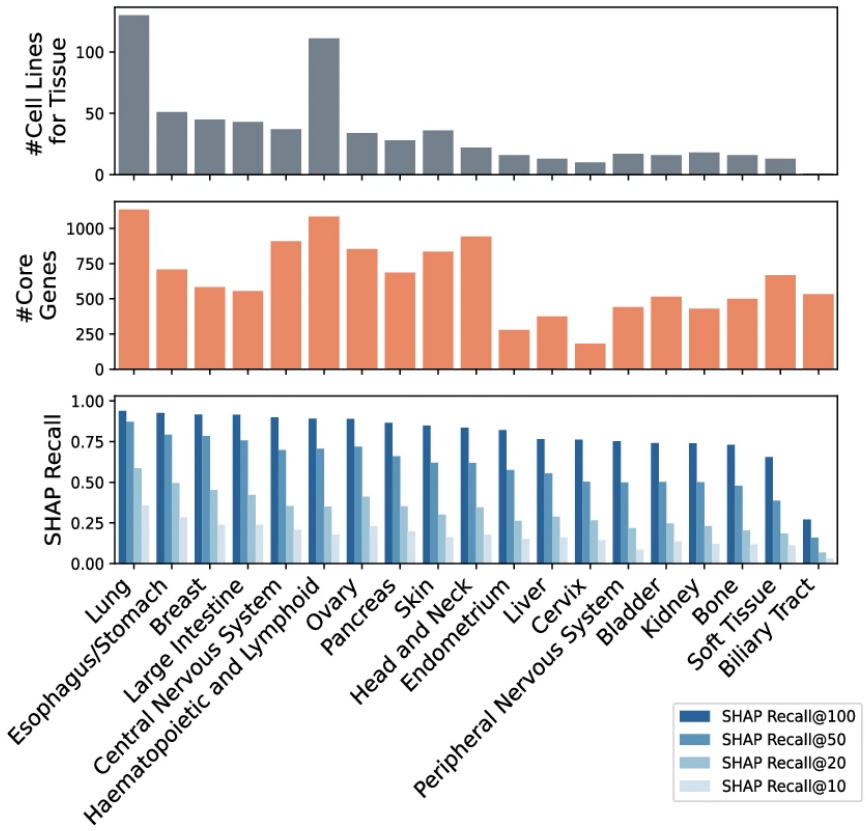

Figure 6.10: **Tissue-specific recovery of essential genes from predictive features.** (Top) Number of cell lines available per tissue. (Middle) Number of core essential genes identified in dependency maps. (Bottom) Recall of essential genes among the top k most important genes (SHAP-ranked) across tissues, evaluated at thresholds of $k = 10, 20, 50, 100$.

Regulator) emerged as the top contributors to predictive performance when ranked by their average SHAP importance across drug models. Both genes not only displayed high predictive relevance but also exhibited elevated network connectivity in the STRING protein–protein interaction network, highlighting their central role in cellular processes (Fig. 6.11; Supplementary Data 6). This convergence between network centrality and model importance indicates that drug sensitivity is captured not only through drug-specific mechanism-of-action signals but also through the perturbation of core vulnerabilities essential for cell survival and proliferation.

6.4 Scaling explainable drug sensitivity prediction to the PRISM dataset

Scaling to PRISM. We applied the same per-drug modelling strategy to the much larger *PRISM* dataset, which includes more than six thousand compounds tested across nearly nine hundred cancer cell lines. In total, we trained 6,337 drug-specific models, a computationally demanding task that required high-performance GPUs and extensive parallelisation to complete efficiently. From this large collection, 762 models achieved

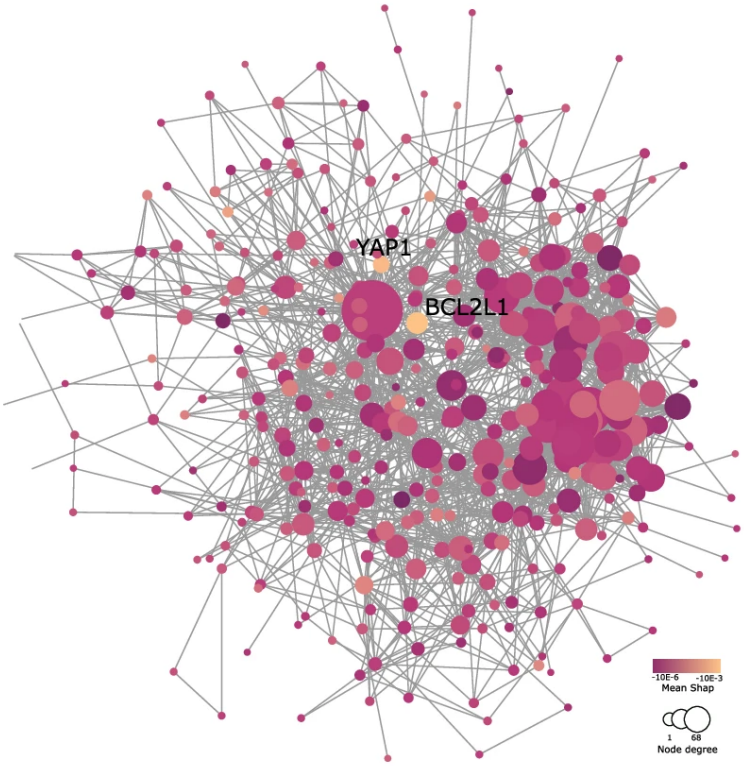

Figure 6.11: **Network connectivity of essential, prediction-important genes in lung cancer.** STRING protein–protein interaction (PPI) network of lung core essential genes recovered by SHAP importances. Node diameters are proportional to network degree, while node colors reflect average SHAP values across drug models (brighter colors denote higher importance). Notably, *BCL2L1* and *YAP1* emerge as highly connected hubs with strong predictive importance.

a Pearson correlation above $\rho = 0.2$, a performance threshold commonly used in prior PRISM studies to indicate predictive value [70]. Among these successful models, kinase inhibitors were the most frequently represented, consistent with their well-established role in targeted cancer therapies. Other enzymatic classes, such as additional kinases and epigenetic regulators, also appeared prominently, while G protein-coupled receptor (GPCR) modulators contributed to the set of reliably predictable drugs. Across all retained models, the overall predictive performance reached a median correlation of $\rho = 0.80$ with a MSE of 1.18 (Fig. 6.12). Importantly, these results were obtained despite the intrinsic limitations of the PRISM dataset, which is based on single-dose viability assays and contains many compounds with only weak activity. The ability of the framework to recover known drug–target associations under such conditions highlights both its robustness and its scalability to very large pharmacogenomic screens.

Variability of drug response and model performance. In the PRISM dataset, we found that many compounds show very little variation in their log-fold change (LFC) values across cell lines. This limited variability makes accurate prediction inherently difficult, since the models have fewer differences in response to learn from. When we

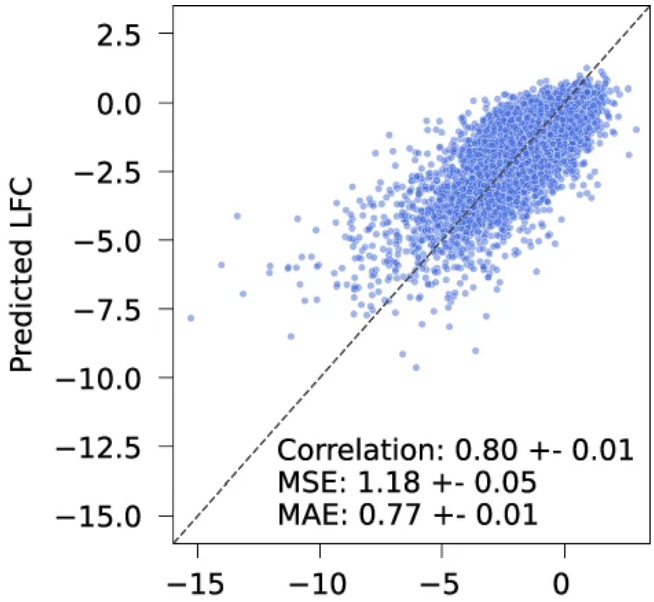

Figure 6.12: **Predictive performance of PRISM drug-specific models.** Scatter plot of predicted versus experimental log-fold change (LFC) values from models surpassing the correlation threshold of $\rho \geq 0.2$. Shown are representative predictions with performance metrics (Pearson correlation, mean squared error, and mean absolute error) annotated in the panel. The diagonal line indicates perfect agreement between predicted and observed responses.

compared compounds with broader response ranges to those with narrower ones, the effect was clear: drugs with an interquartile range (IQR) in LFC greater than 1 achieved a much higher median predictive performance (Pearson $\rho \approx 0.24$), whereas the median correlation across all compounds was only $\rho \approx 0.04$. This indicates that models perform better when the underlying drug responses span a wider dynamic range, because transcriptional features can more effectively separate sensitive from resistant cell lines. Consistently, the joint distribution of performance metrics (Pearson correlation vs. MSE) shows that higher-IQR compounds (warmer colors) concentrate in the region of higher correlations and lower errors, while lower-IQR compounds (cooler colors) cluster toward poorer performance; marginal density plots further highlight the shift in both metrics for the IQR> 1 subset (Fig. 6.13). For this reason, we restricted subsequent analyses to models with a correlation above $\rho > 0.2$, ensuring that we focused on predictions supported by meaningful biological signal rather than noise.

Drug classes with strongest predictive signal. Kinase inhibitors stood out as the drug class with the largest number of models that achieved strong predictive performance, both in raw counts and when adjusted for the number of compounds within each target class. These drugs were also those for which the models most often recovered the expected targets among the top-ranked genes identified by SHAP and permutation importance. This pattern is consistent with the fact that kinase inhibitors typically act through well-defined mechanisms of action that are reflected in transcriptomic changes, making them easier for the models to detect. At the same time, strong performance was not limited to kinase inhibitors: we also observed recurrently accurate models for other

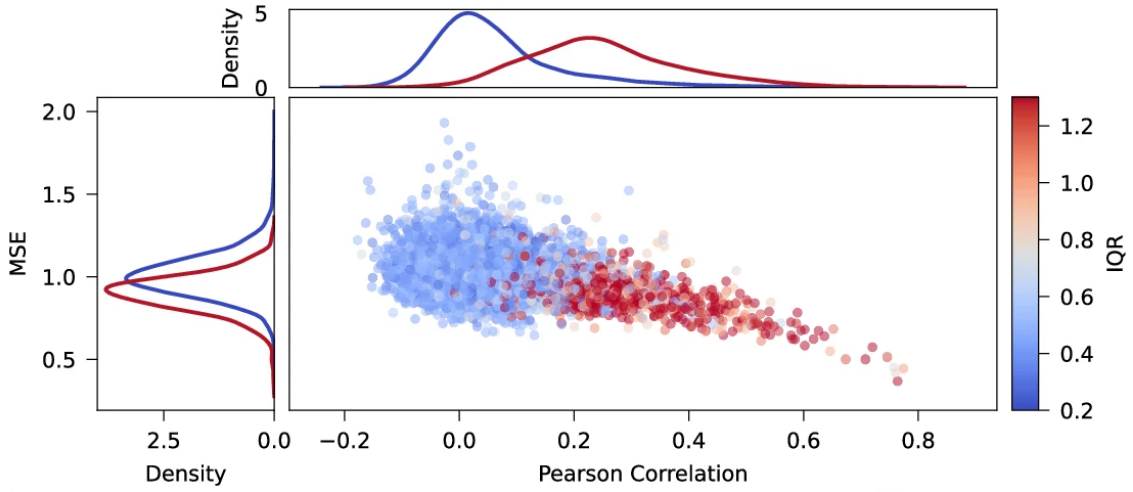

Figure 6.13: **Response variability governs predictive performance in PRISM.** Scatter plot of drug-specific models showing Pearson correlation (x-axis) versus mean squared error (MSE; y-axis). Each point represents a PRISM compound–specific model, colored by the interquartile range (IQR) of its LFC profile across cell lines (blue to red indicates increasing IQR). The top and left marginal density plots compare the distributions of correlation and MSE, respectively, between all models (blue) and the subset with IQR>1 (red), illustrating that higher response variability is associated with higher correlations and lower errors.

target categories, such as broad-spectrum enzymes, epigenetic regulators, and GPCR ligands. Consistently, stratifying PRISM models by putative target family shows the highest counts for kinases, followed by enzymes and epigenetic regulators; the subset with target recovery mirrors this ranking (salmon vs. red bars in Fig. 6.14).

Significance of scaling to PRISM. This large-scale deployment represents, to our knowledge, the first interpretable ML modelling effort spanning the full PRISM compound library. The ability to recover known targets and MOA-consistent pathways across thousands of structurally and pharmacologically heterogeneous compounds highlights the generalisability of the framework. Moreover, the broader MOA diversity uncovered in PRISM underscores the potential of such large-scale, explainable models to inform drug repurposing, polypharmacology mapping, and novel hypothesis generation in oncology and beyond. Consistent with this, the set of significantly enriched MOA–pathways recovered from PRISM models substantially extends that observed in GDSC, with 314 pathways shared, 465 PRISM-specific, and 133 GDSC-specific (Fig. 6.15).

6.5 Knowledge-driven "MOA-primed" models

From all-genes to MOA-driven feature selection. Although the models trained on the full set of available genes (*all-genes models*) already achieved good predictive accuracy, we reasoned that narrowing the input features to those genes directly involved in a drug's MOA could yield further improvements in both performance and interpretability. To implement this idea, we used the LLM-curated MOA–pathway associations generated

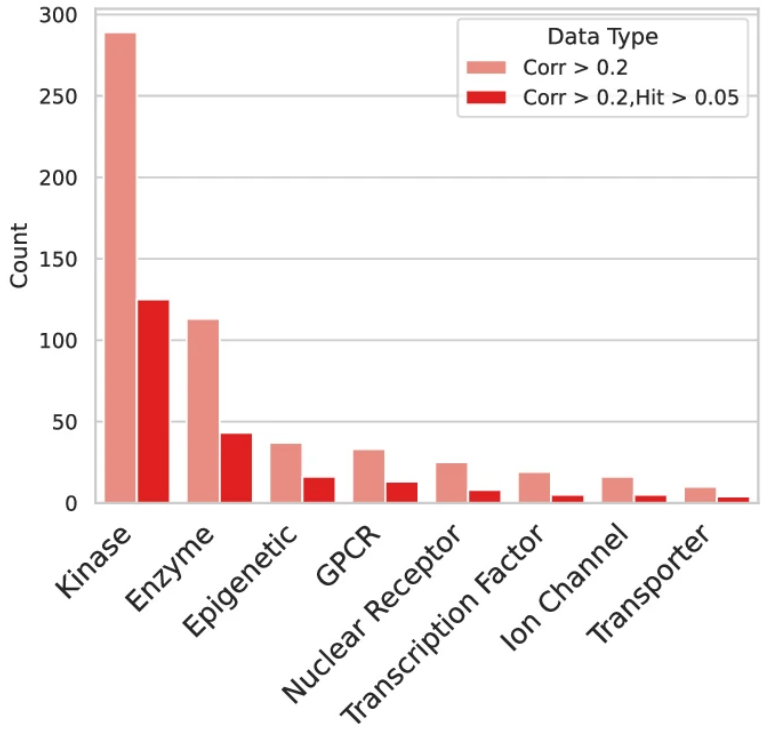

Figure 6.14: **Drug classes with strongest predictive signal in PRISM.** Bar plot of drug-specific models stratified by putative target protein families. Salmon bars show the number of models with predictive performance $\rho > 0.2$; red bars show the subset that both achieved $\rho > 0.2$ and recovered the annotated target among top-ranked genes by SHAP and permutation importance. Kinase inhibitors dominate in both counts, followed by enzymes and epigenetic regulators, with additional contributions from GPCRs, nuclear receptors, transcription factors, ion channels, and transporters.

by our Mixtral-based pipeline (see Chapter 6.2), which systematically links each compound to the Reactome pathways most likely to influence its activity. For every drug, we restricted the model inputs to the genes belonging to its curated MOA–pathways. This focused approach substantially reduced the dimensionality of the feature space, lowering the average number of predictors from 18,174 to approximately 4,117 per model—a reduction of about 4.4-fold—thereby allowing more thorough hyperparameter optimization and decreasing the risk of overfitting.

Model architecture and training. The MOA-primed models were built as ensembles of three separate XGBoost regressors, each one trained on a different partition of the data in a three-fold cross-validation setting. Within every regressor, learning was carried out using five boosted decision trees in parallel. The hyperparameters that control the behaviour of these trees were not chosen arbitrarily but were optimised through a multi-objective procedure that balanced two key criteria: the correlation between predicted and observed responses, and the mean squared error. By combining several XGBoost models in this way, the framework maintained the transparency and feature interpretability of tree-based approaches, while simultaneously taking advantage of the reduced feature space to achieve higher statistical efficiency.

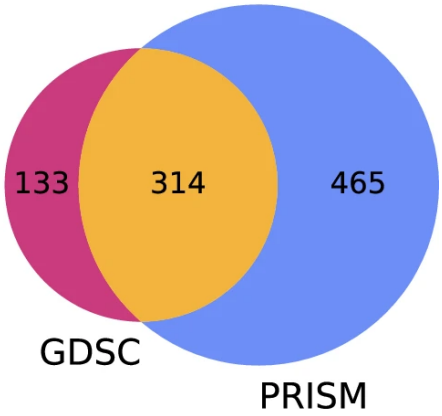

Figure 6.15: **Overlap of enriched MOA–pathways in PRISM and GDSC.** Venn diagram comparing the sets of significantly enriched MOA-pathways identified from drug-specific models in PRISM and GDSC, based on genes deemed important by the models. Numbers indicate pathway counts: 314 shared, 465 PRISM-specific, and 133 GDSC-specific. The larger PRISM-only segment reflects the broader MOA coverage achieved when scaling to the full PRISM library.

Performance gains on GDSC. When applied to the GDSC dataset, models trained with MOA-guided gene subsets substantially outperformed models that used all available genes (Fig. 6.16A–C). Across the 286 evaluated drugs, the median Pearson correlation between predicted and observed responses on held-out test sets rose from approximately $\rho = 0.40$ to $\rho = 0.50$ (Fig. 6.16A). When aggregating predictions across all drug–cell line pairs, these MOA-primed models reached a correlation of $\rho = 0.89$, accompanied by the lowest mean squared error ($\mathrm{MSE} = 1.52$) among all tested model configurations (Fig. 6.16B). The magnitude of improvement varied across drugs: for example, the androgen receptor antagonist Bicalutamide showed nearly a twofold increase in predictive correlation when restricted to its annotated MOA-related genes, whereas a small subset of compounds (e.g., BX795, Gemcitabine, Savolitinib) experienced slight performance decreases, likely reflecting incomplete or overly narrow mechanism annotations (Fig. 6.16C).

Performance gains on PRISM. The PRISM dataset was more challenging to model because it contains a wider variety of compounds and many drugs that show little variation in their activity across cell lines. To focus on drugs where meaningful predictions were possible, we limited the analysis to those with greater variability in response (interquartile range above 1). Within this subset, the models trained on MOA-informed features consistently outperformed those trained on all genes. As shown in Fig. 6.17**A**, the distribution of per-drug correlations shifts to higher values for MOA-primed models; the typical correlation between predicted and observed responses increased from about $\rho = 0.24$ to $\rho = 0.32$. At the aggregate level, predictions closely track measured responses (Fig. 6.17**B**), yielding an overall correlation of $\rho = 0.93$. This filtering nearly doubled the number of PRISM drug models that surpassed the utility threshold of $\rho > 0.2$, increasing from 762 to 1,254. Gains were particularly evident for drugs with well-defined

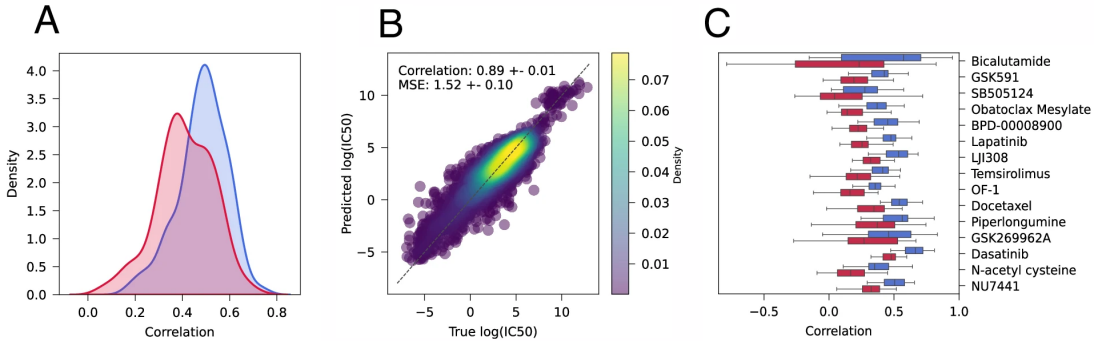

Figure 6.16: **MOA-primed models improve drug response prediction in GDSC. (A)** Distribution of Pearson correlations for all-genes (red) versus MOA-primed (blue) models across 286 drugs, showing a rightward shift with MOA guidance. **(B)** Predicted versus experimental log(IC_{50}) for MOA-primed models; point color encodes local point density. The pooled correlation is $\rho \approx 0.89$ with $\mathrm{MSE} \approx 1.52$. **(C)** Per-drug boxplots highlighting compounds with the largest correlation gains under MOA-priming (blue) compared with all-genes baselines (red); each box shows median, interquartile range, and whiskers for variability across splits.

targets, such as the neurokinin-1 receptor antagonist Rolapitant (Fig. 6.17**C**), where the mechanism of action is well captured by the selected gene sets.

6.6 Patient-level inference at scale on TCGA

After establishing and validating the models on preclinical cell line datasets, we extended the pipeline to patient samples by applying it to transcriptomic profiles from The Cancer Genome Atlas (TCGA, $n \approx 10{,}000$). Because these patient samples were generated with bulk RNA-seq, we first used *Celligner* [380] to harmonize them with the cell line expression space. This step ensured that patient and cell line data became directly comparable, minimizing systematic differences in measurement platforms and biological composition. Once aligned, we applied the drug-specific models trained on cell lines to the TCGA data, enabling us to generate predicted inhibitory concentrations (IC_{50}) for every compound in our model collection across thousands of patient tumors.

Validation against approved indications. We next applied our drug–response models to TCGA in order to predict which approved therapies might be effective for individual patient tumors based on their transcriptomic profiles. To establish a reference, we compiled a list of FDA–approved drugs and linked each compound to its corresponding cancer type as annotated in the National Cancer Institute database. This yielded 41 drugs from the GDSC library with approvals spanning 23 distinct cancer types. For every approved drug, we generated predictions across all TCGA samples and ranked the top 600 predicted responders using two complementary criteria (see Figure 6.18). The first criterion was the predicted log(IC_{50}), which reflects absolute sensitivity. The second was the quantile score, a metric designed to balance potency with selectivity by evaluating how strongly a drug is predicted to act on a given sample compared to all other samples. We

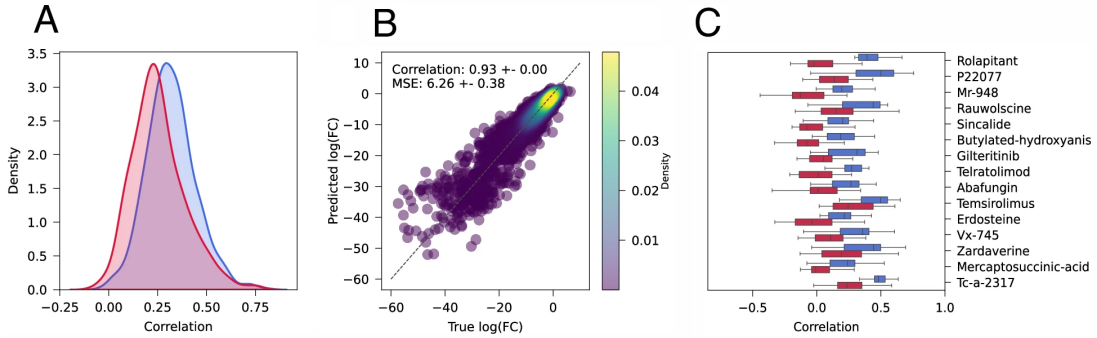

Figure 6.17: **MOA-primed models improve drug response prediction in PRISM.** **(A)** Distribution of Pearson correlations for all-genes (red) versus MOA-primed (blue) models across variable-response drugs (IQR $>$ 1), showing a rightward shift with MOA guidance. **(B)** Predicted versus experimental log fold-change (logFC) for MOA-primed models; point color encodes local point density. The pooled correlation is $\rho \approx 0.93$ with MSE $\approx$ 6.26. **(C)** Per-drug boxplots highlighting compounds with the largest correlation gains under MOA-priming (blue) compared with all-genes baselines (red); each box shows median, interquartile range, and whiskers for variability across splits. Notably, Rolapitant exhibits a pronounced improvement.

chose the cutoff of 600 top-ranked samples because it maximized performance in a binary classification task, where the goal was to recover patient tumors from cancer types for which the drug is clinically approved. Both ranking strategies effectively prioritized patients from the correct indications (see Figure 6.19). For several drugs—including Cytarabine, Venetoclax, and 5-azacytidine—the models achieved particularly high recall for the relevant cancer types. More broadly, 37 out of the 41 approved drugs (90%) yielded models that ranked at least one patient from their corresponding indication among the top 600 predictions (see Figure 6.20). In many cases, the majority of highly ranked samples came from the target cancer type, such as Fulvestrant in breast cancer (BRCA), BCL2 inhibitors and Cyclophosphamide in breast cancer or acute myeloid leukemia (LAML), and the BRAF inhibitor Dabrafenib and MEK1/2 inhibitor Trametinib in skin cutaneous melanoma (SKCM) (see Figure 6.20).Dabrafenib offered an instructive example of the model's interpretability. The drug is only effective against tumors carrying the BRAF V600E mutation, a dependency determined by mutation status rather than absolute expression level. Accordingly, the model did not highlight *BRAF* itself or its associated pathways as important features. Instead, among the top 600 ranked TCGA samples, the model consistently prioritized tumors harboring BRAF mutations (see Supplementary Figure. 6.21B). While most of these were melanomas, the model also surfaced BRAF-mutant tumors from other cancer types such as thyroid carcinoma (THCA) and diffuse large B-cell lymphoma (DLBC) (see Supplementary Figure. 6.21B). This suggests that the framework can detect mutation-driven transcriptional signatures and may help identify repurposing opportunities for targeted therapies in less common molecular contexts [339].

Prediction of combination therapies. Across the 33 cancer types represented in TCGA, our models identified 10,500 patient samples that appeared within the top 600

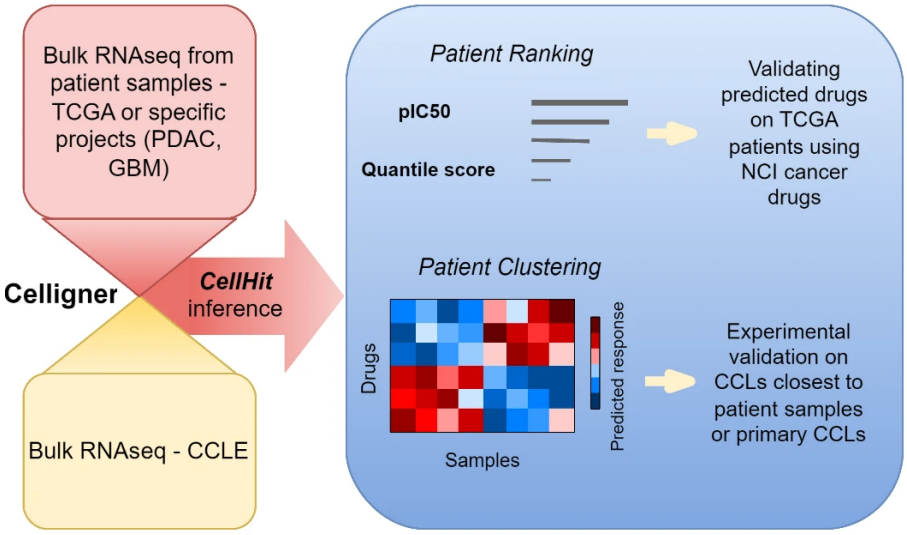

Figure 6.18: **Workflow for drug response prediction on TCGA.** Bulk RNA sequencing data from TCGA patients, as well as from PDAC and GBM cohorts, are harmonized using *Celligner* and processed through *CellHit* to infer drug responses. Patients are ranked by predicted log(IC_{50}) and quantile score, and clustered by response profiles. Validation involves comparison with NCI drug approvals and experimental testing on cell lines closest to patient tumors.

predicted responders for more than one drug, indicating that these tumors might benefit from combination treatments (see Figure 6.22). To explore this systematically, we ranked all predicted drug–drug pairs based on the number of overlapping samples they jointly prioritized. Many of the highest-scoring pairs corresponded to combinations that are already clinically approved, including Trametinib with Dabrafenib in melanoma (SKCM), Venetoclax with either Cytarabine or 5-azacytidine in acute myeloid leukemia (LAML), Fulvestrant with CDK4/6 inhibitors such as AZD5363, Alpelisib, or Palbociclib in breast cancer (BRCA), and Oxaliplatin with 5-Fluorouracil in colon adenocarcinoma (COAD). Beyond these established regimens, the analysis also revealed additional drug pairs predicted to be effective in overlapping patient groups for the same indication, pointing to new opportunities for rational design of combination therapies (see Figure 6.22).

Extension to non-oncology drugs. We next used the PRISM-trained models to investigate whether non-oncological drugs could be repurposed for cancer treatment by applying them to TCGA tumors. For each compound, we ranked all patient samples by their predicted sensitivity and selected the top 600 most responsive cases. To ensure that these predictions were reliable, we restricted the analysis to the 20 non-oncological drugs with the strongest model performance, which included eight enzyme inhibitors and six ligands of G-protein–coupled receptors (GPCRs) (see Figure 6.23). This analysis revealed distinct patterns of sensitivity that were specific to particular cancer types. A notable example was provided by the adenosine receptor antagonists CGS-15943 and MRS-1220, which have recently been suggested as potential treatments in several cancers [16, 71]. Both compounds were consistently predicted to be effective in subsets of breast (BRCA), liver (LIHC), prostate (PRAD), and gastric (STAD) tumors (see Figure

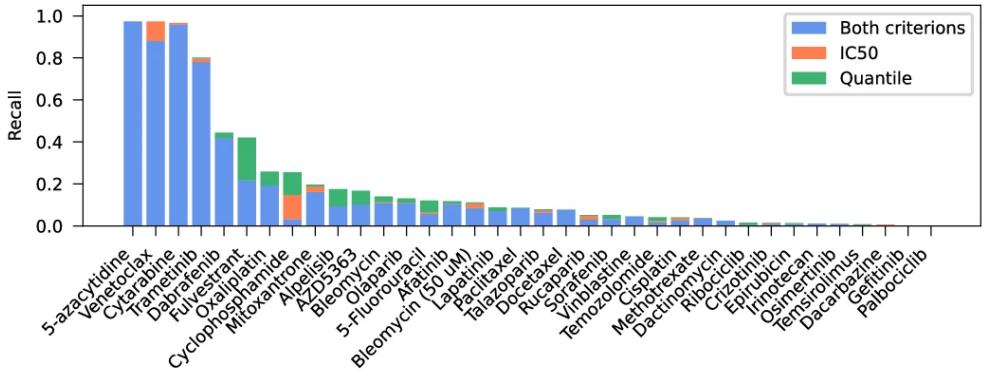

Figure 6.19: **Recovery of approved drug indications in TCGA.** Recall of FDA–approved drug indications across TCGA tumors among the top 600 predicted responders. Each bar shows the fraction of tumors from the approved cancer type correctly recovered for a given drug, using either the predicted log(IC_{50}) (orange), the quantile score (green), or both criteria (blue). Drugs such as 5-azacytidine, Venetoclax, and Cytarabine reached near-complete recall for their approved indications.

6.23).

6.7 Prospective wet-lab validation of model predictions

To assess the translational potential of the CellHit framework, we conducted prospective experimental validations on two highly lethal and therapeutically challenging solid tumours: pancreatic ductal adenocarcinoma (PDAC) and glioblastoma multiforme (GBM). In both cases, predictions generated from patient-derived transcriptomic profiles were tested in wet-lab assays, enabling a direct evaluation of the model's capacity to identify subtype- or sample-specific drug sensitivities.

Subtype-specific drug predictions in PDAC. We applied CellHit, trained on GDSC data, to PDAC transcriptomic profiles stratified into recently defined morphological–molecular subtypes: Glandular (GL), Transitional (TR), and Undifferentiated (UN). The analysis revealed distinct predicted sensitivity patterns between subtypes, with hierarchical clustering of predicted IC_{50} values segregating GL and TR samples into distinct response groups (Figure 6.24). Importantly, the model identified two clinically approved topoisomerase inhibitors, Irinotecan and Teniposide (and its analogous Etoposide), as showing higher predicted efficacy in GL compared to TR samples (Figure 6.25A). To experimentally validate these predictions, we selected PDAC cell lines transcriptionally closest to each subtype (CFPAC-1 for GL-like, PANC-1 for TR-like) and measured viability after drug treatment at clinically relevant concentrations (Figure 6.25B). Both irinotecan and etoposide demonstrated markedly greater cytotoxicity in CFPAC-1 than in PANC-1, confirming the subtype-specific sensitivity predicted by the model.

We next applied CellHit to primary cultures derived from glioblastoma (GBM) patient tumors, testing its ability to generalize beyond established cell lines. Using transcriptomic profiles from 64 GBM samples, the model highlighted two representative cases

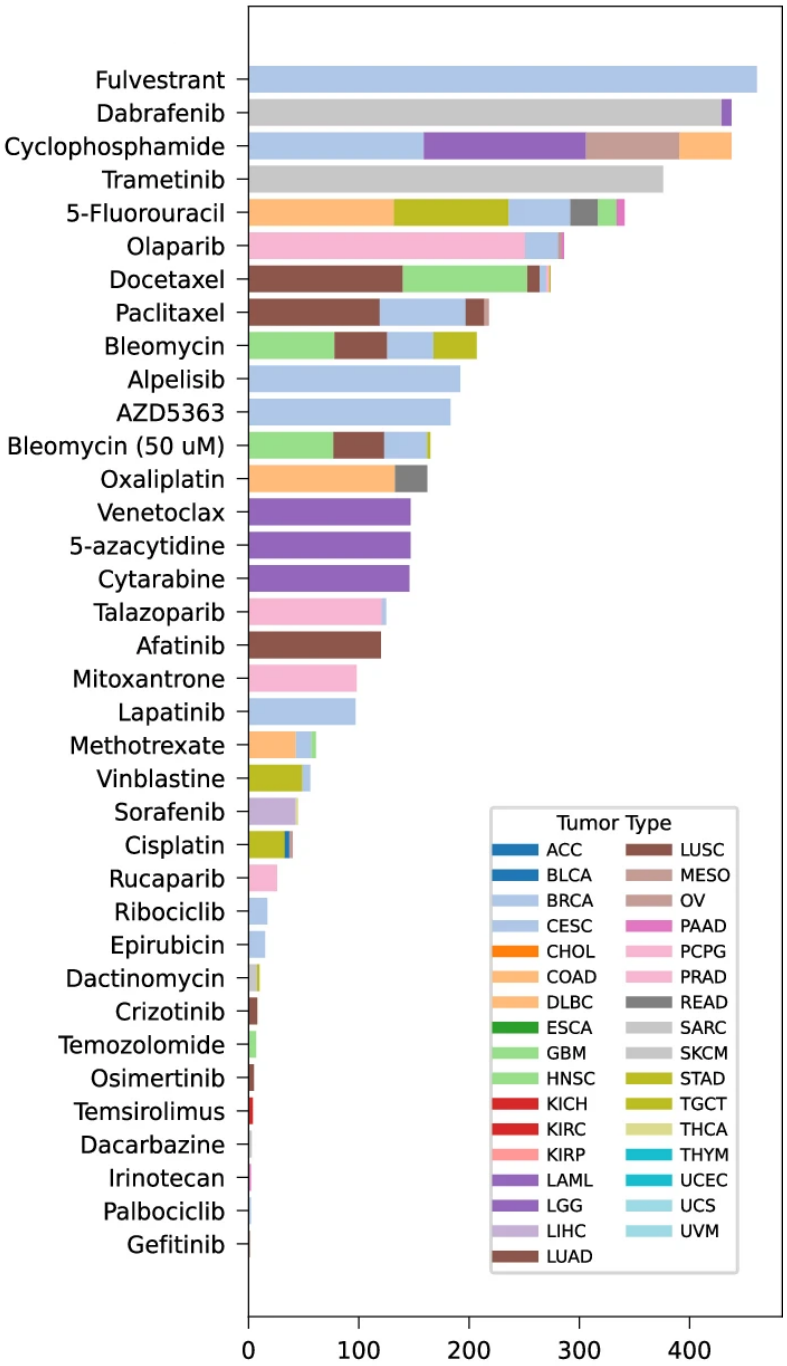

Figure 6.20: **Recovery of approved indications in TCGA predictions.** Barplot showing the distribution of the top 600 predicted responders per drug, stratified by tumor type, for FDA–approved compounds present in the GDSC library. Each bar corresponds to one drug, with colors denoting tumor types according to TCGA abbreviations. For most drugs, samples from the cancer type of approval are strongly enriched among the highest-ranked predictions, exemplified by Fulvestrant in breast cancer (BRCA), Venetoclax and Cytarabine in acute myeloid leukemia (LAML), Cyclophosphamide in BRCA and LAML, and Dabrafenib/Trametinib in skin cutaneous melanoma (SKCM). Overall, 37 of 41 drugs (90%) successfully retrieved patients from their approved indications among the top-ranked predictions.

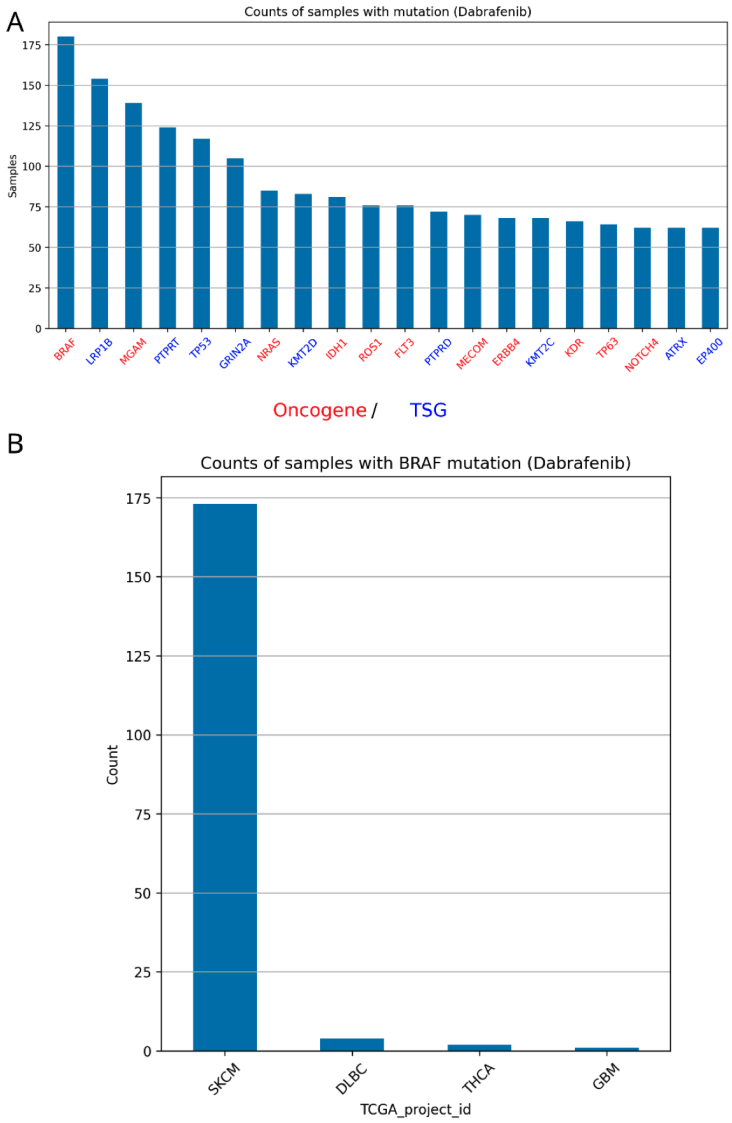

Figure 6.21: **Mutational burden of top-ranked TCGA patients for Dabrafenib predictions.** (**A**) Distribution of mutation counts across known oncogenic drivers in the top 600 TCGA patient samples prioritized by the Dabrafenib model. (**B**) Number of BRAF-mutant samples identified among the top 600 ranked patients, stratified by cancer type. In line with the drug's known mechanism of action, the model preferentially selected tumors carrying the *BRAF* V600E mutation, predominantly in melanoma (SKCM), but also in thyroid carcinoma (THCA) and diffuse large B-cell lymphoma (DLBC)

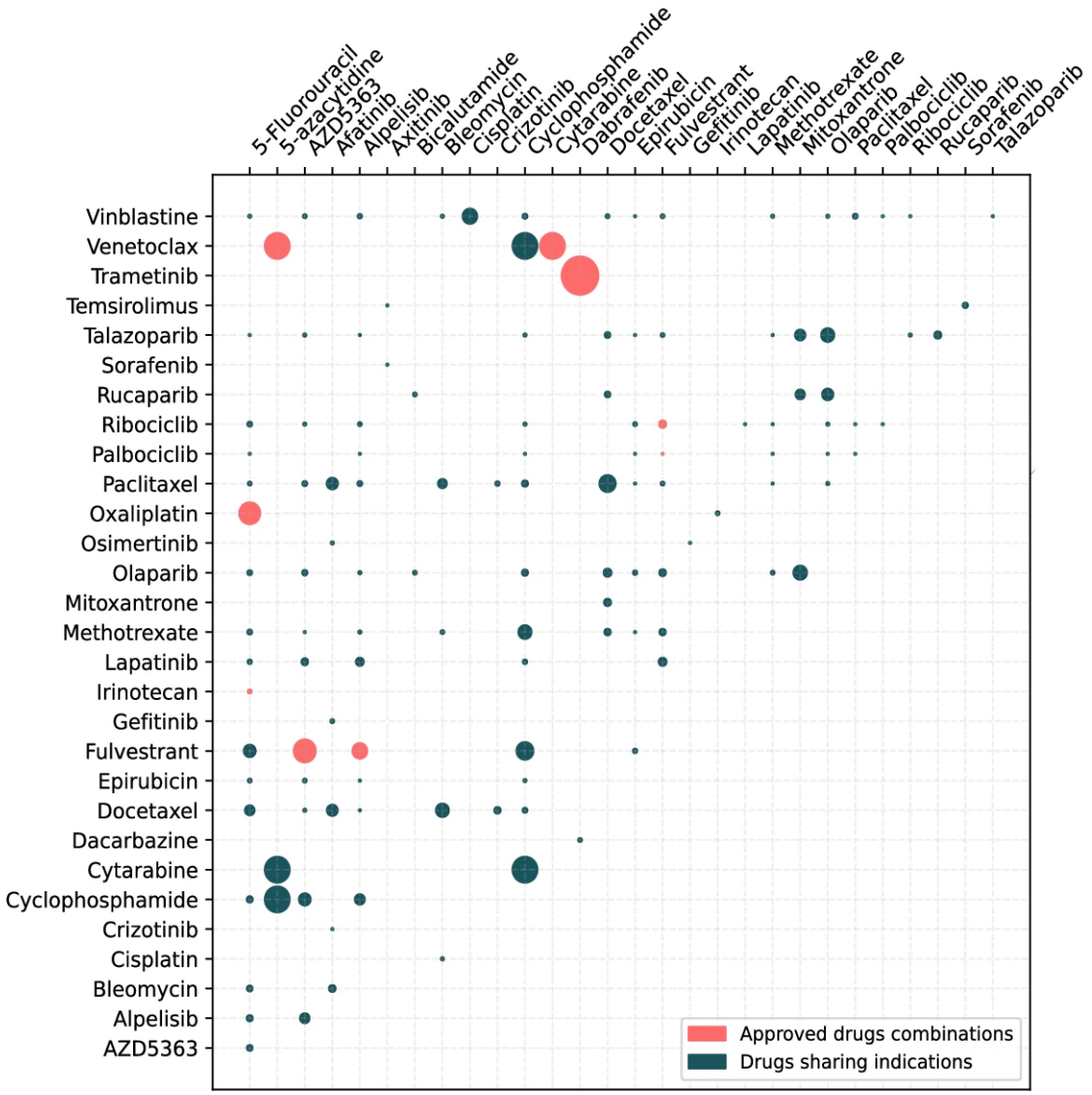

Figure 6.22: **Predicted drug combinations across TCGA tumors.** Each circle represents a drug–drug pair, with diameter proportional to the number of patient samples (within the top 600 predicted responders) jointly prioritized by both drug models. Colors denote the level of support: red highlights clinically approved combinations, while dark green indicates pairs sharing an approved indication for the same cancer type.

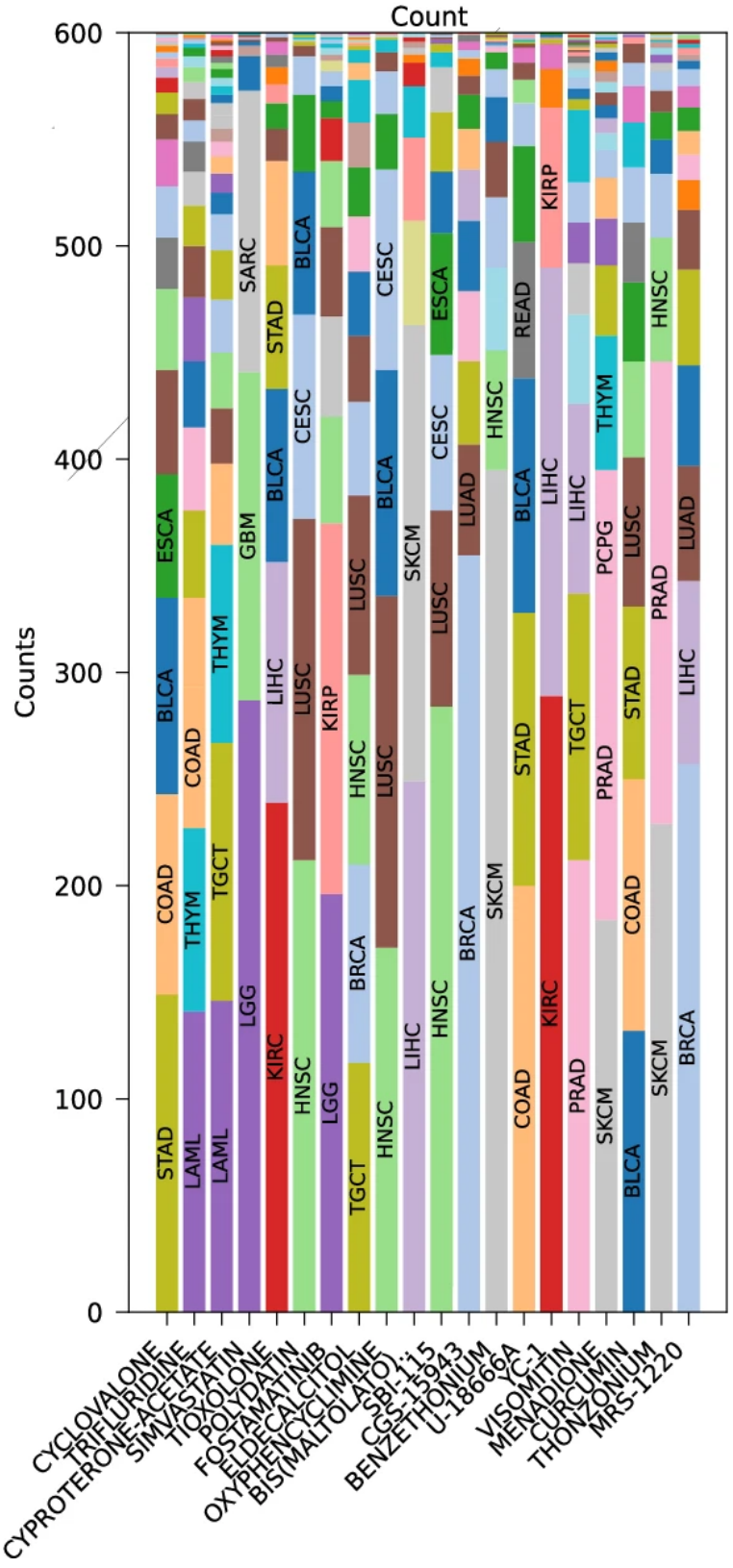

Figure 6.23: **Inference of TCGA tumors for non-oncological drugs** Inference on TCGA tumors using the 20 best performing non-oncological drug models trained on PRISM data. Each bar represents one drug, with the height of the stacked segments corresponding to the number of top-600 predicted samples, and the color denoting the associated cancer type. This highlights tumor type–specific sensitivity patterns and suggests opportunities for drug repurposing.

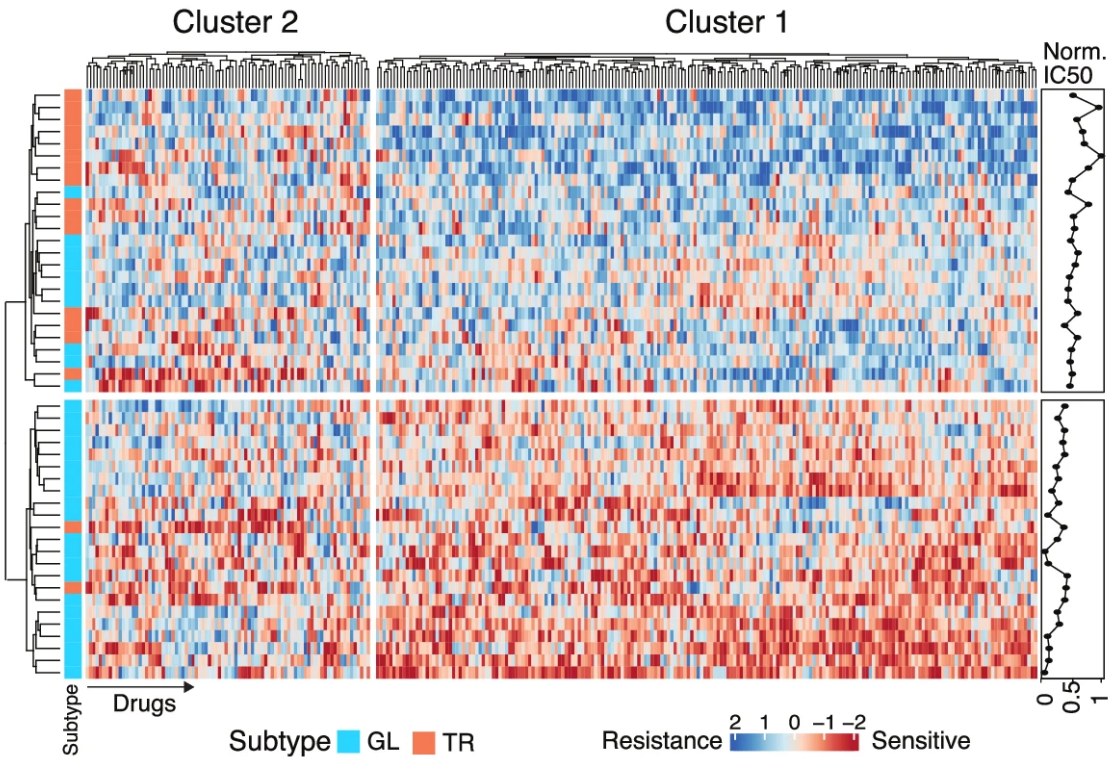

Figure 6.24: **Subtype-specific drug response patterns in PDAC inferred by CellHit.** Heatmap of predicted IC_{50} ($predIC_{50}$) values for GDSC drugs in PDAC samples. K-means clustering ($n = 2$, Euclidean distance) grouped samples into two major clusters. Subtype annotations (GL = Glandular, TR = Transitional) are shown alongside the heatmap, illustrating the separation of GL and TR subtypes into distinct response groups. Color scale denotes relative drug sensitivity, with blue indicating resistance and red indicating sensitivity.

(Gb130 and Gb107) with notably different predicted responses. Specifically, Gb130 was expected to be more sensitive to the Mcl-1 inhibitor AZD5991, whereas both samples were predicted to respond similarly to the XIAP inhibitor AZD5582. Model predictions of $\ln(IC_{50})$ and associated uncertainty are shown in 6.26A, with corresponding Quantile Scores in 6.26B. Experimental dose–response assays in the patient-derived cultures supported these predictions. As illustrated in Figure 6.26C–D, AZD5991 displayed stronger cytotoxicity in Gb130 than in Gb107. By contrast, AZD5582 produced comparable effects in both samples (6.26E–F), mirroring the model's ranking. The modest discrepancies between predicted and measured $\ln(IC_{50})$ values, especially for AZD5582, likely stem from patient-specific transcriptional programs not fully represented in the training data.

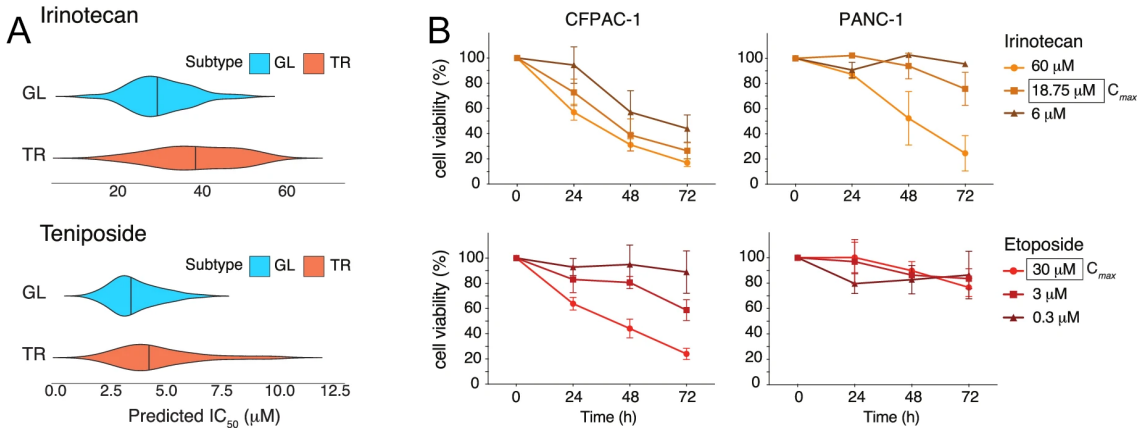

Figure 6.25: **Subtype-specific sensitivity to topoisomerase inhibitors.** (A) Violin plots showing the predicted IC_{50} ($predIC_{50}$) values of Irinotecan and Teniposide for the Glandular (GL, blue) and Transitional (TR, orange) PDAC subtypes. (B) Cell viability assays in CFPAC-1 (GL-like) and PANC-1 (TR-like) cells treated with increasing concentrations of Irinotecan or Etoposide at 24, 48, and 72 hours. Data represent the mean of three independent experiments ± SD. The results confirm the higher sensitivity of GL-like cells, consistent with model predictions.

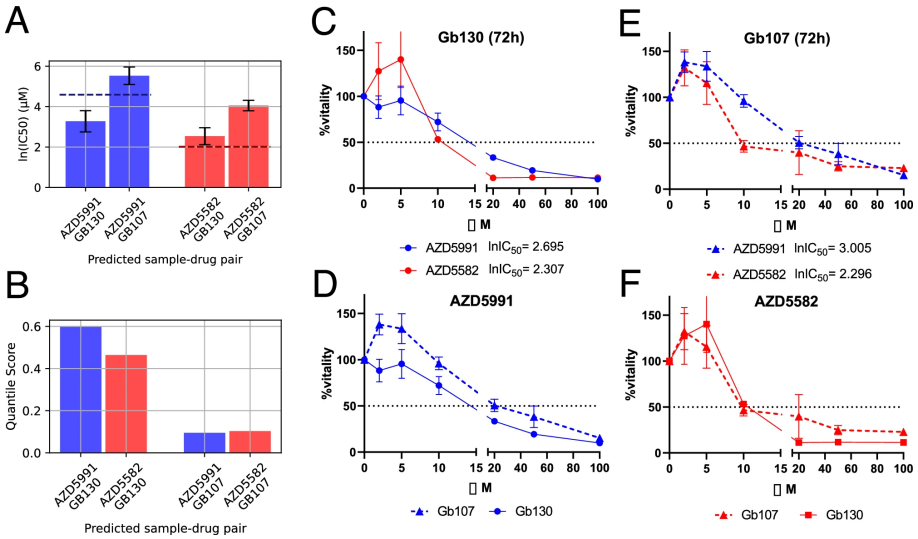

Figure 6.26: **CellHit predictions and validation in primary GBM cultures. (A)** Predicted $\ln(IC_{50})$ for AZD5991 (blue) and AZD5582 (red) in Gb130 and Gb107, with GDSC medians as reference (dashed lines) and ensemble uncertainty (error bars). **(B)** Predicted Quantile Scores for the same pairs. **(C,E)** Dose–response curves (72h) for Gb130 (**C**) and Gb107 (**E**), with fitted $\ln(IC_{50})$ values. **(D,F)** Cross-sample comparisons for AZD5991 (**D**) and AZD5582 (**F**). Gb107 is shown with dashed lines/triangles and Gb130 with solid lines/circles. The dotted line marks 50% viability. Error bars show triplicate assay variation.

Chapter 7

A web server to predict and analyze cancer patients' drug responsiveness

This chapter is based on the journal paper F. Carli, N. De Oliveira Rosa, S. Blotas, P. Di Chiaro, L. Bisceglia, M. Morelli, F. Lessi, A. L. Di Stefano, C. M. Mazzanti, G. Natoli, et al. Cellhit: a web server to predict and analyze cancer patients' drug responsiveness. *Nucleic Acids Research*, page gkaf414, 2025

7.1 A public end-to-end web server for transcriptomics-based drug response prediction

Overview of the CellHit Web Server We developed and deployed *CellHit* (https://cellhit.bioinfolab.sns.it/), an open-access web server designed to predict cancer drug sensitivity directly from bulk RNA-seq data (see Figure 7.1).. The platform is intended for researchers without programming expertise and provides a fully automated pipeline from data upload to interactive analysis. By integrating large-scale pharmacogenomic datasets from GDSC2 and PRISM with transcriptomic reference profiles from CCLE and TCGA, *CellHit* embeds new samples in a biologically meaningful context. The service is freely available, does not require user registration, and returns interactive results that can support hypothesis generation, patient stratification, and exploratory research

Data Input and Preprocessing Users can upload bulk RNA-seq expression profiles in CSV, ZIP, or GZ formats. Expression values should be provided as log-transformed TPM ($\log_2(\mathrm{TPM}+1)$), with gene identifiers specified as either HGNC symbols or Ensembl IDs. To ensure compatibility with the TCGA reference panel, the server applies `pyComBat`-based batch correction, optionally conditioned on user-supplied tumor type labels. Missing genes are automatically imputed using a gradient-boosted decision tree model (`XGBoost`) trained on synthetically masked TCGA profiles, with hyperparameters optimized through `Optuna`.

Sample Alignment and Visualization After preprocessing, uploaded samples are embedded into a joint CCLE–TCGA reference space using an enhanced *Celligner* work-

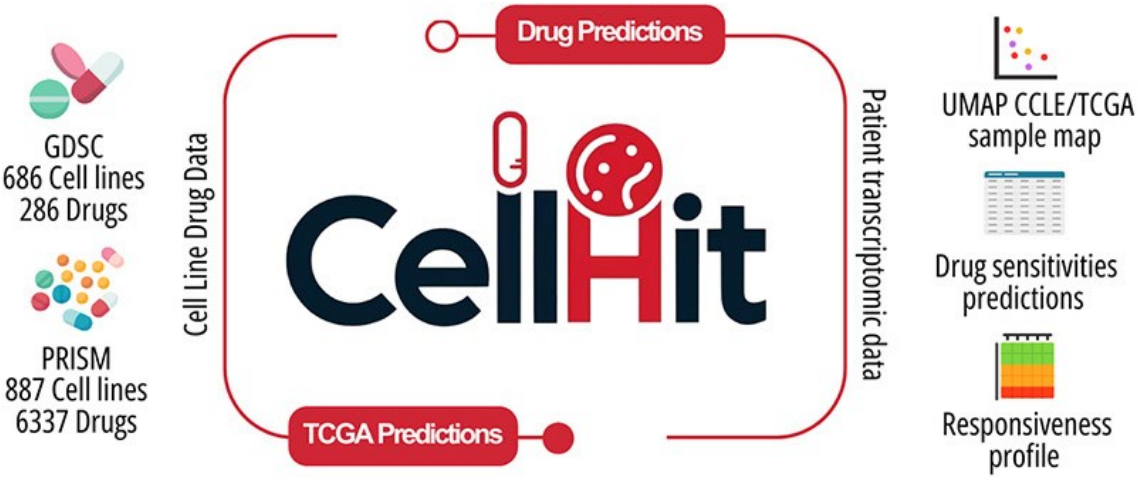

Figure 7.1: **Graphical overview of the *CellHit* web server.** The platform integrates large-scale cell line drug response datasets (GDSC, PRISM) with patient transcriptomic data (TCGA, CCLE) to enable drug sensitivity predictions. Key functionalities include parametric UMAP sample embedding, automated drug response modeling, and interactive visualization of predicted sensitivities and responsiveness profiles.

flow coupled with *Parametric UMAP*. Unlike standard UMAP, which recomputes the embedding whenever new data are added, the parametric formulation learns a mapping function that remains fixed after training. This allows any new sample to be projected directly into the established reference space without rerunning the entire alignment, ensuring both efficiency and reproducibility. The resulting two-dimensional embedding can be annotated with tissue type and OncoTree classifications, making it possible to assess sample placement, alignment quality, and biological similarity to well-characterized reference tumors in a consistent way.

Drug Sensitivity Prediction and Contextualization The predictive core of *CellHit* applies pre-trained, drug-specific machine learning models to estimate therapeutic sensitivity. For compounds in the GDSC2 panel, the system predicts natural log-transformed IC_{50} values, whereas for PRISM drugs it estimates log-fold change (LFC) in viability. Each prediction is accompanied by uncertainty estimates and quantile scores. Aditionally, each sample is contextualized by identifying the most transcriptionally or response-similar samples in TCGA and CCLE, using FAISS-based nearest-neighbor search. To aid interpretability, the server computes SHAP values for every sample–drug pair, highlighting the genes most influential in determining sensitivity or resistance.

Interactive Analysis and Quality Control Results are presented through linked, interactive visualizations: (i) a parametric UMAP embedding showing sample placement in the reference space; (ii) sortable and filterable prediction tables; (iii) per-sample *SHAP* importance plots for the top 15 genes; (iv) KDE-based diagnostics to distinguish selective efficacy from general cytotoxicity both within-sample (drug vs. other drugs) and across samples (drug vs. other samples); and (v) clustered heatmaps of drug responses with customizable scaling (median subtraction or z-scoring) and drug subset selection based on variance.

Precomputed TCGA Resource To facilitate immediate exploration, the web server hosts a large precomputed dataset comprising millions of predictions for all TCGA samples across both GDSC2 and PRISM drug panels. This resource includes ln(IC50) or LFC values, empirical drug statistics, `OncoTree` classifications, and the top SHAP genes per prediction, enabling rapid in silico hypothesis generation without requiring any user-uploaded data.

Implementation and Reusability The backend is implemented in `FastAPI` with `Celery`-based orchestration, a `MySQL+GraphQL` data layer, and task queuing via `Redis`. The frontend is built with `ReactJS` and `Plotly.js` for interactive visualizations, combined with `PrimeReact` and `React-Bootstrap` UI components. All core computational modules, including the GPU-enabled reimplementations of `Celligner` and `Parametric UMAP`, are released as standalone Python packages on `PyPI`, ensuring that the methodology is reusable beyond the web interface.

7.2 Enhanced cross-domain alignment between patient tumors and cell lines

We developed an enhanced version of the *Celligner* framework [380] to improve the alignment of patient tumor transcriptomes from TCGA with cancer cell line profiles from CCLE. The new implementation introduces three key advances: it eliminates information leakage during preprocessing, modernizes the computational backbone for greater efficiency and reproducibility, and incorporates systematic hyperparameter tuning to improve the biological validity of the aligned space.

Dataset-specific standardization to prevent leakage. Mean and standard deviation statistics were computed separately for each reference dataset (TCGA and CCLE) prior to transformation. New patient samples are standardized using only the statistics of the target reference domain, ensuring independence between datasets and avoiding bias in the alignment.

Full Python/PyTorch re-implementation of cPCA and DE steps. The contrastive principal component analysis (cPCA) [3] and differential expression (DE) steps were re-implemented entirely in Python using PyTorch [269]. The original version relied on a combination of `scikit-learn` and the R package `limma`, which made reproducibility and integration with machine learning workflows more difficult. By consolidating these steps into a single PyTorch-based implementation, we enabled GPU acceleration, streamlined execution, and simplified downstream integration with modern AI pipelines.

Reinstatement and tuning of the α hyperparameter. The α parameter in cPCA regulates the balance between removing dataset-specific variation and preserving biologically meaningful variance. In the upgraded framework, we reinstated α as a tunable parameter rather than fixing it, which allows the method to adapt more flexibly to different levels of divergence between datasets. This adjustment improves the quality of

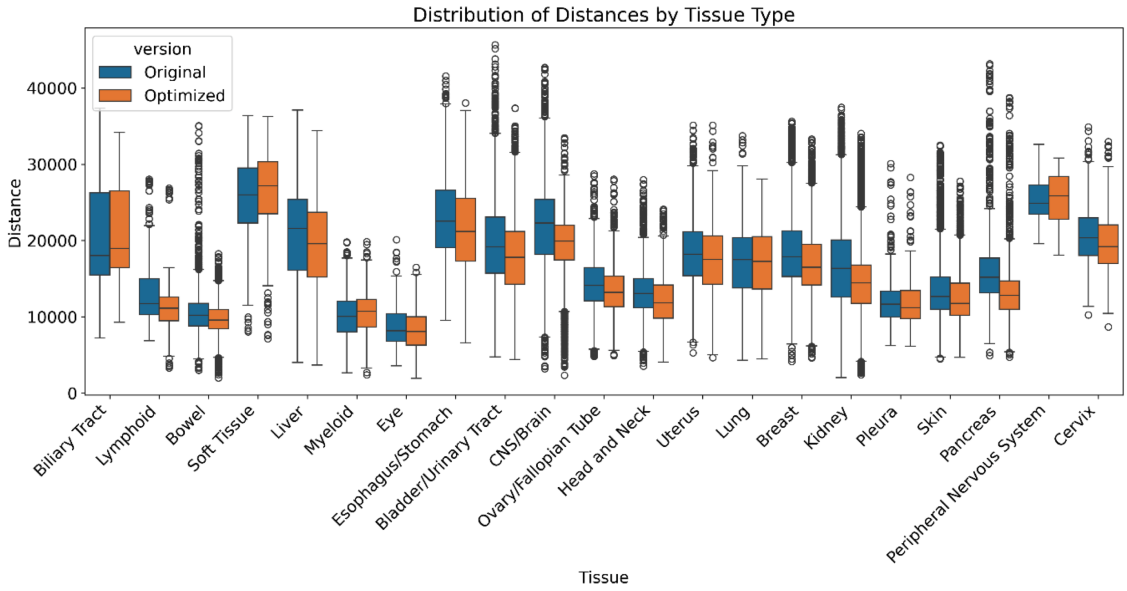

Figure 7.2: **Comparison of alignment strategies across tissues.** Distribution of Euclidean distances between CCLE cell lines and TCGA tumors, stratified by tissue type, under the original and optimized alignment procedures. The optimized strategy, guided by neighborhood consistency, achieves reduced distances in most tissues, indicating more biologically coherent alignment.

the aligned space by maintaining essential biological structure while minimizing confounding technical noise.

Hyperparameter optimization via neighborhood consistency. To improve the robustness of the alignment, we introduced a new objective function based on a *Neighborhood Consistency* score. This score measures how well biologically meaningful neighborhoods (defined by transcriptomic similarity) are preserved when projecting samples into the joint TCGA–CCLE space. The rationale is that cell lines from the same tissue should cluster together both before and after alignment. In practice, however, some CCLE cell lines display inconsistent or unreliable annotations even prior to alignment. To address this, we quantified for each cell line the proportion of its nearest neighbors that share the same tissue label. Cell lines for which fewer than 50% of neighbors belonged to the same tissue were considered low-quality and excluded. The remaining cell lines, which showed high agreement with their neighbors in the CCLE transcriptomic space, were retained as a *high-confidence* set. Hyperparameter optimization of the updated *Celligner* implementation was then carried out with the explicit goal of minimizing the distance between these high-confidence CCLE cell lines and the corresponding TCGA tumor samples from the same tissue. As shown in Figure 7.2, this optimization reduces tissue-specific distances across the majority of cancer types, enforcing biologically coherent alignment in a reproducible and robust manner.

7.3 Parametric UMAP for stable and consistent embeddings

Fixed Reference Space Construction. We implemented a *Parametric UMAP* model [306] in PyTorch to project transcriptomic profiles from The Cancer Genome Atlas (TCGA) and the Cancer Cell Line Encyclopedia (CCLE) into a shared low-dimensional reference space. Unlike standard UMAP [235], which must recompute embeddings whenever new data are added, the parametric version trains a neural network to approximate the UMAP transformation. This yields a *fixed* embedding space, so new samples can be placed directly without altering the global layout. The result is a stable and reproducible representation that avoids distortions from stochastic initialisation or dataset-specific batch effects.

Model Architecture and Training. To embed TCGA and CCLE transcriptomes into a shared low-dimensional space, we implemented a parametric version of UMAP in PyTorch [269] following the original formulation of Parametric UMAP from Sainburg et al. [306]. The model is a three-layer multilayer perceptron that maps $\log_2(\text{TPM}+1)$ gene expression profiles to two-dimensional coordinates. Unlike conventional UMAP, which requires recomputing the embedding when new samples are added, the parametric formulation learns a continuous mapping, enabling reproducible projections and fast GPU-accelerated inference. Training proceeds by first constructing a k-nearest neighbor (k-NN) graph with FAISS for scalable similarity search on datasets exceeding one million samples. Local bandwidths are then optimized via vectorized binary search to compute fuzzy simplicial set probabilities, which are symmetrized into the target neighborhood graph. The network is trained end-to-end on mini-batches of positive neighbor edges, with additional negative edges sampled at a 5:1 ratio to enforce repulsion. Optimization uses the UMAP cross-entropy loss directly, augmented by a Pearson correlation term to preserve global transcriptomic distances, and is performed with the AdamW optimizer on GPUs. Sparse matrix operations are employed throughout to reduce memory overhead, and the resulting model can be saved and reloaded for efficient application to new samples without re-embedding the reference.

Online Projection and Integration. New transcriptomic samples are processed using the same pipeline as the reference data—gene symbol harmonisation, optional batch correction with `pyComBat` [30], and dataset-specific scaling—before being passed through the parametric network. Because the reference space is fixed, projection can be performed online without recomputing the global embedding, which allows rapid and consistent alignment of external data.

Advantages for Interpretability and Reproducibility. A fixed embedding space ensures that sample positions remain comparable across different runs or datasets, preserving the biological meaning of neighbourhoods in the joint space. Moreover, inference requires only a single forward pass through the neural network, which substantially reduces computational cost and makes the approach suitable for interactive web deployment.

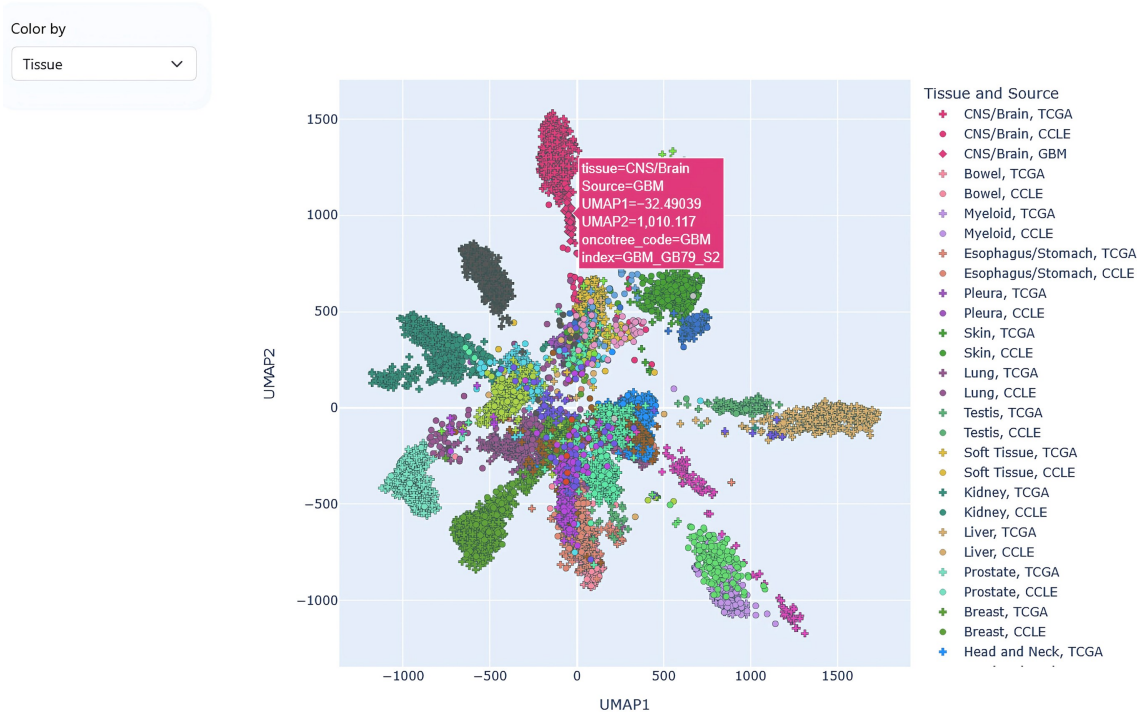

Figure 7.3: **Graphical abstract of Parametric UMAP alignment.** Fixed reference embedding space generated with Parametric UMAP, integrating TCGA tumors, CCLE cell lines, and patient-derived primary cultures. Diamonds indicate patient-derived GBM cultures aligned within the transcriptomic neighborhood of TCGA (crosses) and CCLE (circles) glioblastoma samples. Colors represent tissue origins, providing intuitive assessment of alignment quality and biological consistency.

Integration into the CellHit Aligner. The trained parametric UMAP model was incorporated as the final dimensionality reduction step in the enhanced *CellHit* pipeline (see Figure 7.3). It is available both for large-scale prediction tasks and through an *online aligner*, where users can upload bulk RNA-seq profiles and immediately visualise their position relative to TCGA tumors and CCLE cell lines. This integration enables efficient, reproducible, and biologically coherent sample contextualisation for translational applications.

7.4 Robust preprocessing stack for real-world transcriptomic inputs

Preprocessing framework. A key element of the web server data pipeline is a preprocessing framework that standardizes and completes input transcriptomic data before downstream analysis (see Figure 7.4). Real-world bulk RNA-seq profiles (whether from newly sequenced patients or external studies) often differ from reference compendia such as TCGA and CCLE in both technical and biological aspects. To ensure compatibility, we apply a two-step procedure: first, we correct for batch effects while preserving tissue-specific signals; second, we impute missing gene expression values using a machine-learning model trained on reference data.

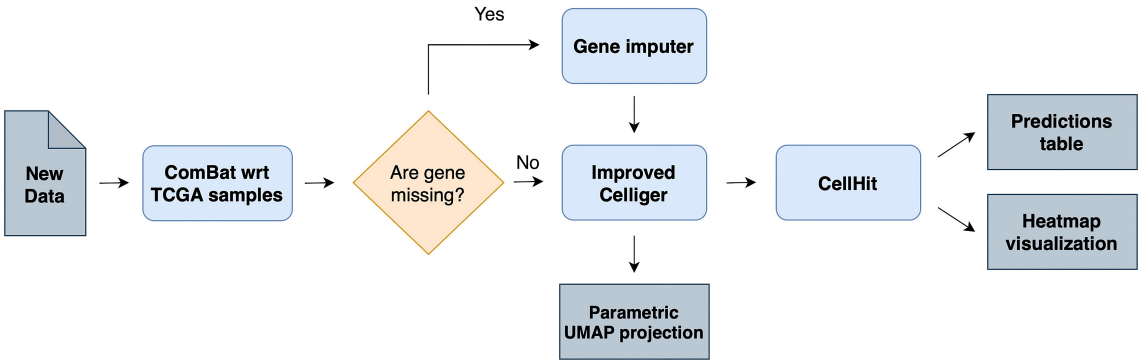

Figure 7.4: **Overview of the computational pipeline.** Bulk RNA-seq data are pre-aligned to TCGA reference samples using ComBat. If necessary, missing genes are imputed prior to processing with improved version of Celligner and the CellHit pipeline. The resulting outputs include a parametric UMAP projection of the aligned data, assay-specific predictions of drug sensitivity (tabular output), and a clustermap visualization highlighting drug response profiles across samples and drugs.

Batch Correction with Metadata Conditioning Incoming expression matrices are batch-corrected against TCGA references using the `pyComBat` implementation of the empirical Bayes ComBat algorithm [30, 409]. When available, the user-supplied TCGA tumor code (e.g., BRCA, LUAD, LAML) is passed to the `mod` parameter of `pyComBat` to perform conditional correction, preserving biologically meaningful variation associated with tumor type while removing technical offsets. This approach is particularly advantageous when integrating datasets from distinct sequencing facilities or library preparation protocols, as it mitigates systematic differences without erasing tissue-specific transcriptional programs. Inputs must be provided in $\log_2(\mathrm{TPM} + 1)$ scale, and the procedure assumes bulk tumor RNA-seq rather than cell-line–derived profiles.

Machine-learning–based gene imputation. To handle incomplete gene coverage caused by targeted panels, low read depth, or preprocessing filters, the pipeline uses a regression model based on XGBoost [60]. The model is trained on TCGA data where 10–20% of genes are randomly masked to mimic realistic dropout patterns. Hyperparameters are optimized with Optuna [10] using a Tree-structured Parzen Estimator [381] and three-fold cross-validation across stratified 400-gene subsets. During inference, the model predicts missing values from the observed genes and the tumor label, producing complete profiles that remain consistent with the reference distribution.

Impact on downstream analysis. This preprocessing strategy, combining metadata-aware batch correction with learned imputation, ensures that external samples are projected into the joint TCGA–CCLE reference space in a reliable way. By addressing both technical variability and missing data, it stabilizes the input profiles used for alignment (via the enhanced *Celligner* module) and drug-response prediction. As a result, the pipeline remains robust when applied to real-world datasets, which are often heterogeneous and incomplete compared to controlled training compendia.

7.5 Precomputed TCGA drug response resource

Uncertainty and quantile scoring. Each prediction is enriched with two reliability measures. First, we report uncertainty, estimated as the standard deviation across cross-validation folds, which captures the stability of the prediction. Second, we provide a *Quantile Score*, which places the prediction in the context of the empirical distribution of responses for that drug. These complementary metrics allow users to prioritize high-confidence, high-specificity predictions and to distinguish robust findings from noise.

Contextualization via transcriptomic and response neighbors. To support biological interpretation, each prediction is linked to a neighborhood context. Transcriptomic neighbors are identified using FAISS-based approximate nearest-neighbor search on aligned CCLE and TCGA expression profiles, highlighting samples that are molecularly similar. In parallel, response neighbors are computed by comparing predicted drug sensitivity patterns, regardless of gene expression. Together, these contexts help distinguish predictions driven by transcriptional similarity from those that reflect convergent response phenotypes.

Rich metadata integration. Predictions are connected to extensive metadata. Each TCGA sample is annotated with *OncoTree* codes, tissue of origin, and clinical descriptors, enabling stratification by histotype or subtype. At the drug level, summary statistics such as the minimum, median, and maximum predicted responses are provided, offering reference points to judge whether a given sample's prediction is typical or extreme within the cohort.

Mechanistic insights via SHAP attributions. For every sample–drug pair, we computed gene-level explanations using the `TreeExplainer` SHAP algorithm, retaining the 15 genes with the highest absolute importance values. These gene sets highlight molecular features that drive sensitivity or resistance, often recovering known or putative drug targets. By explicitly linking predictions to interpretable gene signatures, the resource connects statistical modeling to mechanistic hypotheses.

Utility and reuse. The resulting repository offers a comprehensive *in silico* landscape of drug sensitivity across TCGA tumors. Its combination of quantitative predictions, uncertainty estimates, contextual neighborhoods, disease metadata, and mechanistic gene attributions provides a multi-layered foundation for exploratory analysis, patient stratification, and therapeutic hypothesis generation. The dataset is openly reusable and designed to support both computational discovery and translational oncology research.

7.6 Built-in Interpretability and Quality Control

Interpretability and quality control. A central element of the framework is that predictive outputs are never presented in isolation. Every drug–sample prediction is paired with complementary layers of explanation and diagnostic evidence, so that users

can evaluate both the biological plausibility and the robustness of the result. This integration of interpretability and quality control metrics is achieved through three complementary modules: gene-level SHAP attributions, KDE–based diagnostics, and interactive cohort-level heatmaps. Together, these components allow users not only to interpret model predictions mechanistically but also to identify potential artifacts or spurious associations.

SHAP-based local explanations. For each prediction, whether expressed as a log-transformed IC_{50} or as a log-fold change in viability, the framework computes SHAP values using the `shap` Python library's `TreeExplainer`. SHAP decomposes the model's output into additive contributions from individual genes. A positive SHAP value indicates that the corresponding gene pushes the prediction toward resistance, while a negative value points to genes associated with sensitivity. For every sample–drug pair, the system reports the fifteen most influential genes ranked by absolute SHAP importance. These local explanations help users see which transcriptional drivers underlie the predicted response and provide a straightforward way to assess whether the model is capturing biologically meaningful signals or instead relying on spurious correlations (see Fig. 7.5).

KDE-based selectivity diagnostics. Predicted potency alone can conflate specific drug sensitivity with nonspecific cytotoxicity. To address this, the framework generates two kernel density estimation plots for each prediction. The first situates a drug's predicted effect relative to all other compounds within the same sample, making it clear whether a compound stands out as unusually effective in that particular transcriptional context. The second situates the same drug's predicted effect across the entire cohort of samples, highlighting cases where uniformly low viability values suggest broad toxicity rather than selective activity. Considering both perspectives helps ensure that prioritized drugs are not just generally toxic but exhibit meaningful selectivity for specific molecular contexts. As an example, Figure 7.6 shows kernel density estimation plots for SAR405838 (panel A) and for the GBM sample (panel B). In panel A, SAR405838's prediction lies far to the left of the cohort-wide distribution, indicating a strong viability reduction relative to most drugs, which could suggest broad cytotoxicity. However, panel B reveals that within the GBM sample, SAR405838 also stands out compared to the background of all other compounds, meaning that the predicted effect is unusually strong in this particular transcriptional context. The combination of both perspectives indicates that the compound is not merely toxic across the board but exhibits selective potency in at least one biologically relevant setting.

Interactive cohort-level heatmaps. Beyond individual predictions, the framework supports exploratory analysis at the cohort level. It generates interactive clustered heatmaps where rows correspond to patient samples and columns to drugs (Figure 7.7). Hierarchical clustering reveals groups of tumors with shared response profiles, while interactive filters enable researchers to focus on compounds that are most variable or clinically relevant. Two scaling modes are available: median-centered scaling, which emphasizes how each sample's predicted sensitivity deviates from a drug's typical activity, and within-sample standardization, which highlights relative drug rankings for a given transcrip-

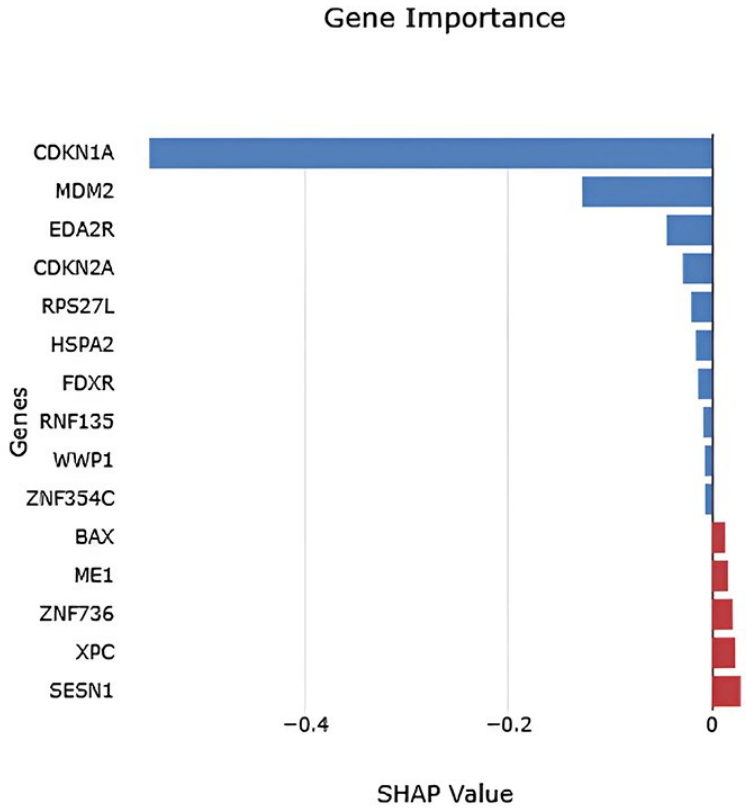

Figure 7.5: **Interactive prediction summary with putative target and key genes.** Example of SHAP-based feature attributions for a drug–sample prediction. Bars indicate the relative contribution of each gene to the predicted response, with negative SHAP values corresponding to sensitivity drivers and positive values to resistance drivers. The top fifteen genes ranked by absolute SHAP importance are displayed.

tome. As illustrated in Figure 7.7, these visualizations provide an intuitive way to detect outliers, uncover latent subtypes, and guide biomarker discovery. In the PDAC cohort shown, the highlighted region reveals a cluster enriched with TR-type samples that are predicted to be more resistant than the GL group, underscoring how the approach can stratify patients based on distinct molecular response patterns.

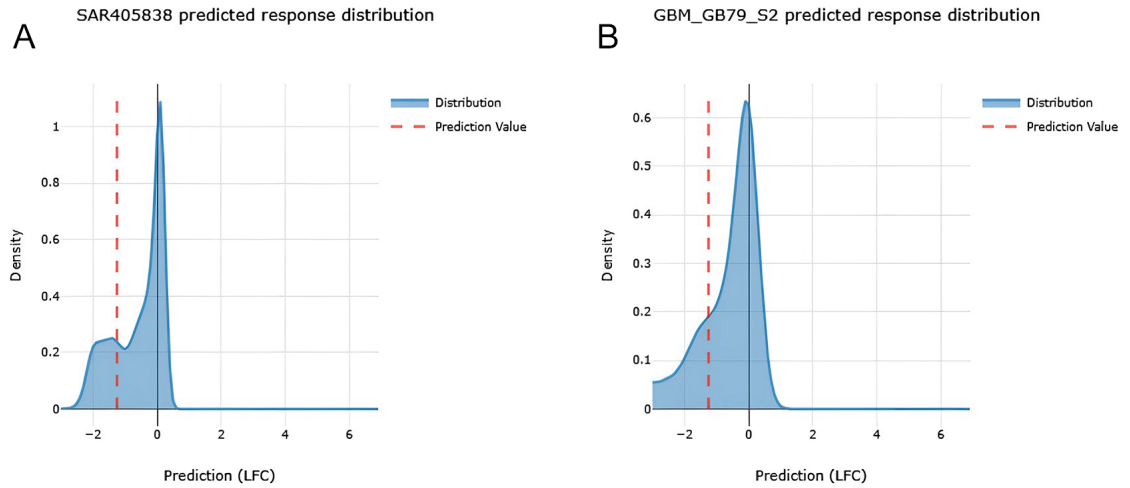

Figure 7.6: **Kernel density diagnostics for selectivity vs. general toxicity.** (A) Predicted response distribution for SAR405838 across the cohort of samples; the red dashed line marks the focal prediction, which lies at the extreme left tail, suggesting strong viability reduction. (B) Predicted response distribution of all compounds within the GBM_GB79_S2 sample; the red dashed line shows SAR405838 standing out from the background, indicating unusually strong activity in this transcriptional context. Taken together, the two perspectives demonstrate how the framework distinguishes between nonspecific cytotoxicity and context-dependent selective responses. Predictions are expressed as log-fold change (LFC), where lower values indicate higher predicted sensitivity.

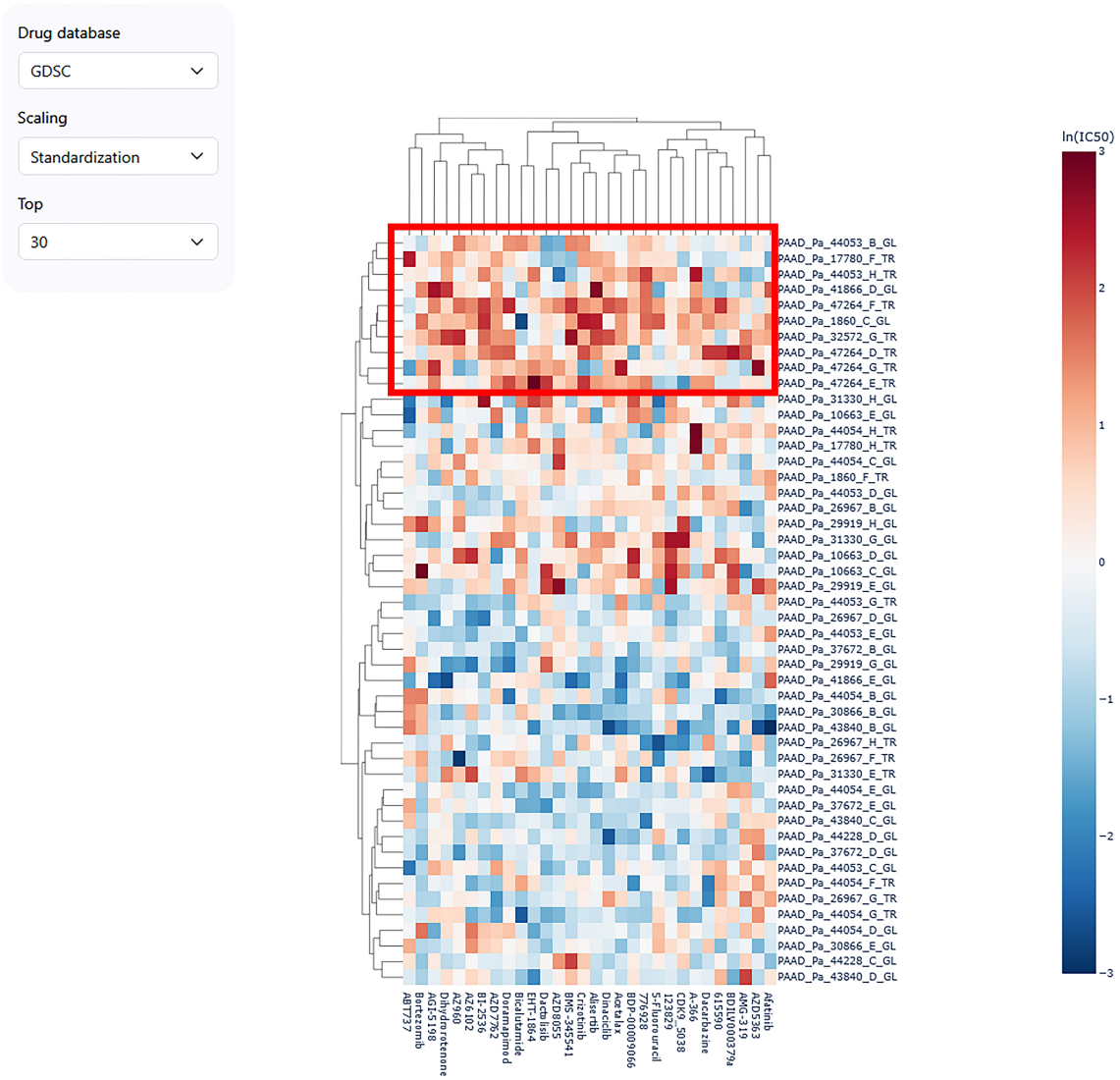

Figure 7.7: **Cohort-level drug response heatmap.** Clustermap displaying drug response predictions for PDAC patients. Rows correspond to patient samples and columns to drugs, with hierarchical clustering revealing shared response profiles. The color scale represents scaled ln(IC_{50}) values, where blue indicates higher sensitivity and red greater resistance. The highlighted region (rectangular box) marks a cluster enriched with TR-type samples, which are predicted to be more resistant compared to the GL group. Users can interactively adjust scaling methods and the number of top drugs displayed to explore response patterns.

Part IV

Discussion and conclusion

Chapter 8

Discussion and Open Research Directions

8.1 Biochemical level

Limits of augmentation techinques

A potential limitation of the present framework arises from the reliance on strategies such as loss rebalancing, negative sample generation, careful data splitting, and related augmentation procedures to mitigate dataset incompleteness in DTI modeling. While these approaches can partially alleviate class imbalance and compensate for sparsely annotated chemical-protein pairs, they cannot fully resolve the systematic biases embedded in current resources. Public bioactivity datasets, such as ChEMBL [405] or BindingDB [215], disproportionately represent well-studied targets (particularly kinases) whereas many biologically relevant proteins remain "dark" or "orphan," with little or no interaction data available [312, 324]. Importantly, even within extensively investigated receptor families such as GPCRs, a substantial proportion of orphan proteins persists, meaning that performance estimates derived from annotated members of the family may be overly optimistic and biased when extrapolated to orphan receptors [131]. In these cases, augmentation and rebalancing strategies cannot compensate for the complete absence of ground-truth labels, since information on the true interaction profiles of these proteins is fundamentally missing. As a result, models trained predominantly on historically studied targets risk learning biased representations that generalize poorly to understudied or novel proteins, thereby perpetuating existing knowledge gaps in drug discovery. Addressing this challenge requires not only improved augmentation schemes but also systematic efforts to expand experimental coverage of understudied proteins.

Foundation models and better representations

Another area of improvement concerns the representations adopted on both the protein and ligand sides. As introduced in chapter 3.3, recent foundation models indicate that richer, more transferable embeddings could strengthen drug-target prediction. On the protein side, large protein language models learn structure- and evolution-aware features directly from sequence and already enable near–state-of-the-art structure inference (e.g., ESM-2/ESM-Fold) [209, 292]. More recent multimodal generative models (e.g., ESM-3 [133]) further integrate sequence, structure and functional constraints, suggesting additional gains in downstream generalization when appropriately aligned to binding tasks . On the small-molecule side, chemical foundation models trained at scale (e.g., MolE [236]) provide atom-level graph embeddings that outperform task-specific

baselines across diverse ADMET endpoints, supporting their use as general-purpose compound encoders. Notwithstanding these advances, preliminary experiments within our framework indicated limited benefits when such representations are transferred naively to DTI; this likely reflects a mismatch between pretraining objectives and the binding-specific signal of interest. One promising approach is to utilize models that work directly at the atomic level and explicitly represent 3D interactions. For instance, architectures capable of jointly predicting and reasoning about protein-ligand contacts such as AlphaFold 3 [4], generative docking and co-folding models like Boltz-2 [268], DynamicBind [220], NeuralPLexer [280], and ATOMICA [96] can be adopted. Additionally, these models can be fine-tuned using structure-aware objectives, including contrastive pocket-ligand alignment, pose consistency, and pocket-conditioned masking, all designed to enhance binding fidelity. A systematic benchmarking of these representations within the current pipeline, specifically focusing on their ability to generalize to poorly annotated or orphan targets, is a concrete and technically feasible next step.

Pocket-Conditioned Generative Screening

A promising strategy for advancing structure-based drug discovery lies in the use of pocket-conditioned generative models, such as Pocket2Mol [274], PocketFlow [164], and Lingo3DMol [98] and DiffSBDD [318]. These approaches directly synthesize candidate ligands within the geometric and physicochemical context of a binding site, thereby embedding structural priors at the point of generation. Rather than relying exclusively on existing ligand datasets, this paradigm enables the creation of tailored molecular scaffolds that are inherently compatible with target pockets. These generated molecules could be used as "warm-up" samples to train lightweight screening models, enabling them to internalize structural binding features in the absence of empirical ligands. Once primed, these streamlined models can efficiently screen expansive compound libraries, reserving computationally heavy or experimental evaluation for a refined shortlist. This hierarchical pipeline of generation, lightweight screening, and refined validation may provide a scalable strategy for ZINC-scale virtual screening by balancing structural fidelity, efficiency, and throughput.

Active Learning strategies

A further limitation of the present framework stems from its passive reliance on fixed training datasets, which constrains the chemical and protein spaces that can be effectively explored. One promising strategy to overcome this bottleneck is the adoption of *active learning* paradigms [287], in which the model iteratively identifies areas of uncertainty and queries an external oracle for additional information. By prioritizing the acquisition of labels for compounds or targets that most reduce predictive uncertainty, active learning has been shown to accelerate model improvement in low-data regimes and to enhance coverage of underexplored regions of chemical space [95, 332]. In the context of drug-target interaction modeling, the oracle could be instantiated not only as a human expert but also as a computationally intensive model that provides high-fidelity interaction estimates at scale, for example through physics-informed simulations or generative co-folding architectures [4, 268]. This hybrid design would enable lightweight predictors to bootstrap from sparse supervision while selectively incorporating information from more accurate, but expensive, methods. In practice, such strategies could be used to guide the systematic filling of "dark spots" in the interaction landscape, ensuring that the learned representations are not only data-efficient but also better aligned with regions of genuine biomedical interest. Integrating active learning into the pipeline therefore represents a concrete avenue to counteract dataset incompleteness and im-

prove generalization beyond historically overrepresented targets.

8.2 Cellular level

8.2.1 Expanding Cell Hit to new modalities

Transcriptomics: strengths and rationale

Gene expression remains one of the most informative molecular readouts for predicting drug response [100], as it reflects the functional state of the cell and captures regulatory programs that are not directly observable from static genomic alterations. Transcriptomic data offers several practical advantages that justify its central role in *CellHit*: it provides a comprehensive snapshot of cellular activity, is widely available across large pharmacogenomic resources, and has demonstrated strong predictive performance in benchmark comparisons. Indeed, recent large-scale evaluations have shown that the predictive power of the proteome for drug response is very similar to that of the transcriptome [119], supporting the use of gene expression as a robust and efficient starting point for drug sensitivity modelling.

Limitations of transcriptomics-only models

Nevertheless, transcriptomics-only approaches have inherent limitations. First, transcript abundance does not always correlate with protein levels due to post-transcriptional regulation, translation efficiency, and protein degradation [119]. Second, many tumor types are strongly associated with molecular layers that are not captured by gene expression alone, such as copy-number variations in breast cancer [334] or DNA methylation profiles in gliomas [397]. Moreover, gene mutations have been shown in ablation studies to exert a greater influence on drug response prediction than other omics data types in certain contexts [375], suggesting that some therapeutic targets may be better captured through genomic or proteomic features.

Multi-omic resources and integration frameworks

A promising way forward lies in the development of integrative approaches capable of combining multiple omics into unified latent representations, thereby capturing both shared and modality-specific biological signals. Resources such as CCLE and GDSC now provide rich, matched multi-omic profiles (transcriptomics, copy number variations, mutations, proteomics, and metabolomics) for hundreds of cell lines, offering an ideal substrate for multi-modal model development [375]. Canonical examples of integrative frameworks include MOFA [14] and its successor MOFA+ [15], which provide scalable, factor-analytic frameworks for unsupervised integration, as well as supervised alternatives such as DIABLO [329]. More recently, synthetic augmentation frameworks such as MOSA [46] have demonstrated the utility of deep generative models in expanding and harmonizing multi-omics datasets, supporting downstream predictive tasks.

Clinical successes of multi-omic models

Multi-omic machine learning models for therapy response have already achieved remarkable results in clinical settings. For example, Sammut et al. [309] combined clinical, digital pathology, genomic, and transcriptomic profiles of pre-treatment breast cancer biopsies and demonstrated that response to therapy is determined by baseline characteristics of the tumour ecosystem; their multi-omic model achieved an area under the curve of 0.87 for predicting pathological complete response [309]. Similarly, proteogenomic characterizations of small cell lung cancer and acute myeloid leukaemia have identified multi-omic subtypes with subtype-specific therapeutic vulnerabilities that would not be discernible from transcriptomics alone [214, 276]. These studies suggest that integrating complementary modalities can expose drug-response patterns that are invisible to

single-omic models.

Challenges and future directions for CellHit

However, as already introduced in the methods section 4.2.1, multi-omics data is inherently high-dimensional: multi-omics profiles typically involve tens of thousands of features per modality, while the number of available samples remains modest. This imbalance increases the risk of overfitting and limits model generalizability across independent cohorts. A key challenge for such integrative frameworks is also the preservation of interpretability. Latent variable models often trade direct feature correspondence for compact representations, complicating biological validation and downstream mechanistic interpretation. For *CellHit*, future extensions toward multi-omics integration will need to balance representational power with the requirement of retaining transparent links to original features. Only by maintaining this interpretability can predictive models remain actionable in translational contexts, where mechanistic plausibility is as critical as predictive accuracy.

8.2.2 Expand Cell Hit with new predictive models

TabPFN improves results

Beyond traditional gradient-boosted models, recent progress in transformer-based tabular prediction frameworks, such as TabPFN [144], offers the prospect of improved accuracy for cell line drug sensitivity prediction. Indeed, preliminary analyses suggest that TabPFN can capture higher-order dependencies among gene expression features, yielding competitive or superior performance compared to tree-based methods. However, a current limitation of TabPFN is its restricted input dimensionality [144], which prevents the direct use of genome-wide transcriptomic profiles. This necessitates a feature selection or filtering step to reduce dimensionality prior to model training, a procedure that introduces an additional layer of methodological dependency and may inadvertently discard biologically relevant signals. Future work could explore scalable adaptations of TabPFN, such as hierarchical feature grouping or sparse attention mechanisms, that would allow models to accommodate larger gene sets without prohibitive computational costs. Such developments could enable the direct incorporation of transcriptome-scale inputs and reduce the risk of information loss at the preprocessing stage.

Interpretability trade-offs

A second challenge concerns interpretability. The current design of *CellHit* relies on gradient-boosted decision trees, which naturally lend themselves to efficient computation of Shapley values [227]. These local attributions have been essential to characterizing drug-specific molecular predictors and ensuring that the model's internal logic remains biologically transparent. Transformer-based approaches, while more expressive, lack comparably efficient and reliable interpretability pipelines. Although approximation methods for Shapley values exist in the context of deep models, they are computationally intensive [227]. Developing interpretability tools that are both computationally tractable and biologically faithful will be crucial if transformer-based architectures are to become viable replacements or complements to tree-based models in translational settings.

8.2.3 Improving cell line-tumor alignment

Inherent limitations of cell line models

Aligning cell line transcriptomes to tumor profiles faces fundamental challenges rooted in the inherent limitations of cancer cell lines as tumor models. Several factors compromise how faithfully cell lines represent primary tumors: genetic drift under in-vitro

selective pressures leads to continuous molecular divergence from the tumors of origin; selection bias favors the establishment of cell lines from more aggressive subtypes due to their predisposition to grow successfully in culture; and the complete absence of tumor microenvironment components (immune cells, fibroblasts, extracellular matrix) eliminates critical biological context [359]. These intrinsic differences manifest as systematic transcriptional shifts that confound direct comparison between in-vitro models and patient samples.

Tumor purity confounds bulk alignment

Computational methods developed to bridge this gap generally pursue three interconnected objectives: integration of cell line and tumor data into comparable feature spaces, scoring of cell lines by their suitability as tumor models, and selection of representative models for defined tumor types or subtypes [359]. Unsupervised alignment procedures such as Celligner [380] address the first objective by applying contrastive principal component analysis followed by mutual nearest neighbor correction to create a joint embedding space. However, bulk tumor measurements remain confounded by variable tumor purity and microenvironmental composition, which dilute malignant signals and bias similarity assessments. This issue is particularly acute in cohorts with strong stromal and immune infiltration, where alignment may overfit to non-malignant variation rather than tumor-intrinsic programs. Current mitigation strategies, including removal of genes correlated with purity estimates or application of batch correction methods, incompletely address this fundamental issue: the first contrastive principal components removed by Celligner are enriched in immune pathways and correlate strongly with tumor purity estimates, yet residual microenvironment effects persist and require additional correction steps [359].

Single-cell references for malignant signal isolation

A promising avenue to more directly isolate malignant signal would leverage single-cell technologies, which provide a unique opportunity to circumvent some of these intrinsic limitations by enabling comparison between cell lines and pure populations of malignant cells from patient tumors [359]. Pan-cancer single-cell atlases have revealed that many recurrent drivers of transcriptional heterogeneity observed in tumors are also present in cell lines, suggesting that individual cell lines might serve as models for specific components of intra-tumor heterogeneity [185]. Building on this foundation, tumor-type-specific malignant reference profiles derived from cancer-specific single-cell atlases could be used to deconvolute bulk RNA-seq before alignment. Single-cell studies provide high-resolution definitions of malignant cell states discriminated from stromal and immune populations, offering stable gene programs that can serve as malignant "fingerprints" [253, 355]. Integrating these references into a digital cytometry framework (e.g., CIBERSORTx) would enable the estimation of cell-type fractions and the imputation of malignant-cell expression from bulk mixtures, effectively purifying tumor-intrinsic signal prior to cross-system mapping [250].

Methodologically, a practical pipeline would involve:

1. curating single-cell compendia per tumor type;

2. defining malignant-state signatures with explicit controls for patient-specific and tumor microenvironment variability;

3. applying reference-based deconvolution to reconstruct malignant expression from bulk samples; and

4. rerunning alignment between cell lines and the deconvolved malignant profiles.

Validation and broader impact

This design isolates the biological quantity of interest (malignant programs) and reduces dependence on *ad-hoc* batch correction, while remaining compatible with current alignment tooling. Such an approach aligns with the emerging consensus that large single-cell atlases for both in-vitro models and tumor patients offer a unique opportunity to create integrated references enabling direct comparisons with single-cell resolution [359]. Evaluation should quantify improvements in tumor-cell line correspondence and downstream predictive validity (e.g., increased concordance between aligned similarity and genotype- or dependency-based benchmarks), as well as robustness across purity strata and immune/stromal compositions.

8.2.4 Refining the characterization of drug MOAs

Updating LLM-based pipelines

A potential limitation of the current pipeline lies in the reliance on a LLM that, while state-of-the-art at the time of implementation, is now almost two years old. In the rapidly evolving landscape of LLMs, this represents a significant gap, as more recent models demonstrate markedly improved factual accuracy, reduced hallucination rates, and enhanced capacity for reasoning over structured biomedical knowledge [349, 396, 406]. Updating the pipeline with contemporary architectures would likely yield more reliable and fine-grained mechanistic annotations, thereby strengthening downstream analyses.

Granularity of pathway selection on Reactome

A further limitation concerns the biological resolution at which drug mechanisms were annotated. As outlined in Chapter 6.2, the selection of candidate pathways modulating drug action was performed at a relatively coarse level, identifying broad Reactome categories while lacking the granularity necessary to resolve specific sub-pathways, signaling branches, or context-dependent interactions. The current selection procedure relies exclusively on the generated drug MOA description, the putative target, and supporting information extracted from PubMed searches. Crucially, no textual description of the pathways themselves, nor details of their constituent genes, are provided to the LLM. Consequently, the model is forced to select pathways solely on the basis of their names and its internal prior knowledge, an abstraction that risks obscuring subtle but clinically relevant mechanistic distinctions. Future iterations of the pipeline could incorporate more fine-grained ontologies, curated pathway databases, and structured representations of pathway content. Such improvements would enable a more systematic and scalable curation of drug associations to mechanistic biological processes, with the potential to capture clinically actionable biological nuance that is currently lost at higher levels of abstraction.

On-demand biological resources

Recent methodological advances open opportunities for more efficient pipelines. The declining costs of LLM inference, together with the emergence of modular retrieval protocols such as the Model Context Protocol (MCP[1]), enable on-demand access to structured biological resources through resources such as Kuehl et al. [194]. Incorporating these systems into the characterization pipeline would allow LLMs to dynamically ground predictions in up-to-date curated knowledge, minimizing hallucinations and ensuring interpretability. A modular architecture combining generative models with retrieval-augmented querying of biological databases could yield scalable, transparent, and continuously improvable annotations of drug mechanisms.

[1] https://modelcontextprotocol.io

Toward efficient and reproducible pipelines

Taken together, these developments suggest a clear path toward refining the current framework. By retraining with newer LLM architectures, incorporating fine-grained biological ontologies, and embedding retrieval-augmented mechanisms, future pipelines may achieve higher accuracy, improved interpretability, and greater reproducibility. Beyond methodological refinement, such advances would also support translational applications, as more precise mechanistic mappings are essential for linking biochemical drug effects to patient-level outcomes in oncology.

8.2.5 Integrating interpretability with functional validation

From observational to functional validation

The interpretability studies in CellHit have primarily evaluated whether feature importance analyses recover biologically plausible aspects of a drug's mechanism of action. While this provides a useful sanity check, such analyses remain observational and do not establish the functional relevance of identified features. A natural extension involves comparing model-identified important genes with essentiality profiles from large-scale CRISPR–Cas9 knockout screens. In Chapter 6.3, we demonstrated that pooling tissue-wise SHAP importance scores from CellHit predictions recovers most tissue-specific genes identified by Pacini et al. [265]. However, this classification system allows genes to be flagged as specific to multiple tissues simultaneously. To obtain more refined results, future work should focus on genes specific to individual tissues exclusively, excluding those identified as pan-cancer or multi-tissue specific. Additionally, since our goal is to elucidate drug-specific or disease-specific mechanisms, we must filter out essential housekeeping genes involved in core cellular processes. While these genes are indispensable for cell survival, they provide limited insight into targeted therapeutic mechanisms.

Convergence with mechanistic modeling

A further opportunity lies in the systematic integration of interpretability-derived signals with mechanistic explanation models. Frameworks such as CARNIVAL [293], COSMOS [88] and CORNETO [293] provide causal reasoning over molecular interaction networks, explicitly embedding prior biological knowledge to reconstruct context-specific signaling dynamics. Intersecting the gene- or pathway-level insights produced by *CellHit* with the predictions of such mechanistic models would allow testing for convergence between data-driven and knowledge-driven evidence. Agreement across these complementary approaches would strengthen confidence in the validity of candidate mechanisms, while discrepancies could highlight areas where either the model or the prior knowledge is incomplete. Embedding this type of cross-validation into the analytic workflow could transform interpretability from a diagnostic add-on into a structured mechanism-discovery tool.

Toward mechanistically grounded interpretability

Overall, expanding beyond observational consistency checks toward functional validation and convergence with causal models represents a promising direction for future work. Such a framework would not only enhance confidence in the biological plausibility of model-derived features but also bridge predictive modeling with mechanistic interpretation. By systematically aligning interpretability signals with both functional genomics and prior-knowledge–based reasoning, *CellHit* could evolve into a platform that delivers mechanistically grounded predictions, supporting more reliable translational insights.

8.3 Translational level

8.3.1 Extending Validation to Tumor Subtypes

Validation scope limited to tumor types

A potential limitation of the validation strategy in tumor samples proposed in 6.6 is that drug response predictions were mainly compared against broad tumor types, specifically TCGA project categories. Although this level of granularity offers a first-order assessment of predictive performance, it overlooks the marked heterogeneity that exists within each type of cancer. Molecularly defined subtypes such as PAM50 classes in breast cancer, IDH mutant versus IDH wild-type gliomas, or KRAS-driven subsets in colorectal cancer have been repeatedly shown to exhibit distinct pharmacological sensitivities [22, 47, 141]. Consequently, limiting the evaluation to coarse tumor categories may mask drug–subtype associations of therapeutic relevance, particularly in cases where activity is confined to narrowly defined molecular contexts.

Subtype annotation as future direction

Future work could extend validation efforts by systematically annotating and incorporating tumor subtype labels across preclinical and clinical datasets. This would enable a more precise validation of the predictive model, assessing whether the model is capable of identifying subtype-specific vulnerabilities. Methodologically, this requires harmonizing subtype definitions across resources, accounting for discrepancies between clinical and preclinical taxonomies, and ensuring sufficient sample representation to support statistically robust checks. By integrating subtype-level annotations from actual tumor samples, predictive models trained on cell lines could be evaluated for their ability to capture clinically relevant subtleties that emerge only in patient-derived contexts. This refinement directly addresses the ultimate goal of such frameworks: enabling patient stratification and the discovery of molecularly defined tumor subgroups. Beyond enhancing the interpretability of predictions, testing performance at the subtype resolution in real tumors provides a stringent benchmark for translational utility, ensuring that the model's predictive signal extends beyond controlled in vitro systems to the heterogeneous and clinically complex landscape of patient tumors.

8.3.2 Incorporating Toxicity and Dosing Information into Drug Ranking

IC_{50} and quantile score limitations

The prioritization criteria used in Carli et al. [51] stage depended on potency measures such as IC_{50}, along with a quantile score designed to balance specificity and overall activity across the set of compounds. While this approach marks progress toward a more nuanced ranking of candidate therapies, a key limitation is that potency alone is not an adequate measure of therapeutic suitability. Importantly, classical viability-based metrics are sensitive to cell division rates, which can skew rankings in favor of broadly cytotoxic agents [126, 127]. Consequently, drugs with inherently higher cytotoxicity profiles may achieve lower IC_{50} values, even if they lack clinical viability. This can result in rankings that do not accurately reflect translational priorities.

Need for systematic toxicity data

A key limitation of the current ranking framework is the absence of systematic toxicity and clinical dosing information, which risks over-prioritizing compounds that appear potent in vitro but are clinically unsuitable due to narrow therapeutic windows or high systemic toxicity. A promising avenue for improvement involves integrating structured

dosing and label information from resources such as DrugBank [387], together with post-marketing safety evidence from pharmacovigilance repositories such as FAERS and its curated derivative AEOLUS [23]. By contextualizing potency values with clinically established dose ranges and tolerability profiles, predictions could be normalized according to therapeutic index, allowing for a clearer distinction between compounds whose low IC_{50} values reflect specific pharmacological targeting versus those driven by non-specific cytotoxicity.

Therapeutic index–aware rescaling

Methodologically, rescaling potency-based predictions by approximations of the therapeutic window, linking exposure, efficacy, and safety, could reduce the weight of non-specific cytotoxins and align in silico classifications with clinical plausibility [244]. This adjustment complements the robust response metrics of the division rate and supports clinically significant compound prioritization [126, 127].

8.4 Multi-scale integration

From siloed to integrated scales

The methods developed so far operate largely within a single scale, focusing either on biochemical binding or on cellular drug response, and therefore do not yet exchange information across these layers. However, recent advances indicate that representations learned at one level can be made explicitly *context-aware* and transferred across scales, opening the door to principled multi-scale coupling. Building on these insights, our current frameworks could evolve toward genuine cross-scale integration with only modest conceptual extensions.

8.4.1 Incorporating context embeddings inside protein and molecule representations

Context-aware proteins and molecules

A central limitation of sequence- or structure-only protein embeddings is that they are *context-free*: the same representation is used irrespective of the cellular or tissue environment in which the protein functions. Geometric deep learning applied to context-specific PPI networks trained on single-cell atlases may help address this gap. For instance, PINNACLE [205] constructs cell-type–aware PPI graphs across tissues and learns *contextualized protein embeddings* that reflect cellular and tissue hierarchies, improving downstream tasks such as target nomination and drug effect analysis compared to context-free baselines. Integrating this type of embedding into biochemical-level models could make protein representations sensitive to the biological context, thereby narrowing plausible ligand–target interactions to those that are functionally relevant in specific tissues or cell types. At the same time, the Chemical Checker (CC) [90] shows that the principle of molecular similarity can be extended beyond chemical structure to a hierarchy of *bioactivity signatures* encompassing targets, pathways, cellular phenotypes, and clinical outcomes.

Contextual BindSight

Combining context-aware protein embeddings with multi-level compound signatures could create a unified representation in which biochemical and cellular contexts naturally intersect. Training BindSight on these enriched embeddings would test whether aligning context-sensitive protein spaces with biologically informed compound signatures enhances predictive performance. Even modest improvements in contextual coherence would suggest that incorporating biologically grounded inductive biases, link-

ing protein and compound representations through shared contextual structure, offers a practical path to extend current drug–target interaction models beyond purely molecular representations.

8.4.2 Incorporating structural and interaction priors into cellular representations

Protein-aware embeddings as structural priors

Another direction exploiting multi-scale integration methods would be to explore whether protein-aware cellular embeddings, inspired by SATURN [297], could provide CellHit with a principled way to incorporate structural and interaction priors into its cellular representations. SATURN combines single-cell transcriptomes with protein language model embeddings and learns higher-level features called "macrogenes." Each macrogene groups together genes that are functionally related in protein space, representing every cell as a nonlinear combination of these macrogenes. This design captures remote homology and shared molecular functions that may not be visible from expression data alone. Compared with conventional bulk RNA features, macrogene embeddings could capture more conserved biological programs underlying drug response such as coordinated regulation of pathways around drug targets or essential cellular processes—while being less sensitive to confounding factors like tumor purity or platform-specific effects. Because SATURN-like embeddings have also proven effective for label transfer and for identifying conserved and species-specific cell types, they could summarize intratumoral heterogeneity into functionally meaningful axes that facilitate patient stratification.

Macrogene space for cross-scale alignment

Embedding CellHit's inputs in macrogene space might enhance patient-level predictions for two main reasons. First, macrogene features encode implicit protein-level constraints on structure, function, and potential interaction neighborhoods that introduce biologically grounded inductive biases and help stabilize learning under distribution shift, such as when transferring from cell lines to tumors. Second, macrogene mappings are inherently robust to gene-panel mismatches and missing genes, as they do not depend on strict one-to-one gene correspondence. This could simplify the alignment between cell-line training data and heterogeneous patient bulk profiles that CellHit currently achieves with Celligner before inference. Under this formulation, CellHit would operate on cellular representations already enriched with protein-level priors, potentially improving both generalization and mechanistic coherence of its per-drug models.

Interpretability challenges and recovery strategies

A likely challenge of this approach would be interpretability. Although macrogenes are interpretable by design, their biological meaning may not always be clearly defined due to incomplete annotations or context-dependent functions. Preserving CellHit's explanatory power would therefore require additional interpretability layers. Possible strategies include computing SHAP or permutation importances at the macrogene level and mapping them back to genes through macrogene weights, projecting macrogenes onto curated pathway resources to recover mechanism-of-action explanations, and performing targeted perturbations in macrogene space to test the causal relevance of inferred programs.

8.4.3 Integrating CellHit and BindSight

One of the main strengths of the CellHit framework is its ability to identify which genes play the largest role in shaping a cancer cell line's response to a drug. BindSight, by contrast, operates at the biochemical level and focuses on identifying compounds that bind to a specific protein. Integrating the two frameworks could provide a deeper layer of interpretability for CellHit. While most drugs modeled by CellHit are FDA-approved compounds with well-characterized mechanisms of action, BindSight could help uncover potential off-target effects by linking influential genes to unexpected protein interactions. In addition, BindSight could be used to identify promising drug combinations. By examining which genes most strongly enhance or reduce the predicted efficacy of a given compound, as revealed by CellHit's SHAP-based interpretability, we could search chemical libraries for molecules that modulate those same genes, potentially improving therapeutic response.

8.4.4 Agentic integration across scales.

LLM agents as multi-scale orchestrators

Another orthogonal direction of multi-scale interaction can be represented by LLM-based agentic systems. Rather than being monolithic models where the integration phase occurs implicitly during training, these systems act as orchestrators that autonomously plan, write code, and invoke specialized analytical tools to complete a given task. Recent studies have demonstrated that such agents can dynamically design computational workflows, retrieve relevant resources, and execute domain-specific pipelines, combining retrieval-augmented reasoning with program synthesis to achieve complex biomedical objectives [150, 376]

Dynamic, on-demand multi-scale analysis

Within the context of drug–target interaction and drug response prediction, these capabilities could provide a new route toward automated, on-demand multi-scale integration. For instance, an agent could first query biochemical-level predictors such as docking or binding affinity models, then condition subsequent cellular-level analyses (e.g., transcriptomic perturbation modeling or pathway enrichment) on those results, and finally synthesize the outputs into a patient-level interpretation. Instead of predefining a fixed pipeline, the agent could iteratively refine its strategy, choosing tools, adjusting parameters, and cross-validating intermediate results to produce a coherent, reproducible multi-scale report.

Bridging modeling scales and enhancing interpretability

In principle, this flexibility could bridge traditionally disconnected modeling scales: molecular docking could inform the interpretation of transcriptional signatures, while cellular response models could guide prioritization of binding targets or drug combinations. Beyond technical integration, agentic systems may also enhance interpretability by explicitly documenting each reasoning step, thereby enabling auditability of cross-scale conclusions. While these ideas remain speculative, the rapid progress of biomedical agents suggests that autonomous systems capable of dynamically constructing and executing such multi-scale analyses may soon provide a practical framework for integrating drug–biosystem interactions across levels of biological organization. [150, 376]

Chapter 9

Conclusion

Thesis Motivation and Hypothesis

The challenge of translating preclinical discoveries into effective patient therapies remains a significant bottleneck in modern drug discovery. This thesis confronted this challenge by developing multiple computational framework, leveraging artificial intelligence to model drug-biosystem interactions from the biochemical level to the patient level. Our central hypothesis was that by creating accurate, interpretable, and accessible AI models, we could systematically navigate the vast chemical and biological landscapes, ultimately accelerating the identification of novel therapeutic strategies. The work presented herein validates this hypothesis through a series of integrated contributions that bridge computational prediction with experimental validation.

Biochemical Scale: BindSight

At the biochemical scale, we developed *BindSight*, a modular framework for drug–target interaction prediction that integrates data curation, representation learning, model evaluation, and scalable deployment. By combining rigorous scaffold-aware splitting and protein promiscuity stratification with a two-phase prediction strategy that first performs rapid library-wide screening and then applies TabPFN re-scoring, the framework ensures both efficiency and robust generalization. Its CLIP-style architecture allows proteins and compounds to be embedded in a shared latent space, while support for diverse molecular and protein representations, advanced loss functions, and distributed training makes the system adaptable to future methodological advances. While the results at this stage remain preliminary, they provide encouraging evidence that *BindSight* can serve as a solid foundation for future biochemical-level modeling and possibly integration with higher-scale analyses.

Cellular Scale: CellHit

At the *cellular scale*, we introduced *CellHit*, an interpretable framework that predicts drug responses from transcriptomic profiles of cancer cell lines and extends them to patient tumors. By training on large pharmacogenomic resources (GDSC, PRISM) and aligning them with patient bulk RNA-seq through Celligner, the framework uncovered transcriptional programs underpinning drug sensitivity and recovered known drug–target relationships. Incorporating LLM-curated mechanism-of-action pathways enhanced predictive power.

Translational Impact and Open Science

The translational potential of this framework was realized at the *patient scale*. By applying our models to over 10,000 patient transcriptomes from The Cancer Genome Atlas (TCGA), we successfully recovered a majority of approved drug-indication pairs, providing strong *in silico* validation. Importantly, we bridged the gap from computational hypotheses to experimental confirmation through prospective wet-lab experiments, which validated the novel vulnerabilities predicted by our models in pancreatic and glioblas-

toma cell lines. This shows the tangible real-world impact of our approach and its potential to guide preclinical research. To maximize the impact and accessibility of these methods, all the tools developed in this thesis have been released as open-source software and are available through a public web server.

Limitations and Future Directions

While this work represents a significant step forward, we acknowledge its limitations. Our models are primarily based on *in vitro* cell line data and rely on transcriptomics, which, while informative, captures only one dimension of cellular complexity. The integration of multi-omics data, including genomics, proteomics, and epigenomics, presents a clear and promising avenue for future work. Such an approach would enable a more holistic understanding of drug response and resistance. Further research should also focus on extending this framework to predict the efficacy of drug combinations and to model the dynamics of acquired resistance, two of the most pressing challenges in clinical oncology.

Cross-scale Integration and Orchestration

Beyond these directions, a broader challenge concerns the realization of a truly *end-to-end*, cross-scale framework. Current resources rarely provide *vertically profiled* cohorts that jointly capture biochemical features, cellular phenotypes, and longitudinal clinical outcomes, leading to systematic sparsity and missingness across scales. Yet genuine mechanistic explanations must flow across levels: biophysical constraints at the biochemical scale should propagate upward to shape cellular programs, while patient-level heterogeneity and treatment histories feed back to contextualize biochemical effects. Meeting this challenge will require modular, hierarchy-aware models capable of:

- representing multi-modal signals with both shared and scale-specific latent factors;
- learning under structured missingness and distribution shifts;
- propagating effects and uncertainty across scales;
- encoding causal assumptions to distinguish mechanism from correlation.

In this landscape, modern multi-modal LLM based agentic systems appear promising not as monolithic solvers but as *orchestrators*: they can plan analyses, call specialized tools (e.g., protein structure predictors, pathway simulators, deconvolution frameworks), reconcile outputs, and maintain provenance. Domain-specific solvers remain indispensable: no language model will infer protein folding or kinetic parameters from text alone. However, tool-augmented agents provide a path toward scalable, semi-automated orchestration of multi-scale reasoning while preserving mechanistic fidelity.

List of Figures

Bibliography

[1] Pharmacogenomic agreement between two cancer cell line data sets. *Nature*, 528 (7580):84–87, 2015. doi: 10.1038/nature15736.

[2] S. Abdallah, M. Sharifa, M. K. I. Almadhoun, M. M. Khawar, U. Shaikh, K. M. Balabel, I. Saleh, A. Manzoor, A. K. Mandal, O. Ekomwereren, W. M. Khine, and O. T. Oyelaja. The impact of artificial intelligence on optimizing diagnosis and treatment plans for rare genetic disorders. *Cureus*, 15, 2023. URL https://api.semanticscholar.org/CorpusId:265812574.

[3] A. Abid, M. J. Zhang, V. K. Bagaria, and J. Zou. Exploring patterns enriched in a dataset with contrastive principal component analysis. *Nature communications*, 9 (1):2134, 2018.

[4] J. Abramson, J. Adler, J. Dunger, and et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 2024. doi: 10.1038/s41586-024-07487-w.

[5] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[6] G. Adam, L. Rampášek, Z. Safikhani, P. Smirnov, B. Haibe-Kains, and A. Goldenberg. Machine learning approaches to drug response prediction: challenges and recent progress. *NPJ precision oncology*, 4(1):19, 2020.

[7] M. F. Adasme, K. L. Linnemann, S. N. Bolz, F. Kaiser, S. Salentin, V. J. Haupt, and M. Schroeder. Plip 2021: expanding the scope of the protein–ligand interaction profiler to dna and rna. *Nucleic Acids Research*, 49(W1):W530–W534, 2021. doi: 10.1093/nar/gkab294.

[8] F. E. Agamah et al. Network-based integrative multi-omics approach reveals biosignatures of covid-19 disease states. *Frontiers in Molecular Biosciences*, 11:1393240, 2024. doi: 10.3389/fmolb.2024.1393240. URL https://www.frontiersin.org/articles/10.3389/fmolb.2024.1393240/full.

[9] W. Ahmad, E. Simon, S. Chithrananda, G. Grand, and B. Ramsundar. Chemberta-2: Towards chemical foundation models, 2022. URL https://arxiv.org/abs/2209.01712.

[10] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama. Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2623–2631, 2019.

[11] A. Alexa, J. Rahnenführer, and T. Lengauer. Improved scoring of functional groups from gene expression data by decorrelating GO graph structure. *Bioinformatics*, 22(13):1600–1607, 2006. doi: 10.1093/bioinformatics/btl140.

[12] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402, 1997.

[13] D. W. Apley and J. Zhu. Visualizing the effects of predictor variables in black box supervised learning models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):1059–1086, 2020. doi: 10.1111/rssb.12377.

[14] R. Argelaguet, B. Velten, D. Arnol, S. Dietrich, T. Zenz, J. C. Marioni, F. Buettner, W. Huber, and O. Stegle. Multi-omics factor analysis—a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*, 14(6):e8124, 2018. doi: 10.15252/msb.20178124.

[15] R. Argelaguet, D. Arnol, D. Bredikhin, Y. Deloro, B. Velten, J. C. Marioni, and O. Stegle. Mofa+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biology*, 21(1):111, 2020. doi: 10.1186/s13059-020-02015-1.

[16] C. Arora, M. Matic, L. Bisceglia, P. Di Chiaro, N. D. O. Rosa, F. Carli, L. Clubb, L. A. N. Fard, G. Kargas, G. R. Diaferia, et al. The landscape of cancer-rewired gpcr signaling axes. *Cell Genomics*, 4(5), 2024.

[17] C. P. Austin, C. M. Cutillo, L. P. L. Lau, and et al. Future of rare diseases research 2017–2027: An irdirc perspective. *Clinical and Translational Science*, 11(1):21–27, 2018. doi: 10.1111/cts.12500.

[18] F. Azzarello, F. Carli, V. De Lorenzi, M. Tesi, P. Marchetti, F. Beltram, F. Raimondi, and F. Cardarelli. Machine-learning-guided recognition of α and β cells from label-free infrared micrographs of living human islets of langerhans. *Scientific Reports*, 14(1):14235, 2024.

[19] J. L. Ba, J. R. Kiros, and G. E. Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[20] M. Baek, F. DiMaio, I. Anishchenko, J. Dauparas, S. Ovchinnikov, G. R. Lee, J. Wang, Q. Cong, L. N. Kinch, R. D. Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557): 871–876, 2021.

[21] J. B. Baell and M. A. Walters. Chemical con artists foil drug discovery. *Nat. Rev. Drug Discov.*, 16(7):484–498, 2017. On PAINS and assay interference.

[22] M. H. Bailey, C. Tokheim, E. Porta-Pardo, S. Sengupta, D. Bertrand, A. Weerasinghe, A. Colaprico, and et al. Comprehensive characterization of cancer driver genes and mutations. *Cell*, 173(2):371–385.e18, 2018. doi: 10.1016/j.cell.2018.02.060.

[23] J. M. Banda, L. Evans, R. Vanguri, and et al. A curated and standardized adverse drug event resource to accelerate drug safety research. *Scientific Data*, 3:160026, 2016. doi: 10.1038/sdata.2016.26.

[24] A.-L. Barabási, N. Gulbahce, and J. Loscalzo. Network medicine: a network-based approach to human disease. *Nature Reviews Genetics*, 12(1):56–68, 2011.

[25] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, M. Kim, and others. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391):603–607, 2012. doi: 10.1038/nature11003.

[26] J. Barretina, G. Caponigro, N. Stransky, K. Venkatesan, A. A. Margolin, S. Kim, C. J. Wilson, J. Lehár, G. V. Kryukov, D. Sonkin, et al. The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, 483(7391): 603–607, 2012. doi: 10.1038/nature11003.

[27] A. Basu, N. E. Bodycombe, J. H. Cheah, E. V. Price, K. Liu, G. I. Schaefer, R. Y. Ebright, M. L. Stewart, D. Ito, S. Wang, et al. An interactive resource to identify cancer genetic and lineage dependencies targeted by small molecules. *Cell*, 154(5): 1151–1161, 2013.

[28] P. W. Battaglia, J. B. Hamrick, V. Bapst, A. Sanchez-Gonzalez, V. Zambaldi, M. Malinowski, A. Tacchetti, D. Raposo, A. Santoro, R. Faulkner, C. Gulcehre, F. Song, A. Ballard, J. Gilmer, G. Dahl, A. Vaswani, K. Allen, C. Nash, V. Langston, C. Dyer, N. Heess, D. Wierstra, P. Kohli, M. Botvinick, O. Vinyals, Y. Li, and R. Pascanu. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[29] E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44, 2019.

[30] A. Behdenna, M. Colange, J. Haziza, A. Gema, G. Appé, C.-A. Azencott, and A. Nordor. pycombat, a python tool for batch effects correction in high-throughput molecular data using empirical bayes methods. *BMC bioinformatics*, 24(1):459, 2023.

[31] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

[32] A. P. Bento, A. Hersey, E. Félix, G. Landrum, A. Gaulton, F. Atkinson, L. J. Bellis, M. De Veij, and A. R. Leach. An open source chemical structure curation pipeline using rdkit. *Journal of Cheminformatics*, 12(1):51, 2020.

[33] T. Bepler and B. Berger. Learning the protein language: Evolution, structure, and function. *Cell systems*, 12(6):654–669, 2021.

[34] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne. The protein data bank. *Nucleic acids research*, 28(1): 235–242, 2000.

[35] M. Binnewies, E. W. Roberts, K. Kersten, V. Chan, D. F. Fearon, M. Merad, L. M. Coussens, D. I. Gabrilovich, S. Ostrand-Rosenberg, C. C. Hedrick, et al. Understanding the tumor immune microenvironment (time) for effective therapy. *Nature medicine*, 24(5):541–550, 2018.

[36] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, and et al. On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258, 2021. URL https://arxiv.org/abs/2108.07258.

[37] M.-A. Bray, S. Singh, H. Han, C. T. Davis, B. Borgeson, C. Hartland, and others. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9):1757–1774, 2016. doi: 10.1038/nprot.2016.105.

[38] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[39] L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001. doi: 10.1023/A:1010933404324.

[40] P. Brennecke and et al. Accounting for technical noise in single-cell rna-seq experiments. *Nature Methods*, 10:1093–1095, 2013. doi: 10.1038/nmeth.2645. URL https://www.nature.com/articles/nmeth.2645.

[41] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[42] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1877–1901, 2020. URL https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[43] B. G. Buchanan and E. H. Shortliffe, editors. *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Addison-Wesley, Reading, MA, 1984. ISBN 978-0201101720.

[44] B. G. Buchanan and R. G. Smith. Fundamentals of expert systems. *Annual Review of Computer Science*, 2003.

[45] D. Buterez, I. Bica, I. Tariq, H. Andrés-Terré, and P. Liò. Cellvgae: an unsupervised scrna-seq analysis workflow with graph attention networks. *Bioinformatics*, 38(5): 1277–1286, 2022.

[46] Z. Cai, S. Apolinário, A. R. Baião, C. Pacini, M. D. Sousa, S. Vinga, R. R. Reddel, P. J. Robinson, M. J. Garnett, Q. Zhong, and E. Gonçalves. Synthetic augmentation of cancer cell line multi-omic datasets using unsupervised deep learning. *Nature Communications*, 15:10390, 2024. doi: 10.1038/s41467-024-54771-4.

[47] Cancer Genome Atlas Network. Comprehensive molecular portraits of human breast tumours. *Nature*, 490(7418):61–70, 2012. doi: 10.1038/nature11412.

[48] E. Cano-Gamez and G. Trynka. From gwas to function: Using functional genomics to identify the mechanisms underlying complex diseases. *Frontiers in Genetics*, 11, 2020. URL https://www.ncbi.nlm.nih.gov/pubmed/32477401.

[49] L. Cantini and et al. Benchmarking joint multi-omics dimensionality reduction methods for cancer study. *Nature Communications*, 12:124, 2021. doi: 10.1038/s41467-021-23896-0. URL https://www.nature.com/articles/s41467-021-23896-0.

[50] F. Carli, N. De Oliveira Rosa, S. Blotas, P. Di Chiaro, L. Bisceglia, M. Morelli, F. Lessi, A. L. Di Stefano, C. M. Mazzanti, G. Natoli, et al. Cellhit: a web server to predict and analyze cancer patients' drug responsiveness. *Nucleic Acids Research*, page gkaf414, 2025.

[51] F. Carli, P. Di Chiaro, M. Morelli, C. Arora, L. Bisceglia, N. De Oliveira Rosa, A. Cortesi, S. Franceschi, F. Lessi, A. L. Di Stefano, et al. Learning and actioning general principles of cancer cell drug sensitivity. *Nature Communications*, 16 (1):1654, 2025.

[52] E. Cerami, J. Gao, U. Dogrusoz, B. E. Gross, S. O. Sumer, B. A. Aksoy, A. Jacobsen, C. J. Byrne, M. L. Heuer, E. Larsson, Y. Antipin, B. Reva, A. P. Goldberg, C. Sander, and N. Schultz. The cbio cancer genomics portal: An open platform for exploring multidimensional cancer genomics data. *Cancer Discovery*, 2(5):401–404, 2012. doi: 10.1158/2159-8290.CD-12-0095.

[53] A. Chatterjee, R. Walters, Z. Shafi, O. S. Ahmed, M. Sebek, D. Gysi, R. Yu, T. Eliassi-Rad, A.-L. Barabási, and G. Menichetti. Improving the generalizability of protein-ligand binding predictions with ai-bind. *Nature communications*, 14(1):1989, 2023.

[54] S. Chawla, A. Rockstroh, M. Lehman, E. Ratther, A. Jain, A. Anand, A. Gupta, N. Bhattacharya, S. Poonia, P. Rai, et al. Gene expression based inference of cancer drug sensitivity. *Nature communications*, 13(1):5680, 2022.

[55] C. Chen et al. Applications of multi-omics analysis in human diseases. *MedComm*, ...:..., 2023. doi: 10.1002/mco2.211. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC10390758/.

[56] J. Chen and L. Zhang. A survey and systematic assessment of computational methods for drug response prediction. *Briefings in bioinformatics*, 22(1):232–246, 2021.

[57] J.-Y. Chen, J.-F. Wang, Y. Hu, X.-H. Li, Y.-R. Qian, and C.-L. Song. Evaluating the advancements in protein language models for encoding strategies in protein function prediction: a comprehensive review. *Frontiers in bioengineering and biotechnology*, 13:1506508, 2025.

[58] J. Y. Chen et al. A comprehensive review of protein language models. *Frontiers in Bioengineering and Biotechnology*, 2025.

[59] R. J. Chen and et al. Algorithm fairness in artificial intelligence for medicine and healthcare. *Nature Biomedical Engineering*, 7:1308–1325, 2023. doi: 10.1038/s41551-023-01171-3. URL https://www.nature.com/articles/s41551-023-01171-3.

[60] T. Chen and C. Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[61] Z. Chen, P. Zhao, F. Li, A. Leier, T. T. Marquez-Lago, Y. Wang, G. I. Webb, A. I. Smith, R. J. Daly, K.-C. Chou, et al. ifeature: a python package and web server for features extraction and selection from protein and peptide sequences. *Bioinformatics*, 34(14):2499–2502, 2018.

[62] Z. Chen, X. Liu, P. Zhao, C. Li, Y. Wang, F. Li, T. Akutsu, C. Bain, R. B. Gasser, J. Li, et al. ifeatureomega: an integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets. *Nucleic acids research*, 50(W1):W434–W447, 2022.

[63] T. Ching, D. S. Himmelstein, B. K. Beaulieu-Jones, and et al. Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface*, 15(141):20170387, 2018. doi: 10.1098/rsif.2017.0387.

[64] S. Chithrananda, G. Grand, and B. Ramsundar. Chemberta: large-scale self-supervised pretraining for molecular property prediction. *arXiv preprint arXiv:2010.09885*, 2020.

[65] H.-Y. Chuang, E. Lee, Y.-T. Liu, D. Lee, and T. Ideker. Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, 3:140, 2007. doi: 10.1038/msb4100180.

[66] M. M. Clark, A. Hildreth, S. Batalov, and et al. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Science Translational Medicine*, 11(489):eaat6177, 2019. doi: 10.1126/scitranslmed.aat6177.

[67] G. S. Collins and et al. Tripod+ai statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385: e078378, 2024. doi: 10.1136/bmj-2023-078378. URL https://www.bmj.com/content/385/bmj-2023-078378.

[68] G. E. R. Consortium and . G. P. Investigators. 100,000 genomes pilot on rare-disease diagnosis in a national health system. *New England Journal of Medicine*, 385(19): 1868–1880, 2021. doi: 10.1056/NEJMoa2035790.

[69] U. Consortium. Uniprot: a hub for protein information. *Nucleic acids research*, 43 (D1):D204–D212, 2015.

[70] S. M. Corsello, R. T. Nagari, R. D. Spangler, J. Rossen, M. Kocak, J. G. Bryan, R. Humeidi, D. Peck, X. Wu, A. A. Tang, et al. Discovering the anticancer potential of non-oncology drugs by systematic viability profiling. *Nature cancer*, 1 (2):235–248, 2020.

[71] S. M. Corsello, R. D. Spangler, R. Humeidi, C. N. Harrington, R. T. Nagari, R. Singh, V. Wang, M. Kocak, J. Rossen, A. Madec, et al. Adenosine receptor antagonists exhibit potent and selective off-target killing of foxa1-high cancers. *Cancer Research*, 80(16_Supplement):3400–3400, 2020.

[72] C. Cortes and V. Vapnik. Support-vector networks. *Machine Learning*, 20(3):273–297, 1995. doi: 10.1007/BF00994018.

[73] J. C. Costello, L. M. Heiser, E. Georgii, M. Gönen, M. P. Menden, N. J. Wang, and others. A community effort to assess and improve drug sensitivity prediction algorithms. *Nature Biotechnology*, 32(12):1202–1212, 2014. doi: 10.1038/nbt.2877.

[74] L. Cowen, T. Ideker, B. J. Raphael, and R. Sharan. Network propagation: a universal amplifier of genetic associations. *Nature Reviews Genetics*, 18(9):551–562, 2017. doi: 10.1038/nrg.2017.38.

[75] S. Cruz Rivera and et al. Guidelines for clinical trial protocols for interventions involving artificial intelligence: the spirit-ai extension. *Nature Medicine*, 26:1351–1363, 2020. doi: 10.1038/s41591-020-1037-7. URL https://www.nature.com/articles/s41591-020-1037-7.

[76] H. Cui, C. Wang, H. Maan, and et al. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, 21:1470–1480, 2024. doi: 10.1038/s41592-024-02201-0.

[77] I. Dagogo-Jack and A. T. Shaw. Tumour heterogeneity and resistance to cancer therapies. *Nature reviews Clinical oncology*, 15(2):81–94, 2018.

[78] Datamol-io. Datamol. https://docs.datamol.io/, .

[79] Datamol-io. Datamol. https://molfeat.datamol.io, .

[80] M. I. Davis, J. P. Hunt, S. Herrgard, P. Ciceri, L. M. Wodicka, G. Pallares, M. Hocker, D. K. Treiber, and P. P. Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature biotechnology*, 29(11):1046–1051, 2011.

[81] A. J. DeGrave, J. D. Janizek, and S.-I. Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3:610–619, 2021. doi: 10.1038/s42256-021-00338-7. URL https://www.nature.com/articles/s42256-021-00338-7.

[82] J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*, pages 4171–4186, 2019. URL https://aclanthology.org/N19-1423.

[83] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.

[84] S. Dimitrieva, R. Janssens, G. Li, A. Szalata, R. Gopalakrishnan, C. Parmar, A. Kauffmann, and E. Y. Durand. Biologically relevant integration of transcriptomics profiles from cancer cell lines, patient-derived xenografts, and clinical tumors using deep learning. *Science Advances*, 11(3):eadn5596, 2025.

[85] J. M. Doležal and et al. Uncertainty-informed deep learning models enable high-accuracy histopathology classification. *Nature Communications*, 13:6138, 2022. doi: 10.1038/s41467-022-34025-x. URL https://www.nature.com/articles/s41467-022-34025-x.

[86] P. Domingos. A few useful things to know about machine learning. *Communications of the ACM*, 55(10):78–87, 2012. doi: 10.1145/2347736.2347755.

[87] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou. The faiss library. *arXiv preprint arXiv:2401.08281*, 2024.

[88] A. Dugourd, C. Kuppe, M. Sciacovelli, E. Gjerga, A. Gabor, K. B. Emdal, V. Vieira, D. B. Bekker-Jensen, J. Kranz, E. M. Bindels, et al. Causal integration of multi-omics data with prior knowledge to generate mechanistic hypotheses. *Molecular Systems Biology*, 17(1):e9730, 2021.

[89] J. Durairaj, Y. Adeshina, Z. Cao, X. Zhang, V. Oleinikovas, T. Duignan, Z. McClure, X. Robin, G. Studer, D. Kovtun, et al. Plinder: The protein-ligand interactions dataset and evaluation resource. *bioRxiv*, pages 2024–07, 2024.

[90] M. Duran-Frigola, E. Pauls, O. Guitart-Pla, M. Bertoni, V. Alcalde, D. Amat, T. Juan-Blanco, and P. Aloy. Extending the small-molecule similarity principle to all levels of biology with the chemical checker. *Nature Biotechnology*, 38(9):1087–1096, 2020.

[91] D. K. Duvenaud, D. Maclaurin, J. Iparraguirre, R. Bombarell, T. Hirzel, A. Aspuru-Guzik, and R. P. Adams. Convolutional networks on graphs for learning molecular fingerprints. *Advances in neural information processing systems*, 28, 2015.

[92] A. Elnaggar, M. Heinzinger, C. Dallago, et al. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

[93] A. Elnaggar et al. Prottrans: Towards cracking the language of life's code through self-supervised learning. *bioRxiv*, 2021.

[94] ESM Team. ESM Cambrian: Revealing the Mysteries of Proteins with Unsupervised Learning. https://evolutionaryscale.ai/blog/esm-cambrian, 2024. (EvolutionaryScale Website).

[95] N. S. Eyke, W. Green, and K. Jensen. Iterative experimental design based on active machine learning reduces the experimental burden associated with reaction screening. *Reaction Chemistry & Engineering*, 2020. URL https://pdfs.semanticscholar.org/965d/7bb325ba2afccda3c630ff831a25e3522e98.pdf.

[96] A. Fang, Z. Zhang, A. Zhou, and M. Zitnik. Atomica: Learning universal representations of intermolecular interactions. *bioRxiv*, pages 2025–04, 2025.

[97] E. A. Feigenbaum. The art of artificial intelligence: I. themes and case studies of knowledge engineering. Technical report, Stanford University, Computer Science Department, 1977. URL https://stacks.stanford.edu/file/druid:bg342cm2034/bg342cm2034.pdf. IJCAI-5 report.

[98] W. Feng, L. Wang, Z. Lin, Y. Zhu, H. Wang, J. Dong, R. Bai, H. Wang, J. Zhou, W. Peng, et al. Generation of 3d molecules in pockets via a language model. *Nature Machine Intelligence*, 6(1):62–73, 2024.

[99] R. D. Finn, A. Bateman, J. Clements, et al. Pfam: the protein families database. *Nucleic Acids Research*, 42(D1):D222–D230, 2014. doi: 10.1093/nar/gkt1223.

[100] F. Firoozbakht, B. Yousefi, and B. Schwikowski. An overview of machine learning methods for monotherapy drug response prediction. *Briefings in bioinformatics*, 23(1), 2022.

[101] A. Fisher, C. Rudin, and F. Dominici. All models are wrong, but many are useful: Learning a variable's importance by studying an entire class of prediction models simultaneously. *Journal of Machine Learning Research*, 20(177):1–81, 2019.

[102] D. Fourches, E. Muratov, and A. Tropsha. Trust, but verify: On the importance of chemical structure curation in cheminformatics and qsar modeling research. *J. Chem. Inf. Model.*, 50(7):1189–1204, 2010. doi: 10.1021/ci100176x.

[103] D. Fourches, E. Muratov, and A. Tropsha. Trust, but verify ii: A practical guide to chemogenomics data curation. *J. Chem. Inf. Model.*, 56(7):1243–1252, 2016. doi: 10.1021/acs.jcim.6b00129.

[104] N. C. Frey, R. Soklaski, S. Axelrod, S. Samsi, R. Gomez-Bombarelli, C. W. Coley, and V. Gadepally. Neural scaling of deep chemical models. *Nature Machine Intelligence*, 5(11):1297–1305, 2023.

[105] J. H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001. doi: 10.1214/aos/1013203451.

[106] F. B. Fuchs, D. E. Worrall, V. Fischer, and M. Welling. Se(3)-transformers: 3d roto-translation equivariant attention networks. *arXiv preprint arXiv:2006.10503*, 2020. URL https://arxiv.org/abs/2006.10503.

[107] M. Gallagher and A. Chen-Plotkin. The post-gwas era: From association to function. *American journal of human genetics*, 102 5:717–730, 2018. URL https://www.ncbi.nlm.nih.gov/pubmed/29727686.

[108] J. Gao, B. A. Aksoy, U. Dogrusoz, G. Dresdner, B. Gross, S. O. Sumer, Y. Sun, A. Jacobsen, R. Sinha, E. Larsson, E. Cerami, C. Sander, and N. Schultz. Integrative analysis of complex cancer genomics and clinical profiles using the cbioportal. *Science Signaling*, 6(269):pl1, 2013. doi: 10.1126/scisignal.2004088.

[109] S. Gao, A. Fang, Y. Huang, V. Giunchiglia, A. Noori, J. R. Schwarz, Y. Ektefaie, J. Kondic, and M. Zitnik. Empowering biomedical discovery with ai agents. *Cell*, 187(22):6125–6151, 2024.

[110] M. J. Garnett, E. J. Edelman, S. J. Heidorn, A. Dastur, K. W. Lau, P. Greninger, and others. Systematic identification of genomic markers of drug sensitivity in cancer cells. *Nature*, 483(7391):570–575, 2012. doi: 10.1038/nature11005.

[111] T. Gaudelet, B. Day, A. R. Jamasb, J. Soman, C. Regep, G. Liu, J. B. Hayter, R. Vickers, C. Roberts, J. Tang, et al. Utilizing graph machine learning within drug discovery and development. *Briefings in bioinformatics*, 22(6):bbab159, 2021.

[112] A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, et al. Chembl: a large-scale bioactivity database for drug discovery. *Nucleic acids research*, 40(D1):D1100–D1107, 2012.

[113] A. Gayoso, R. Lopez, and et al. A joint model of rna expression and surface protein abundance in single cells. *Nature Methods*, 18:272–282, 2021. doi: 10.1038/s41592-020-01050-0. URL https://www.nature.com/articles/s41592-020-01050-0.

[114] R. Geirhos, J.-H. Jacobsen, C. Michaelis, R. Zemel, W. Brendel, M. Bethge, and F. A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[115] M. Ghandi, F. W. Huang, J. Jané-Valbuena, G. V. Kryukov, C. C. Lo, E. R. McDonald III, J. Barretina, E. T. Gelfand, C. M. Bielski, H. Li, et al. Next-generation characterization of the cancer cell line encyclopedia. *Nature*, 569(7757):503–508, 2019.

[116] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam. The false hope of current approaches to explainable artificial intelligence in health care. *The lancet digital health*, 3(11):e745–e750, 2021.

[117] J. Gilmer, S. S. Schoenholz, P. F. Riley, O. Vinyals, and G. E. Dahl. Neural message passing for quantum chemistry. In *International conference on machine learning*, pages 1263–1272. Pmlr, 2017.

[118] G. Goh, N. Cammarata, C. Voss, S. Carter, M. Petrov, L. Schubert, A. Radford, and C. Olah. Multimodal neurons in artificial neural networks. *Distill*, 6(3):e30, 2021.

[119] E. Gonçalves, R. C. Poulos, Z. Cai, S. Barthorpe, S. S. Manda, N. Lucas, A. Beck, D. Bucio-Noble, M. Dausmann, C. Hall, M. Hecker, J. Koh, H. Lightfoot, S. Mahboob, I. Mali, J. Morris, L. Richardson, A. J. Seneviratne, R. Shepherd, E. Sykes, F. Thomas, S. Valentini, S. G. Williams, Y. Wu, D. Xavier, K. L. MacKenzie, P. G. Hains, B. Tully, P. J. Robinson, Q. Zhong, M. J. Garnett, and R. R. Reddel. Pan-cancer proteomic map of 949 human cell lines. *Cancer Cell*, 40(8):835–849.e8, 2022. doi: 10.1016/j.ccell.2022.06.010.

[120] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio. *Deep learning*, volume 1. MIT press Cambridge, 2016.

[121] C. S. Greene, A. Krishnan, A. K. Wong, E. Ricciotti, R. A. Zelaya, D. S. Himmelstein, R. Zhang, B. M. Hartmann, E. Zaslavsky, S. C. Sealfon, et al. Understanding multicellular function and disease with human tissue-specific networks. *Nature Genetics*, 47(6):569–576, 2015.

[122] L. Grinsztajn, Oyallon, and G. Varoquaux. Why do tree-based models still outperform deep learning on typical tabular data? In *NeurIPS Workshops*, 2022.

[123] GTEx Consortium. The gtex consortium atlas of genetic regulatory effects across human tissues. *Science*, 369(6509):1318–1330, 2020. doi: 10.1126/science.aaz1776.

[124] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti. A survey of methods for explaining black box models. *ACM Computing Surveys*, 51 (5), 2018. doi: 10.1145/3236009.

[125] K. Guu, K. Lee, Z. Tung, P. Pasupat, and M. Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020.

[126] M. Hafner, M. Niepel, M. Chung, and P. K. Sorger. Growth rate inhibition metrics correct for confounders in measuring sensitivity to cancer drugs. *Nature Methods*, 13(6):521–527, 2016. doi: 10.1038/nmeth.3853.

[127] M. Hafner, M. Niepel, and P. K. Sorger. Alternative drug sensitivity metrics improve preclinical cancer pharmacogenomics. *Nature Biotechnology*, 35(6):500–502, 2017. doi: 10.1038/nbt.3882.

[128] B. Haibe-Kains, N. El-Hachem, N. J. Birkbak, A. C. Jin, A. H. Beck, H. J. W. L. Aerts, and J. Quackenbush. Inconsistency in large pharmacogenomic studies. *Nature*, 504 (7480):389–393, 2013. doi: 10.1038/nature12831.

[129] W. L. Hamilton. Graph representation learning. *Synthesis Lectures on Artifical Intelligence and Machine Learning*, 14(3):1–159, 2020.

[130] Y. Hasin, M. Seldin, and A. Lusis. Multi-omics approaches to disease. *Genome Biology*, 18:83, 2017. doi: 10.1186/s13059-017-1215-1.

[131] A. S. Hauser, D. E. Gloriam, H. Bräuner-Osborne, and S. R. Foster. Novel approaches leading towards peptide gpcr de-orphanisation. *British Journal of Pharmacology*, 177(5):961–968, 2020. doi: 10.1111/bph.14950.

[132] M. Hay, D. W. Thomas, J. L. Craighead, C. Economides, and J. Rosenthal. Clinical development success rates for investigational drugs. *Nature Biotechnology*, 32(1): 40–51, 2014. doi: 10.1038/nbt.2786.

[133] T. Hayes, R. Rao, H. Akin, and et al. Simulating 500 million years of evolution with a language model. *bioRxiv*, 2024. doi: 10.1101/2024.07.01.600583.

[134] K. He, X. Zhang, S. Ren, and J. Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 1026–1034. IEEE, 2015. doi: 10.1109/ICCV.2015.123.

[135] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[136] T. He, M. Heidemeyer, F. Ban, A. Cherkasov, and M. Ester. Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. *Journal of cheminformatics*, 9(1):24, 2017.

[137] M. Heinzinger et al. Bilingual language model for protein sequence and structure. *NAR Genomics and Bioinformatics*, 2024.

[138] D. Hendrycks and K. Gimpel. Gaussian error linear units (gelus). *arXiv:1606.08415*, 2016.

[139] M. Hidalgo, F. Amant, A. V. Biankin, E. Budinská, A. T. Byrne, C. Caldas, R. B. Clarke, S. de Jong, J. Jonkers, G. M. Mælandsmo, et al. Patient-derived xenograft models: an emerging platform for translational cancer research. *Cancer discovery*, 4(9):998–1013, 2014.

[140] M. Hidalgo, F. Amant, A. V. Biankin, E. Budinská, A. T. Byrne, C. Caldas, and others. Patient-derived xenograft models: An emerging platform for translational cancer research. *Cancer Discovery*, 4(9):998–1013, 2014. doi: 10.1158/2159-8290.CD-14-0009.

[141] K. A. Hoadley, C. Yau, T. Hinoue, D. M. Wolf, A. J. Lazar, E. Drill, R. Shen, and et al. Cell-of-origin patterns dominate the molecular classification of 10,000 tumors from 33 types of cancer. *Cell*, 173(2):291–304.e6, 2018. doi: 10.1016/j.cell.2018.03.022.

[142] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. d. L. Casas, L. A. Hendricks, J. Welbl, A. Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[143] M. Hofree, J. P. Shen, H. Carter, A. Gross, and T. Ideker. Network-based stratification of tumor mutations. *Nature Methods*, 10(11):1108–1115, 2013. doi: 10.1038/nmeth.2651.

[144] N. Hollmann, S. Müller, L. Purucker, A. Krishnakumar, M. Körfer, S. B. Hoo, R. T. Schirrmeister, and F. Hutter. Accurate predictions on small data with a tabular foundation model. *Nature*, 637:319–327, 2025. doi: 10.1038/s41586-024-08328-6.

[145] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

[146] Z. Hou et al. Learning the protein language of proteome-wide protein–protein interactions. *PMC*, 2023.

[147] W. Hu, B. Liu, J. Gomes, M. Zitnik, P. Liang, V. Pande, and J. Leskovec. Strategies for pre-training graph neural networks. *arXiv preprint arXiv:1905.12265*, 2019.

[148] K. Huang, T. Fu, W. Gao, Y. Zhao, Y. Roohani, J. Leskovec, C. W. Coley, C. Xiao, J. Sun, and M. Zitnik. Therapeutics data commons: Machine learning datasets and tasks for drug discovery and development. *arXiv preprint arXiv:2102.09548*, 2021.

[149] K. Huang, C. Xiao, L. M. Glass, and J. Sun. Moltrans: molecular interaction transformer for drug–target interaction prediction. *Bioinformatics*, 37(6):830–836, 2021.

[150] K. Huang, S. Zhang, H. Wang, Y. Qu, Y. Lu, Y. Roohani, R. Li, L. Qiu, G. Li, J. Zhang, et al. Biomni: A general-purpose biomedical ai agent. *biorxiv*, 2025.

[151] ICGC/TCGA Pan-Cancer Analysis of Whole Genomes Consortium. Pan-cancer analysis of whole genomes. *Nature*, 578:82–93, 2020. doi: 10.1038/s41586-020-1969-6.

[152] V. M. Ingram. Gene mutations in human haemoglobin: the chemical difference between normal and sickle-cell haemoglobin. *Nature*, 180(4581):326–328, 1957. doi: 10.1038/180326a0. URL https://www.nature.com/articles/180326a0.

[153] F. Iorio, T. A. Knijnenburg, D. J. Vis, G. R. Bignell, M. P. Menden, M. Schubert, N. Aben, E. Gonçalves, S. Barthorpe, H. Lightfoot, et al. A landscape of pharmacogenomic interactions in cancer. *Cell*, 166(3):740–754, 2016. doi: 10.1016/j.cell.2016.06.017.

[154] B. W. Irwin, J. R. Levell, T. M. Whitehead, M. D. Segall, and G. J. Conduit. Practical applications of deep learning to impute heterogeneous drug discovery data. *Journal of Chemical Information and Modeling*, 60(6):2848–2857, 2020.

[155] J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield, and R. A. Sayle. Zinc20—a free ultralarge-scale chemical database for ligand discovery. *J. Chem. Inf. Model.*, 2020. doi: 10.1021/acs.jcim.0c00675.

[156] G. Izacard and E. Grave. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*, 2020.

[157] S. Jaeger, S. Fulle, and S. Turk. Mol2vec: unsupervised machine learning approach with chemical intuition. *Journal of chemical information and modeling*, 58(1):27–35, 2018.

[158] S. Jain and B. C. Wallace. Attention is not explanation. *arXiv preprint arXiv:1902.10186*, 2019.

[159] B. Jassal, L. Matthews, G. Viteri, C. Gong, P. Lorente, A. Fabregat, K. Sidiropoulos, J. Cook, M. Gillespie, R. Haw, et al. The reactome pathway knowledgebase. *Nucleic acids research*, 48(D1):D498–D503, 2020.

[160] Y. Ji, L. Zhang, J. Wu, B. Wu, L.-K. Huang, T. Xu, Y. Rong, L. Li, J. Ren, D. Xue, H. Lai, S. Xu, J. Feng, W. Liu, P. Luo, S. Zhou, J. Huang, P. Zhao, and Y. Bian. Drugood: Out-of-distribution (ood) dataset curator and benchmark for ai-aided drug discovery - a focus on affinity prediction problems with noise annotations. *ArXiv*, abs/2201.09637, 2022. URL https://api.semanticscholar.org/CorpusId:261712937.

[161] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[162] A. Q. Jiang, A. Sablayrolles, A. Roux, A. Mensch, B. Savary, C. Bamford, D. S. Chaplot, D. d. l. Casas, E. B. Hanna, F. Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[163] M. Jiang, Z. Li, S. Zhang, S. Wang, X. Wang, Q. Yuan, and Z. Wei. Drug–target affinity prediction using graph neural network and contact maps. *RSC advances*, 10(35):20701–20712, 2020.

[164] Y. Jiang, G. Zhang, J. You, H. Zhang, R. Yao, H. Xie, L. Zhang, Z. Xia, M. Dai, Y. Wu, et al. Pocketflow is a data-and-knowledge-driven structure-based molecular generative model. *Nature Machine Intelligence*, 6(3):326–337, 2024.

[165] J. Jiménez-Luna, F. Grisoni, and G. Schneider. Drug discovery with explainable artificial intelligence. *Nature Machine Intelligence*, 2(10):573–584, 2020.

[166] W. Jin, R. Barzilay, and T. Jaakkola. Junction tree variational autoencoder for molecular graph generation. In *International conference on machine learning*, pages 2323–2332. PMLR, 2018.

[167] B. Jing, S. Eismann, P. Soni, and R. O. Dror. Learning from protein structure with geometric vector perceptrons. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=1FvkSpWosOl.

[168] D. T. Jones. Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology*, 292(2):195–202, 1999. doi: 10.1006/jmbi.1999.3091.

[169] H. M. Jones and K. Rowland-Yeo. Basic concepts in physiologically based pharmacokinetic modeling in drug discovery and development. *CPT: Pharmacometrics & Systems Pharmacology*, 2(8):e63, 2013. doi: 10.1038/psp.2013.41.

[170] H. M. Jones, Y. Chen, C. Gibson, T. Heimbach, N. Parrott, S. A. Peters, J. Snoeys, V. V. Upreti, M. Zheng, and S. D. Hall. Physiologically based pharmacokinetic modeling in drug discovery and development: a pharmaceutical industry perspective. *Clinical Pharmacology & Therapeutics*, 97(3):247–262, 2015. doi: 10.1002/cpt.37.

[171] N. P. Jouppi, C. Young, N. Patil, D. Patterson, G. Agrawal, et al. In-datacenter performance analysis of a tensor processing unit. In *44th Intl. Symposium on Computer Architecture (ISCA)*, pages 1–12, 2017. doi: 10.1145/3079856.3080246. URL https://arxiv.org/abs/1704.04760.

[172] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[173] J. Jumper, R. Evans, A. Pritzel, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2.

[174] Y. Kalakoti, S. Yadav, and D. Sundar. Transdti: transformer-based language models for estimating dtis and building a drug recommendation workflow. *ACS omega*, 7 (3):2706–2717, 2022.

[175] T. Kalliokoski, C. Kramer, A. Vulpetti, and P. Gedeck. Comparability of mixed ic50 data–a statistical analysis. *PloS one*, 8(4):e61007, 2013.

[176] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint*, 2020. URL https://arxiv.org/abs/2001.08361.

[177] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[178] S. Kapoor and A. Narayanan. Leakage and the reproducibility crisis in ml-based science. *Patterns*, 4(9):100747, 2023. doi: 10.1016/j.patter.2023.100747. URL https://www.sciencedirect.com/science/article/pii/S2666389923001599.

[179] K. J. Karczewski and M. P. Snyder. Integrative omics for health and disease. *Nature Reviews Genetics*, 19:299–310, 2018. doi: 10.1038/s41576-018-0024-7. URL https://www.nature.com/articles/s41576-018-0024-7.

[180] V. Karpukhin, B. Oguz, S. Min, P. S. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781, 2020.

[181] V. Karpukhin, B. Oguz, S. Min, P. S. Lewis, L. Wu, S. Edunov, D. Chen, and W.-t. Yih. Dense passage retrieval for open-domain question answering. In *EMNLP (1)*, pages 6769–6781, 2020.

[182] S. Kawashima and M. Kanehisa. Aaindex: amino acid index database. *Nucleic acids research*, 28(1):374–374, 2000.

[183] K. J. Kelleher, T. K. Sheils, S. L. Mathias, and et al. Pharos 2023: an integrated resource for the understudied human proteome. *Nucleic Acids Research*, 51(D1): D1405–D1416, 2023. doi: 10.1093/nar/gkac1033.

[184] P. Khatri, M. Sirota, and A. J. Butte. Ten years of pathway analysis: current approaches and outstanding challenges. *PLoS Computational Biology*, 8(2):e1002375, 2012. doi: 10.1371/journal.pcbi.1002375.

[185] G. S. Kinker, A. C. Greenwald, R. Tal, Z. Orlova, M. S. Cuoco, J. M. McFarland, A. Warren, C. Rodman, J. A. Roth, S. A. Bender, et al. Pan-cancer single-cell RNA-seq identifies recurring programs of cellular heterogeneity. *Nature Genetics*, 52 (11):1208–1218, 2020. doi: 10.1038/s41588-020-00726-6.

[186] T. N. Kipf and M. Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

[187] D. B. Kitchen, H. Decornez, J. R. Furr, and J. Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature Reviews Drug Discovery*, 3(11):935–949, 2004. doi: 10.1038/nrd1549.

[188] C. Knox, M. Wilson, C. M. Klinger, M. Franklin, E. Oler, A. Wilson, A. Pon, J. Cox, N. E. Chin, S. A. Strawbridge, et al. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic acids research*, 52(D1):D1265–D1275, 2024.

[189] I. Kola and J. Landis. Can the pharmaceutical industry reduce attrition rates? *Nature Reviews Drug Discovery*, 3(8):711–715, 2004. doi: 10.1038/nrd1470.

[190] B. Kompa, J. Snoek, and A. L. Beam. A (not so) short introduction to uncertainty in clinical machine learning. *npj Digital Medicine*, 4:39, 2021. doi: 10.1038/s41746-020-00367-3. URL https://www.nature.com/articles/s41746-020-00367-3.

[191] R. Krivák and D. Hoksza. P2rank: machine learning based tool for ligand binding site prediction. *Journal of Cheminformatics*, 10:39, 2018. doi: 10.1186/s13321-018-0281-2.

[192] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems 25 (NeurIPS 2012)*, pages 1097–1105. Curran Associates, Inc., 2012. URL https://papers.nips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.

[193] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012. URL https://proceedings.neurips.cc/paper/4824-imagenet-classification-with-deep-convolutional-neural-networks.pdf.

[194] M. Kuehl, D. P. Schaub, F. Carli, L. Heumos, C. Fernández-Zapata, N. Kaiser, J. Schaul, U. Panzer, S. Bonn, S. Lobentanzer, et al. Community-based biomedical context to unlock agentic systems. *bioRxiv*, pages 2025–07, 2025.

[195] M. Kuhn, C. von Mering, M. Campillos, L. J. Jensen, and P. Bork. Stitch: interaction networks of chemicals and proteins. *Nucleic acids research*, 36(suppl_1):D684–D688, 2007.

[196] R. Kundra, H. Zhang, R. Sheridan, S. J. Sirintrapun, A. Wang, A. Ochoa, M. Wilson, B. Gross, Y. Sun, R. Madupuri, et al. Oncotree: A cancer classification system for precision oncology. *JCO Clinical Cancer Informatics*, 5:221–230, 2021. doi: 10.1200/CCI.20.00108.

[197] D. Lähnemann, A. Kähäri, et al. Eleven grand challenges in single-cell data science. *Genome Biology*, 21:31, 2020.

[198] G. A. Landrum and S. Riniker. Combining ic50 or k i values from different sources is a source of significant noise. *Journal of chemical information and modeling*, 64 (5):1560–1567, 2024.

[199] V. Le Guilloux, P. Schmidtke, and P. Tuffery. Fpocket: An open source platform for ligand pocket detection. *BMC Bioinformatics*, 10:168, 2009. doi: 10.1186/1471-2105-10-168.

[200] Y. LeCun, Y. Bengio, and G. Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015. doi: 10.1038/nature14539.

[201] C. Y. Lee, D. Hubrich, J. K. Varga, C. Schäfer, M. Welzel, E. Schumbera, M. Djokic, J. M. Strom, J. Schönfeld, J. L. Geist, et al. Systematic discovery of protein interaction interfaces using alphafold and experimental validation. *Molecular Systems Biology*, 20(2):75–97, 2024.

[202] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[203] M. M. Li, K. Huang, and M. Zitnik. Graph representation learning in biomedicine. *arXiv preprint arXiv:2104.04883*, 2021.

[204] M. M. Li, K. Huang, and M. Zitnik. Graph representation learning in biomedicine and healthcare. *Nature biomedical engineering*, 6(12):1353–1369, 2022.

[205] M. M. Li, Y. Huang, M. Sumathipala, M. Q. Liang, A. Valdeolivas, A. N. Ananthakrishnan, K. Liao, D. Marbach, and M. Zitnik. Contextual ai models for single-cell protein biology. *Nature Methods*, 21(8):1546–1557, 2024.

[206] Q. Liao, Y. Zhang, Y. Chu, Y. Ding, Z. Liu, X. Zhao, Y. Wang, J. Wan, Y. Ding, P. Tiwari, et al. Application of artificial intelligence in drug-target interactions prediction: a review. *npj biomedical innovations*, 2(1):1, 2025.

[207] J. J. Lica. Effective drug concentration and selectivity depends on cell density and proliferation rate. *Pharmaceuticals (Basel)*, 14(8):8125035, 2021. doi: 10.3390/ph14080812. Highlights need for normalized selectivity index accounting for growth differences.

[208] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017.

[209] Z. Lin, H. Akin, R. Rao, and et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 2023. doi: 10.1126/science.ade2574.

[210] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, N. Smetanin, R. Verkuil, O. Kabeli, Y. Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[211] Z. Lin et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 2023.

[212] C. A. Lipinski, F. Lombardo, B. W. Dominy, and P. J. Feeney. Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. *Advanced drug delivery reviews*, 23(1-3):3–25, 1997.

[213] A. Liu, P. Trairatphisan, E. Gjerga, A. Didangelos, J. Barratt, and J. Saez-Rodriguez. From expression footprints to causal pathways: contextualizing large signaling networks with carnival. *NPJ systems biology and applications*, 5(1):40, 2019.

[214] Q. Liu, J. Zhang, C. Guo, M. Wang, C. Wang, Y. Yan, L. Sun, D. Wang, L. Zhang, H. Yu, L. Hou, C. Wu, Y. Zhu, G. Jiang, H. Zhu, Y. Zhou, S. Fang, T. Zhang, L. Hu, J. Li, Y. Liu, H. Zhang, B. Zhang, L. Ding, A. I. Robles, H. Rodriguez, D. Gao, H. Ji, H. Zhou, and P. Zhang. Proteogenomic characterization of small cell lung cancer identifies biological insights and subtype-specific therapeutic strategies. *Cell*, 187 (1):184–203.e28, 2024. doi: 10.1016/j.cell.2023.12.004.

[215] T. Liu, Y. Lin, X. Wen, R. N. Jorissen, and M. K. Gilson. Bindingdb: a web-accessible database of experimentally determined protein–ligand binding affinities. *Nucleic acids research*, 35(suppl_1):D198–D201, 2007.

[216] T. Liu, L. Hwang, S. K. Burley, C. I. Nitsche, C. Southan, W. P. Walters, and M. K. Gilson. Bindingdb in 2024: a fair knowledgebase of protein-small molecule binding data. *Nucleic acids research*, 53(D1):D1633–D1644, 2025.

[217] X. Liu and et al. Reporting guidelines for clinical trials of artificial intelligence interventions: the consort-ai extension. *BMJ*, 370:m3164, 2020. doi: 10.1136/bmj.m3164. URL https://www.bmj.com/content/370/bmj.m3164.

[218] S. Lobentanzer, P. Rodriguez-Mier, S. Bauer, and J. Saez-Rodriguez. Molecular causality in the advent of foundation models. *Molecular Systems Biology*, 20(8): 848–858, 2024.

[219] R. Lopez, J. Regier, M. B. Cole, M. I. Jordan, and N. Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058, 2018. doi: 10.1038/s41592-018-0229-2. URL https://www.nature.com/articles/s41592-018-0229-2.

[220] W. Lu, J. Zhang, W. Huang, and et al. Dynamicbind: predicting ligand-specific protein–ligand complex structure with a deep equivariant generative model. *Nature Communications*, 2024. doi: 10.1038/s41467-024-45461-2.

[221] M. D. Luecken and F. J. Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular Systems Biology*, 15(6):e8746, 2019. doi: 10.15252/msb.20188746. URL https://www.embopress.org/doi/10.15252/msb.20188746.

[222] S. M. Lundberg and S. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NeurIPS)*, 2017.

[223] S. M. Lundberg and S.-I. Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1705.07874.

[224] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[225] S. M. Lundberg, G. G. Erion, and S.-I. Lee. Consistent individualized feature attribution for tree ensembles. *arXiv preprint arXiv:1802.03888*, 2018.

[226] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. Explainable AI for trees: From local explanations to global understanding. *arXiv preprint arXiv:1905.04610*, 2019.

[227] S. M. Lundberg, G. Erion, H. Chen, A. DeGrave, J. M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S.-I. Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2:56–67, 2020. doi: 10.1038/s42256-019-0138-9.

[228] J. Lyu, S. Wang, T. E. Balius, I. Singh, A. Levit, Y. S. Moroz, M. J. O'Meara, T. Che, E. Algaa, K. Tolmachova, and others. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, 2019. doi: 10.1038/s41586-019-0917-9.

[229] J. Lyu, S. Wang, T. E. Balius, I. Singh, A. Levit, Y. S. Moroz, M. J. O'Meara, T. Che, E. Algaa, K. Tolmachova, et al. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, 2019.

[230] J. Ma, S. H. Fong, Y. Luo, C. J. Bakkenist, J. P. Shen, S. Mourragui, L. F. Wessels, M. Hafner, R. Sharan, J. Peng, et al. Few-shot learning creates predictive models of drug response that translate from high-throughput screens to individual patients. *Nature Cancer*, 2(2):233–244, 2021.

[231] L. v. d. Maaten and G. Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[232] C. A. MacRae and R. T. Peterson. Zebrafish as tools for drug discovery and precision medicine. *Nature Reviews Drug Discovery*, 14(10):721–731, 2015. doi: 10.1038/nrd4627.

[233] S. Mandal, A. Guzmán-Sáenz, N. Haiminen, S. Basu, and L. Parida. A topological data analysis approach on predicting phenotypes from gene expression data. In *International Conference on Algorithms for Computational Biology*, pages 178–187. Springer, 2020.

[234] M. Matic, G. Singh, F. Carli, N. De Oliveira Rosa, P. Miglionico, L. Magni, J. S. Gutkind, R. B. Russell, A. Inoue, and F. Raimondi. Precogx: ex ploring gpcr signaling mechanisms with deep protein representations. *Nucleic Acids Research*, 50 (W1):W598–W610, 2022.

[235] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[236] O. Méndez-Lucio, C. A. Nicolaou, and B. Earnshaw. Mole: a foundation model for molecular graphs using disentangled attention. *Nature Communications*, 15:9431, 2024. doi: 10.1038/s41467-024-53751-y.

[237] L. H. Mervin, A. M. Afzal, G. Drakakis, R. Lewis, O. Engkvist, and A. Bender. Target prediction utilising negative bioactivity data covering large chemical space. *Journal of cheminformatics*, 7(1):51, 2015.

[238] L. H. Mervin, M.-A. Trapotsi, A. M. Afzal, I. P. Barrett, A. Bender, and O. Engkvist. Probabilistic random forest improves bioactivity predictions close to the classification threshold by taking into account experimental uncertainty. *Journal of Cheminformatics*, 13(1):62, 2021.

[239] B. Mészáros, G. Erdős, and Z. Dosztányi. Iupred2a: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Research*, 46(W1):W329–W337, 2018. doi: 10.1093/nar/gky384.

[240] S. Min, B. Lee, and S. Yoon. Deep learning in bioinformatics. *Briefings in Bioinformatics*, 18(5):851–869, 2017. doi: 10.1093/bib/bbw068.

[241] T. M. Mitchell. *Machine Learning*. McGraw-Hill, New York, 1997. ISBN 978-0070428072.

[242] A. Mlinarić, M. Horvat, and V. Šupak Smolčić. Dealing with the positive publication bias: Why you should really publish your negative results. *Biochemia medica*, 27(3):447–452, 2017.

[243] H. L. Morgan. The generation of a unique machine description for chemical structures—a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, 1965. doi: 10.1021/c160017a018.

[244] P. Y. Muller and M. N. Milton. The determination and interpretation of the therapeutic index in drug development. *Nature Reviews Drug Discovery*, 11(10):751–761, 2012. doi: 10.1038/nrd3801.

[245] H. Najgebauer, M. Yang, H. E. Francies, C. Pacini, E. A. Stronach, M. J. Garnett, J. Saez-Rodriguez, and F. Iorio. Cellector: genomics-guided selection of cancer in vitro models. *Cell systems*, 10(5):424–432, 2020.

[246] A. C. Nascimento, R. B. Prudêncio, and I. G. Costa. A multiple kernel learning algorithm for drug-target interaction prediction. *BMC bioinformatics*, 17(1):46, 2016.

[247] J. Nasser, D. T. Bergman, C. Fulco, P. Guckelberger, B. R. Doughty, T. A. Patwardhan, T. Jones, T. H. Nguyen, J. Ulirsch, F. Lekschas, K. S. Mualim, H. Natri, E. M. Weeks, G. Munson, M. Kane, H. Kang, A. Cui, J. P. Ray, T. Eisenhaure, R. L. Collins, K. Dey, H. Pfister, A. Price, C. Epstein, A. Kundaje, R. Xavier, M. Daly, H. Huang, H. Finucane, N. Hacohen, E. Lander, and J. Engreitz. Genome-wide enhancer maps link risk variants to disease genes. *Nature*, 593:238 – 243, 2021. URL https://doi.org/10.1038/s41586-021-03446-x.

[248] M. Nauta, J. Trienes, S. Pathak, E. Nguyen, M. Peters, Y. Schmitt, J. Schlötterer, M. Van Keulen, and C. Seifert. From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable ai. *ACM Computing Surveys*, 55(13s):1–42, 2023.

[249] A. Newell and H. A. Simon. Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, 19(3):113–126, 1976. doi: 10.1145/360018.360022.

[250] A. M. Newman, C. B. Steen, C. L. Liu, A. J. Gentles, A. A. Chaudhuri, F. Scherer, M. S. Khodadoust, M. S. Esfahani, B. A. Luca, D. Steiner, M. Diehn, and A. A. Alizadeh. Determining cell type abundance and expression from bulk tissues with digital cytometry. *Nature Biotechnology*, 37(7):773–782, 2019. doi: 10.1038/s41587-019-0114-2.

[251] D.-T. Nguyen, S. Mathias, C. Bologa, S. Brunak, N. Fernandez, A. Gaulton, A. Hersey, J. Holmes, L. J. Jensen, A. Karlsson, et al. Pharos: collating protein information to shed light on the druggable genome. *Nucleic acids research*, 45(D1):D995–D1002, 2017.

[252] T. Nguyen, H. Le, T. P. Quinn, T. Nguyen, T. D. Le, and S. Venkatesh. Graphdta: predicting drug–target binding affinity with graph neural networks. *Bioinformatics*, 37(8):1140–1147, 2021.

[253] I. Nofech-Mozes, D. Soave, P. Awadalla, S. Abelson, et al. Pan-cancer classification of single cells in the tumour microenvironment. *Nature Communications*, 14:1615, 2023. doi: 10.1038/s41467-023-37353-8.

[254] D. P. Nusinow, J. Szpyt, M. Ghandi, C. M. Rose, E. R. McDonald, M. Kalocsay, J. Jané-Valbuena, E. Gelfand, D. K. Schweppe, M. Jedrychowski, et al. Quantitative proteomics of the cancer cell line encyclopedia. *Cell*, 180(2):387–402, 2020.

[255] N. Oak, A. D. Cherniack, R. J. Mashl, F. R. Hirsch, L. Ding, R. Beroukhim, Z. H. Gümüş, S. E. Plon, and K.-l. Huang. Ancestry-specific predisposing germline variants in cancer. *Genome medicine*, 12(1):51, 2020.

[256] L. Oakden-Rayner, J. Dunnmon, G. Carneiro, and C. Ré. Hidden stratification causes clinically meaningful failures in machine learning for medical imaging. *Proceedings of the ACM Conference on Health, Inference, and Learning / PNAS version*, 2020. URL https://www.pnas.org/doi/10.1073/pnas.2005938117.

[257] D. Ochoa, M. Karim, Bravo, B. Barwick, A. Campbell, B. Chen, N. George, G. Glusman, J. Gonzalez, S. Goodwin, A. Gunasekera, B. Hall, G. Hoxhaj, F. Hunter, A. Jene, W. Luo, C. Malangone, L. Martens, C. Miller, E. Papa, G. Peat, A. Raies, R. Salama, C. Sanderson, A. Sarangi, G. Saunders, S. Seager, O. Shamardina, W. C. Skarnes, K. Smith, E. Stephens, B. Sun, Y. Sun, D. Teixeira, L. Thompson, D. Torre, C. Xu, K. Yusa, E. M. McDonagh, I. Dunham, P. Flicek, and G. Koscielny. Open targets platform: Supporting systematic drug–target identification and prioritisation. *Nucleic Acids Research*, 51(D1):D1304–D1316, 2023. doi: 10.1093/nar/gkac1046.

[258] F. Offensperger, G. Tin, M. Duran-Frigola, E. Hahn, S. Dobner, C. W. a. Ende, J. W. Strohbach, A. Rukavina, V. Brennsteiner, K. Ogilvie, et al. Large-scale chemoproteomics expedites ligand discovery and predicts ligand behavior in cells. *Science*, 384(6694):eadk5864, 2024.

[259] H. Olsson and et al. Estimating diagnostic uncertainty in artificial intelligence for prostate biopsies using conformal prediction. *Nature Communications*, 13: 7097, 2022. doi: 10.1038/s41467-022-34945-8. URL https://www.nature.com/articles/s41467-022-34945-8.

[260] A. v. d. Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

[261] T. I. Oprea, C. G. Bologa, S. Brunak, A. Campbell, G. N. Gan, A. Gaulton, S. M. Gomez, R. Guha, A. Hersey, J. Holmes, and others. Unexplored therapeutic opportunities in the human genome. *Nature Reviews Drug Discovery*, 17(5):317–332, 2018. doi: 10.1038/nrd.2018.14.

[262] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35: 27730–27744, 2022.

[263] J. D. Owens, M. Houston, D. Luebke, S. Green, J. E. Stone, and J. C. Phillips. Gpu computing. *Proceedings of the IEEE*, 96(5):879–899, 2008. doi: 10.1109/JPROC.2008.917757.

[264] H. Öztürk, A. Özgür, and E. Ozkirimli. Deepdta: deep drug–target binding affinity prediction. *Bioinformatics*, 34(17):i821–i829, 2018.

[265] C. Pacini, E. Duncan, E. Gonçalves, J. Gilbert, S. Bhosle, S. Horswell, E. Karakoc, H. Lightfoot, E. Curry, F. Muyas, et al. A comprehensive clinically informed map

of dependencies in cancer cells and framework for target prioritization. *Cancer Cell*, 42(2):301–316, 2024.

[266] T. Pahikkala, A. Airola, S. Pietilä, S. Shakyawar, A. Szwajda, J. Tang, and T. Aittokallio. Toward more realistic drug–target interaction predictions. *Briefings in bioinformatics*, 16(2):325–337, 2015.

[267] J. J. H. Park, E. Siden, M. J. Zoratti, L. Dron, O. Harari, J. Singer, R. T. Lester, K. Thorlund, and E. J. Mills. Systematic review of basket trials, umbrella trials, and platform trials: a landscape analysis of master protocols. *Trials*, 20(1):572, 2019.

[268] S. Passaro, G. Corso, J. Wohlwend, M. Reveiz, S. Thaler, V. R. Somnath, N. Getz, T. Portnoi, J. Roy, H. Stark, et al. Boltz-2: Towards accurate and efficient binding affinity prediction. *BioRxiv*, pages 2025–06, 2025.

[269] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

[270] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, and et al. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3):203–214, 2010. doi: 10.1038/nrd3078.

[271] S. M. Paul, D. S. Mytelka, C. T. Dunwiddie, C. C. Persinger, B. H. Munos, S. R. Lindborg, and A. L. Schacht. How to improve r&d productivity: the pharmaceutical industry's grand challenge. *Nature Reviews Drug Discovery*, 9(3):203–214, 2010. doi: 10.1038/nrd3078.

[272] L. Pauling, H. A. Itano, S. J. Singer, and I. C. Wells. Sickle cell anemia, a molecular disease. *Science*, 110(2865):543–548, 1949. doi: 10.1126/science.110.2865.543. URL https://www.science.org/doi/10.1126/science.110.2865.543.

[273] L. Peng, W. Zhu, B. Liao, Y. Duan, M. Chen, Y. Chen, and J. Yang. Screening drug-target interactions with positive-unlabeled learning. *Scientific reports*, 7(1):8087, 2017.

[274] X. Peng, S. Luo, J. Guan, Q. Xie, J. Peng, and J. Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. In *International conference on machine learning*, pages 17644–17655. PMLR, 2022.

[275] T.-H. Pham, Y. Qiu, J. Zeng, L. Xie, and P. Zhang. A deep learning framework for high-throughput mechanism-driven phenotype compound screening and its application to covid-19 drug repurposing. *Nature machine intelligence*, 3(3):247–257, 2021.

[276] J. C. Pino, C. Posso, S. K. Joshi, M. Nestor, J. Moon, J. R. Hansen, C. Hutchinson-Bunch, M. A. Gritsenko, K. K. Weitz, K. Watanabe-Smith, N. Long, J. E. McDermott, B. J. Druker, T. Liu, J. W. Tyner, A. Agarwal, E. Traer, P. D. Piehowski, C. E. Tognon, K. D. Rodland, and S. J. C. Gosline. Mapping the proteogenomic landscape enables

prediction of drug response in acute myeloid leukemia. *Cell Reports Medicine*, 5 (1):101359, 2024. doi: 10.1016/j.xcrm.2023.101359.

[277] F. Pognan, A. Galetin, et al. The evolving role of investigative toxicology in the pharmaceutical industry. *Nature Reviews Drug Discovery*, 22:317–335, 2023. doi: 10.1038/s41573-022-00633-x.

[278] P. G. Polishchuk, T. I. Madzhidov, and A. Varnek. Estimation of the size of drug-like chemical space based on gdb-17 data. *J. Comput.-Aided Mol. Des.*, 27(8):675–679, 2013. doi: 10.1007/s10822-013-9672-4.

[279] A. B. Popejoy and S. M. Fullerton. Genomics is failing on diversity. *Nature*, 538 (7624):161–164, 2016. doi: 10.1038/538161a. URL https://www.nature.com/articles/538161a.

[280] Z. Qiao, W. Nie, A. Vahdat, T. F. Miller, and A. Anandkumar. State-specific protein–ligand complex structure prediction with a multiscale deep generative model. *Nature Machine Intelligence*, 6:195–208, 2024. doi: 10.1038/s42256-024-00792-z.

[281] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. Improving language understanding by generative pre-training. Technical report, OpenAI, 2018. URL https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

[282] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[283] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[284] R. Rafailov, A. Sharma, E. Mitchell, C. D. Manning, S. Ermon, and C. Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741, 2023.

[285] N. Rappoport and R. Shamir. Multi-omic and multi-view clustering algorithms: review and cancer benchmark. *Nucleic acids research*, 46(20):10546–10562, 2018.

[286] M. Remmert, A. Biegert, A. Hauser, and J. Söding. Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nature methods*, 9(2):173–175, 2012.

[287] P. Ren, Y. Xiao, X. Chang, P.-Y. Huang, Z. Li, B. B. Gupta, X. Chen, and X. Wang. A survey of deep active learning. *ACM computing surveys (CSUR)*, 54(9):1–40, 2021.

[288] J.-L. Reymond. The chemical space project. *Acc. Chem. Res.*, 48(3):722–730, 2015. doi: 10.1021/ar500432k.

[289] M. T. Ribeiro, S. Singh, and C. Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1135–1144, 2016. doi: 10.1145/2939672.2939778.

[290] A. S. Rifaioglu, E. Nalbat, V. Atalay, M. J. Martin, R. Cetin-Atalay, and T. Doğan. Deepscreen: high performance drug–target interaction prediction with convolutional neural networks using 2-d structural compound representations. *Chemical science*, 11(9):2531–2557, 2020.

[291] I. Rivals, L. Personnaz, L. Taing, and M.-C. Potier. Enrichment or depletion of a go category within a class of genes: which test? *Bioinformatics*, 23(4):401–407, 2007. doi: 10.1093/bioinformatics/btl633.

[292] A. Rives, J. Meier, T. Sercu, S. Goyal, Z. Lin, J. Liu, D. Guo, M. Ott, C. L. Zitnick, J. Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[293] P. Rodriguez-Mier, M. Garrido-Rodriguez, A. Gabor, and J. Saez-Rodriguez. Unifying multi-sample network inference from prior knowledge and omics data with corneto. *Nature Machine Intelligence*, pages 1–19, 2025.

[294] D. Rogers and M. Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

[295] D. Rogers and M. Hahn. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

[296] Y. Rong, Y. Bian, T. Xu, W. Xie, Y. Wei, W. Huang, and J. Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571, 2020.

[297] Y. Rosen, M. Brbić, Y. Roohani, K. Swanson, Z. Li, and J. Leskovec. Toward universal cell embeddings: integrating single-cell rna-seq datasets across species with saturn. *Nature Methods*, 21(8):1492–1500, 2024.

[298] F. Rosenblatt. The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6):386–408, 1958.

[299] S. Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

[300] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215, 2019. doi: 10.1038/s42256-019-0048-x.

[301] C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5): 206–215, 2019.

[302] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[303] A. Saadat. Fine-tuning protein language models to understand the ... *ScienceDirect*, 2025.

[304] A. V. Sadybekov and V. Katritch. Computational approaches streamlining drug discovery. *Nature*, 616:673–685, 2023. doi: 10.1038/s41586-023-05905-z.

[305] J. Saez-Rodriguez, L. Wessels, et al. Toward better benchmarking: challenge-based methods assessment in cancer genomics. *Genome Biology*, 15(9):1–12, 2014.

[306] T. Sainburg, L. McInnes, and T. Q. Gentner. Parametric umap embeddings for representation and semisupervised learning. *Neural Computation*, 33(11):2881–2907, 2021.

[307] T. Saito and M. Rehmsmeier. The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets. *PloS one*, 10(3):e0118432, 2015.

[308] S. Salentin, S. Schreiber, V. J. Haupt, M. F. Adasme, and M. Schroeder. Plip: fully automated protein–ligand interaction profiler. *Nucleic Acids Research*, 43(W1): W443–W447, 2015. doi: 10.1093/nar/gkv315.

[309] S.-J. Sammut, M. Crispin-Ortuzar, S.-F. Chin, E. Provenzano, H. A. Bardwell, W. Ma, W. Cope, A. Dariush, S.-J. Dawson, J. E. Abraham, J. Dunn, L. Hiller, J. Thomas, D. A. Cameron, J. M. S. Bartlett, L. Hayward, P. D. Pharoah, F. Markowetz, O. M. Rueda, H. M. Earl, and C. Caldas. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*, 601(7894):623–629, 2022. doi: 10.1038/s41586-021-04278-5.

[310] B. Sánchez-Lengeling and A. Aspuru-Guzik. Inverse molecular design using machine learning: Generative models for matter engineering. *Science*, 361(6400):360–365, 2018. doi: 10.1126/science.aat2663.

[311] F. Sanchez-Vega, M. Mina, J. Armenia, W. K. Chatila, A. Luna, K. La, S. Dimitriadoy, C. Liu, H. S. Kantheti, S. Saghafinia, D. Chakravarty, F. Daian, J. Gao, M. H. Bailey, W.-W. Liang, S. M. Foltz, I. Shmulevich, L. Ding, Z. Heins, A. Ochoa, B. Gross, Q. Gao, et al. Oncogenic signaling pathways in the cancer genome atlas. *Cell*, 173 (2):321–337.e10, 2018. doi: 10.1016/j.cell.2018.03.035.

[312] R. Santos, O. Ursu, A. Gaulton, A. P. Bento, R. S. Donadi, C. G. Bologa, A. Karlsson, B. Al-Lazikani, A. Hersey, T. I. Oprea, and J. P. Overington. A comprehensive map of molecular drug targets. *Nature Reviews Drug Discovery*, 16(1):19–34, 2017. doi: 10.1038/nrd.2016.230.

[313] J. W. Scannell, A. Blanckley, H. Boldon, and B. Warrington. Diagnosing the decline in pharmaceutical r&d efficiency. *Nature Reviews Drug Discovery*, 11(3):191–200, 2012. doi: 10.1038/nrd3681.

[314] L. V. Schaffer and T. Ideker. Mapping the multiscale structure of biological systems. *Cell Systems*, 12(6):622–635, 2021.

[315] L. V. Schaffer and T. Ideker. Mapping the multiscale structure of biological systems. *Cell Systems*, 12(6):622–635, 2021.

[316] R. Schmirler, M. Heinzinger, and B. Rost. Fine-tuning protein language models boosts predictions across diverse tasks. *Nature Communications*, 15(1):7407, 2024.

[317] P. Schneider, W. P. Walters, A. T. Plowright, N. Sieroka, J. Listgarten, R. A. Goodnow, J. Fisher, J. M. Jansen, J. S. Duca, T. S. Rush, M. Zentgraf, J. E. Hill, E. Krutoholow, M. Kohler, J. Blaney, K. Funatsu, C. Luebkemann, and G. Schneider. Rethinking drug design in the artificial intelligence era. *Nature Reviews Drug Discovery*, 19(5):353–364, 2020. doi: 10.1038/s41573-019-0050-3.

[318] A. Schneuing, C. Harris, Y. Du, K. Didi, A. Jamasb, I. Igashov, W. Du, C. Gomes, T. L. Blundell, P. Lio, et al. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4(12):899–909, 2024.

[319] A. M. Schnoes, D. C. Ream, A. W. Thorman, P. C. Babbitt, and I. Friedberg. Biases in the experimental annotations of protein function and their effect on our understanding of protein function space. *PLoS computational biology*, 9(5):e1003063, 2013.

[320] B. Seashore-Ludlow, M. G. Rees, J. H. Cheah, M. Cokol, E. V. Price, M. E. Coletti, V. Jones, N. E. Bodycombe, C. K. Soule, J. Gould, et al. Harnessing connectivity in a large-scale small-molecule sensitivity dataset. *Cancer discovery*, 5(11):1210–1223, 2015.

[321] M. H. S. Segler, M. Preuss, and M. P. Waller. Planning chemical syntheses with deep neural networks and symbolic ai. *Nature*, 555(7698):604–610, 2018. doi: 10.1038/nature25978.

[322] M. L. Shahreza, M. Ghadiri, S. Mousavi, J. Varshosaz, and K. Alirezaei. A review of network-based approaches to drug repositioning. *Briefings in Bioinformatics*, 19 (5):878–892, 2018. doi: 10.1093/bib/bbw020.

[323] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.

[324] T. Sheils, S. L. Mathias, V. B. Siramshetty, G. Bocci, C. G. Bologa, J. J. Yang, A. Waller, N. Southall, D.-T. Nguyen, and T. I. Oprea. How to illuminate the druggable genome using pharos. *Current protocols in bioinformatics*, 69(1):e92, 2020.

[325] T. K. Sheils, S. L. Mathias, V. B. Siramshetty, and et al. Tcrd and pharos 2021: mining the human proteome for disease biology. *Nucleic Acids Research*, 49(D1): D1334–D1346, 2021. doi: 10.1093/nar/gkaa993.

[326] R. P. Sheridan. Time-split cross-validation as a method for estimating the goodness of prospective prediction. *Journal of Chemical Information and Modeling*, 53(4): 783–790, 2013. doi: 10.1021/ci400084k.

[327] R. H. Shoemaker. The nci60 human tumour cell line anticancer drug screen. *Nature Reviews Cancer*, 6(10):813–823, 2006.

[328] R. Shwartz-Ziv and A. Armon. Tabular data: Deep learning is not all you need. *Information Fusion*, 2022. arXiv preprint 2106.03253.

[329] A. Singh, C. P. Shannon, B. Gautier, F. Rohart, M. Vacher, S. J. Tebbutt, and K.-A. Lê Cao. Diablo: an integrative approach for identifying key molecular drivers from multi-omics assays. *Bioinformatics*, 35(17):3055–3062, 2019. doi: 10.1093/bioinformatics/bty1054.

[330] R. Singh, S. Sledzieski, B. Bryson, L. Cowen, and B. Berger. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proceedings of the National Academy of Sciences*, 120(24):e2220778120, 2023.

[331] G. Sirugo, S. M. Williams, and S. A. Tishkoff. The missing diversity in human genetic studies. *Cell*, 177(1):26–31, 2019. doi: 10.1016/j.cell.2019.02.048. URL https://www.cell.com/cell/fulltext/S0092-8674(19)30451-9.

[332] J. S. Smith, N. Lubbers, O. Isayev, B. T. Nebgen, and A. E. Roitberg. Less is more: sampling chemical space with active learning. *The Journal of chemical physics*, 148 24:241733, 2018. URL https://api.semanticscholar.org/CorpusId:4682180.

[333] M. Smyth and J. Martin. x ray crystallography. *Molecular Pathology*, 53(1):8, 2000.

[334] C. D. Steele, A. Abbasi, S. A. Islam, A. L. Bowes, A. Khandekar, K. Haase, S. Hames-Fathi, D. Ajayi, A. Verfaillie, P. Dhami, et al. Signatures of copy number alterations in human cancer. *Nature*, 606(7916):984–991, 2022.

[335] Z. D. Stephens, S. Y. Lee, F. Faghri, and et al. Big data: Astronomical or genomical? *PLOS Biology*, 13(7):e1002195, 2015. doi: 10.1371/journal.pbio.1002195.

[336] J. M. Stokes, K. Yang, K. Swanson, and et al. A deep learning approach to antibiotic discovery. *Cell*, 180(4):688–702.e13, 2020. doi: 10.1016/j.cell.2020.01.021.

[337] D. Stumpfe and J. Bajorath. Exploring activity cliffs in medicinal chemistry: miniperspective. *Journal of medicinal chemistry*, 55(7):2932–2942, 2012.

[338] D. Stumpfe, H. Hu, and J. Bajorath. Evolving concept of activity cliffs. *ACS omega*, 4(11):14360–14368, 2019.

[339] V. Subbiah, R. J. Kreitman, Z. A. Wainberg, A. Gazzah, U. Lassen, A. Stein, P. Y. Wen, S. Dietrich, M. J. de Jonge, J.-Y. Blay, et al. Dabrafenib plus trametinib in brafv600e-mutated rare cancers: the phase 2 roar trial. *Nature medicine*, 29(5): 1103–1112, 2023.

[340] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, A. Paulovich, S. L. Pomeroy, T. R. Golub, E. S. Lander, and J. P. Mesirov. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43): 15545–15550, 2005. doi: 10.1073/pnas.0506580102.

[341] A. Subramanian, R. Narayan, S. M. Corsello, D. Peck, T. Natoli, X. Lu, and others. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, 171(6):1437–1452.e17, 2017. doi: 10.1016/j.cell.2017.10.049.

[342] M. Sundararajan, A. Taly, and Q. Yan. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR, 2017.

[343] R. S. Sutton. The bitter lesson. http://www.incompleteideas.net/IncIdeas/BitterLesson.html, 2019.

[344] M. Sypetkowski, F. Wenkel, F. Poursafaei, N. Dickson, K. Suri, P. Fradkin, and D. Beaini. On the scalability of gnns for molecular graphs. *Advances in Neural Information Processing Systems*, 37:19870–19906, 2024.

[345] D. Szklarczyk, A. Santos, C. Von Mering, L. J. Jensen, P. Bork, and M. Kuhn. Stitch 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic acids research*, 44(D1):D380–D384, 2016.

[346] J. Tanevski, R. O. R. Flores, A. Gabor, D. Schapiro, and J. Saez-Rodriguez. Explainable multiview framework for dissecting spatial relationships from highly multiplexed data. *Genome biology*, 23(1):97, 2022.

[347] J. Tang, A. Szwajda, S. Shakyawar, T. Xu, P. Hintsanen, K. Wennerberg, and T. Aittokallio. Making sense of large-scale kinase inhibitor bioactivity data sets: a comparative and integrative analysis. *Journal of chemical information and modeling*, 54(3):735–743, 2014.

[348] S. Tarazona, A. Arzalluz-Luque, and A. Conesa. Undisclosed, unmet and neglected challenges in multi-omics studies. *Nature Computational Science*, 1(6):395–402, 2021.

[349] K. Team, Y. Bai, Y. Bao, G. Chen, J. Chen, N. Chen, R. Chen, Y. Chen, Y. Chen, Y. Chen, et al. Kimi k2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025.

[350] M. A. Thafar, A. Raies, S. Albaradei, M. Essack, and V. Bajic. Comparison study of computational prediction tools for drug-target binding affinities. *Frontiers in Chemistry*, 7, 2019. URL https://api.semanticscholar.org/CorpusId:208163893.

[351] The ENCODE Project Consortium. Expanded encyclopaedias of dna elements in the human and mouse genomes. *Nature*, 583:699–710, 2020. doi: 10.1038/s41586-020-2493-4.

[352] The Gene Ontology Consortium. The gene ontology resource: enriching a gold mine. *Nucleic Acids Research*, 49(D1):D325–D334, 2021. doi: 10.1093/nar/gkaa1113.

[353] C. Thiel, I. Smit, V. Baier, H. Cordes, B. Fabry, L. M. Blank, and L. Kuepfer. Using quantitative systems pharmacology to evaluate the drug efficacy of cox-2 and 5-lox inhibitors in therapeutic situations. *npj Systems Biology and Applications*, 4:28, 2018. doi: 10.1038/s41540-018-0062-3.

[354] B. I. Tingle, K. G. Tang, M. Castañon, J. J. Gutiérrez, M. Khurelbaatar, C. Dandarchuluun, Y. S. Moroz, and J. J. Irwin. Zinc-22—a free multi-billion-scale database of tangible compounds for ligand discovery. *J. Chem. Inf. Model.*, 63(4):1166–1176, 2023. doi: 10.1021/acs.jcim.2c01253.

[355] I. Tirosh and M. L. Suvà. Cancer cell states: Lessons from ten years of single-cell rna-sequencing of human tumors. *Cancer Cell*, 42(9):1497–1506, 2024. doi: 10.1016/j.ccell.2024.08.005.

[356] R. Todeschini and V. Consonni. Similarity coefficients for binary chemoinformatics data: Overview and extended comparison using simulated and real data sets. *Journal of Chemical Information and Modeling*, 48(3):413–424, 2008. doi: 10.1021/ci700409p.

[357] E. J. Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature Medicine*, 25(1):44–56, 2019. doi: 10.1038/s41591-018-0300-7.

[358] K. A. Tran, O. Kondrashova, A. Bradley, and et al. Deep learning in cancer diagnosis, prognosis and treatment selection. *Genome Medicine*, 13(1):152, 2021. doi: 10.1186/s13073-021-00968-x.

[359] L. Trastulla, J. Noorbakhsh, F. Vazquez, J. McFarland, and F. Iorio. Computational estimation of quality and clinical relevance of cancer cell lines. *Molecular Systems Biology*, 18(7):e11017, 2022. doi: 10.15252/msb.202211017.

[360] A. Tsherniak, F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, S. Gill, W. F. Harrington, S. Pantel, J. M. Krill-Burger, and et al. Defining a cancer dependency map. *Cell*, 170(3):564–576.e16, 2017. doi: 10.1016/j.cell.2017.06.010.

[361] A. Tsherniak, F. Vazquez, P. G. Montgomery, B. A. Weir, G. Kryukov, G. S. Cowley, and others. Defining a cancer dependency map. *Cell*, 170(3):564–576.e16, 2017. doi: 10.1016/j.cell.2017.06.010.

[362] A. M. Tsimberidou, E. Fountzilas, M. Nikanjam, and R. Kurzrock. Review of precision cancer medicine: Evolution of the treatment paradigm. *Cancer treatment reviews*, 86:102019, 2020.

[363] R. M. Turner, B. K. Park, and M. Pirmohamed. Parsing interindividual drug variability: an emerging role for systems pharmacology. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine*, 7(4):221–241, 2015.

[364] J. Vamathevan, D. Clark, P. Czodrowski, and et al. Applications of machine learning in drug discovery and development. *Nature Reviews Drug Discovery*, 18(6): 463–477, 2019. doi: 10.1038/s41573-019-0024-5.

[365] M. Varadi, S. Anyango, M. Deshpande, et al. Alphafold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Research*, 50(D1):D439–D444, 2022. doi: 10.1093/nar/gkab1061.

[366] M. Varadi, D. Bertoni, P. Magana, et al. Alphafold protein structure database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Research*, 52(D1):D368–D375, 2024. doi: 10.1093/nar/gkad1011.

[367] G. Varoquaux and V. Cheplygina. Machine learning for medical imaging. *npj Digital Medicine*, 5:119, 2022. doi: 10.1038/s41746-022-00592-y. URL https://www.nature.com/articles/s41746-022-00592-y.

[368] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[369] P. Veličković. Theoretical foundations of graph neural networks. CST Wednesday Seminar, 2021.

[370] P. Veličković, G. Cucurull, A. Casanova, A. Romero, P. Lio, and Y. Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

[371] L. C. Vieira et al. Medium-sized protein language models perform well at many downstream tasks. *Scientific Reports*, 2025.

[372] G. Vlachogiannis, S. Hedayat, A. Vatsiou, Y. Jamin, J. Fernández-Mateos, K. Khan, A. Lampis, K. Eason, I. Huntingford, R. Burke, et al. Patient-derived organoids model treatment response of metastatic gastrointestinal cancers. *Science*, 359 (6378):920–926, 2018.

[373] I. Wallach and A. Heifets. Most ligand-based classification benchmarks reward memorization rather than generalization. *Journal of chemical information and modeling*, 58(5):916–932, 2018.

[374] B. Wang, A. M. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg. Similarity network fusion for aggregating data types on a genomic scale. *Nature methods*, 11(3):333–337, 2014.

[375] C. Wang, X. Lye, R. Kaalia, P. Kumar, and J. C. Rajapakse. Deep learning and multi-omics approach to predict drug responses in cancer. *BMC bioinformatics*, 22(Suppl 10):632, 2021.

[376] E. Wang, S. Schmidgall, P. F. Jaeger, F. Zhang, R. Pilgrim, Y. Matias, J. Barral, D. Fleet, and S. Azizi. Txgemma: Efficient and agentic llms for therapeutics. *arXiv preprint arXiv:2504.06196*, 2025.

[377] J. Wang, A. Ma, Y. Chang, J. Gong, Y. Jiang, R. Qi, C. Wang, H. Fu, Q. Ma, and D. Xu. scgnn is a novel graph neural network framework for single-cell rna-seq analyses. *Nature communications*, 12(1):1–11, 2021.

[378] L. Wang, Y. Wu, Y. Deng, B. Kim, L. Pierce, G. Krilov, D. Lupyan, S. Robinson, M. K. Dahlgren, J. Greenwood, et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society*, 137 (7):2695–2703, 2015.

[379] R. Wang, X. Fang, Y. Lu, C.-Y. Yang, and S. Wang. The pdbbind database: methodologies and updates. *Journal of medicinal chemistry*, 48(12):4111–4119, 2005.

[380] A. Warren, Y. Chen, A. Jones, T. Shibue, W. C. Hahn, J. S. Boehm, F. Vazquez, A. Tsherniak, and J. M. McFarland. Global computational alignment of tumor and cell line transcriptional profiles. *Nature communications*, 12(1):22, 2021.

[381] S. Watanabe. Tree-structured parzen estimator: Understanding its algorithm components and their roles for better empirical performance. *arXiv preprint arXiv:2304.11127*, 2023.

[382] G. P. Way and C. S. Greene. Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders. In *Pacific Symposium on Biocomputing*, volume 23, pages 80–91, 2018. doi: 10.1142/9789813235533_0008.

[383] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*, 2021.

[384] K. Weissenow and B. Rost. Are protein language models the new universal key? *Current Opinion in Structural Biology*, 2025.

[385] T. Widiandani et al. Cytotoxic activity against breast cancer cell t47d and its comparison to normal (vero) cells: Selectivity index calculation. *Molecules*, 2023. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10365674/. Selectivity index (SI 2 defines tumour selectivity) calculated from IC values in cancer vs normal cell lines.

[386] D. S. Wishart, C. Knox, A. C. Guo, S. Shrivastava, M. Hassanali, P. Stothard, Z. Chang, and J. Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl_1):D668–D672, 2006.

[387] D. S. Wishart, Y. D. Feunang, A. C. Guo, and et al. Drugbank 5.0: a major update to the drugbank database for 2018. *Nucleic Acids Research*, 46(D1):D1074–D1082, 2018. doi: 10.1093/nar/gkx1037.

[388] C. R. Woese and G. E. Fox. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proceedings of the National Academy of Sciences*, 74(11): 5088–5090, 1977.

[389] A. Wonkam. Sequence three million genomes across africa. *Nature*, 590:209–211, 2021. doi: 10.1038/d41586-021-00313-7. URL https://www.nature.com/articles/d41586-021-00313-7.

[390] J. Woodcock and L. M. LaVange. Master protocols to study multiple therapies, multiple diseases, or both. *New England Journal of Medicine*, 377(1):62–70, 2017.

[391] O. J. Wouters, M. McKee, and J. Luyten. Estimated research and development investment needed to bring a new medicine to market, 2009–2018. *JAMA*, 323(9): 844–853, 2020. doi: 10.1001/jama.2020.1166.

[392] Y. Wu and F. Tang. scextract: leveraging large language models for fully automated single-cell rna-seq data annotation and prior-informed multi-dataset integration. *Genome Biology*, 26:174, 2025. doi: 10.1186/s13059-025-03639-x.

[393] Z. Wu, B. Ramsundar, E. N. Feinberg, J. Gomes, C. Geniesse, A. S. Pappu, K. Leswing, and V. Pande. Moleculenet: a benchmark for molecular machine learning. *Chemical Science*, 9(2):513–530, 2018. doi: 10.1039/C7SC02664A.

[394] F. Xia, J. Allen, P. Balaprakash, T. Brettin, C. Garcia-Cardona, A. Clyde, J. Cohn, J. Doroshow, X. Duan, V. Dubinkina, et al. A cross-study analysis of drug response prediction in cancer cell lines. *Briefings in bioinformatics*, 23(1):bbab356, 2022.

[395] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

[396] A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

[397] J. Yang, Q. Wang, Z.-Y. Zhang, L. Long, R. Ezhilarasan, J. M. Karp, A. Tsirigos, M. Snuderl, B. Wiestler, W. Wick, et al. Dna methylation-based epigenetic signatures predict somatic genomic alterations in gliomas. *Nature communications*, 13 (1):4410, 2022.

[398] K. Yang, K. Swanson, W. Jin, C. Coley, P. Eiden, H. Gao, A. Guzman-Perez, T. Hopper, B. Kelley, M. Mathea, et al. Analyzing learned molecular representations for property prediction. *Journal of chemical information and modeling*, 59(8):3370–3388, 2019.

[399] Z. Ying, D. Bourgeois, J. You, M. Zitnik, and J. Leskovec. Gnnexplainer: Generating explanations for graph neural networks. *Advances in neural information processing systems*, 32, 2019.

[400] K. M. Yip, N. Fischer, E. Paknia, A. Chari, and H. Stark. Atomic-resolution protein structure determination by cryo-em. *Nature*, 587(7832):157–161, 2020.

[401] M. D. Young, M. J. Wakefield, G. K. Smyth, and A. Oshlack. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biology*, 11(2):R14, 2010. doi: 10.1186/gb-2010-11-2-r14.

[402] Y. Yu and et al. Assessing and mitigating batch effects in large-scale omics studies. *Genome Biology*, 25:164, 2024. doi: 10.1186/s13059-024-03401-9. URL https://genomebiology.biomedcentral.com/articles/10.1186/s13059-024-03401-9.

[403] Y. Yuan and Z. Bar-Joseph. Gcng: graph convolutional networks for inferring gene interaction from spatial transcriptomics data. *Genome biology*, 21(1):1–16, 2020.

[404] C. Yung-Chi and W. H. Prusoff. Relationship between the inhibition constant (ki) and the concentration of inhibitor which causes 50 per cent inhibition (i50) of an enzymatic reaction. *Biochemical pharmacology*, 22(23):3099–3108, 1973.

[405] B. Zdrazil, E. Felix, F. Hunter, E. J. Manners, J. Blackshaw, S. Corbett, M. De Veij, H. Ioannidis, D. M. Lopez, J. F. Mosquera, et al. The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic acids research*, 52(D1):D1180–D1192, 2024.

[406] A. Zeng, X. Lv, Q. Zheng, Z. Hou, B. Chen, C. Xie, C. Wang, D. Yin, H. Zeng, J. Zhang, et al. Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*, 2025.

[407] Y. Zeng, J. Xie, Z. Wei, Y. Su, and Y. Yang. Cellfm: a large-scale foundation model pre-trained on transcriptomics of 100 million human cells. *Nature Communications*, 16:Article 59926–5, 2025. doi: 10.1038/s41467-025-59926-5.

[408] C. Zhang, X. Zhang, L. Freddolino, and Y. Zhang. Biolip2: an updated structure database for biologically relevant ligand–protein interactions. *Nucleic Acids Research*, 52(D1):D404–D412, 2024. doi: 10.1093/nar/gkad630.

[409] Y. Zhang, G. Parmigiani, and W. E. Johnson. Combat-seq: batch effect adjustment for rna-seq count data. *NAR Genomics and Bioinformatics*, 2(3):lqaa078, 2020. doi: 10.1093/nargab/lqaa078. URL https://academic.oup.com/nargab/article/2/3/lqaa078/5909519.

[410] A. Zhavoronkov, Y. A. Ivanenkov, A. Aliper, and et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature Biotechnology*, 37(9):1038–1040, 2019. doi: 10.1038/s41587-019-0224-x.

[411] Y. Zheng, Y. Liu, J. Yang, L. Dong, R. Zhang, S. Tian, Y. Yu, L. Ren, W. Hou, F. Zhu, et al. Multi-omics data integration using ratio-based quantitative profiling with quartet reference materials. *Nature biotechnology*, 42(7):1133–1149, 2024.

[412] H. Zhu. Big data and artificial intelligence modeling for drug discovery. *Annual Review of Pharmacology and Toxicology*, 60:573–589, 2020. doi: 10.1146/annurev-pharmtox-010919-023324.

[413] M. Zitnik, M. Agrawal, and J. Leskovec. Modeling polypharmacy side effects with graph convolutional networks. *Bioinformatics*, 34(13):i457–i466, 2018.

[414] M. Zitnik, F. Nguyen, B. Wang, J. Leskovec, A. Goldenberg, and M. M. Hoffman. Machine learning for integrating data in biology and medicine: Principles, practice, and opportunities. *Information Fusion*, 50:71–91, 2019.