Chemical Science

Accepted Manuscript

This article can be cited before page numbers have been issued, to do this please use: L. Naef and M. Bronstein, *Chem. Sci.*, 2026, DOI: 10.1039/D6SC01189F.

ROYAL SOCIETY OF CHEMISTRY

rsc.li/chemical-science

Black-Box Data: A new paradigm for biomedicine in the AI era

Luca Naef † and Michael Bronstein ‡,§,†

†Proxima Bio, USA ‡University of Oxford, UK §AITHYRA, Austria

In the past decade, we have witnessed a dramatic expansion of Artificial Intelligence (AI) and Machine Learning (ML) into a broad range of industries and scientific domains, from natural language processing (NLP) and computer vision (CV) to particle physics and chemistry. AI has beaten human players in games such as Go[1] or Starcraft[2], previously thought to be unconquerable. Most recently, it achieved top human-level performance in international math and programming olympiads[3].

What characterizes these achievements is that they are done in software. In experimental fields like the life sciences and biology, where predictions typically need to be validated by a lab experiment, the number of AI successes has so far been more limited. DeepMind's AlphaFold2 (AF2), recognized by the 2024 Nobel Prize in Chemistry, stands out as perhaps the most prominent example, revolutionizing protein monomer structure prediction and, more recently, protein binder design[4].

We have recently hypothesized[5] that AF2's success stems partly from the significant degeneracy of protein structure space: the large, diverse PDB may provide good coverage of naturally occurring folds[6–8]. Supporting this, recent large-scale applications of AF2 predicted only a single new fold across known protein sequences[9], suggesting either limited generalizability or, alternatively, that few novel folds remain to be discovered. Studies on other modalities corroborate the importance of data coverage: performance on protein-protein and protein-ligand complexes, where training data covers less of the interaction space, show strong dependence on similarity to training examples[10–12].

The AlphaFold breakthrough has spurred ambition to replicate this success elsewhere in biology to produce "the next AlphaFold". But it may be more critical to consider "the next PDB," i.e., new experimental data sources[1] for training future biological ML[13]. The PDB

[1] We would like to emphasize our distinction between a *data set* — a fixed collection of data produced by some experimental technology (such as X-ray crystallography or cryogenic electron microscopy, in the case of the PDB) — and a *data source*, a process or technological capability to produce new data for purposes that might not have been envisaged at

originated in the 1970s, predating modern ML and its relevance to biology. It represents five decades of effort by thousands of structural biologists at an estimated cost of up to $50 billion[14]. Building comparable datasets for new domains requires experimental approaches that bridge orders-of-magnitude gaps in cost and throughput.

Taking a historical perspective on algorithmic advances, from first-principles "white-box" models to present-day "black-box" deep learning, we explore the emergence of data sources optimized for ML rather than human consumption. Just as trading handcrafted features for learned representations drove breakthroughs in CV and NLP, we argue that an analogous tradeoff in data generation could unlock the scale needed for the next generation of biological foundation models. We call this paradigm "black-box data" and present a unifying taxonomy[2] with historical and recent examples.

Software 2.0, The Bitter Lesson and Black-Box Methods

The 2012 “ImageNet moment” is generally considered the Rubicon when end-to-end deep learning started to progressively overtake human-designed algorithms across a wide spectrum of domains. ImageNet was a large-scale computer vision competition consisting of classifying images of everyday objects from 1000 classes (such as cats, dogs, etc.), a notoriously hard challenge in the field. At the time, most state-of-the-art computer vision algorithms were based on extracting carefully hand-crafted features from images. In 2012, for the first time, the competition was won by AlexNet, a method that completely removed the human interpretable features in favor of data-driven features directly learned by a deep Convolutional Neural Network[15].

This transition was dubbed the "Software 2.0" paradigm by Andrej Karpathy[16]. In traditional “Software 1.0,” humans explicitly program algorithms that operate on data. Software 2.0 is programmed through data: a neural network learns from data examples without explicit

the time of the technology’s development (e.g. cross-linking mass-spectrometry was not conceived to be used for generative ML).

[2] Here, we use the term "taxonomy" referring to different *experimental data generation strategies*, rather than human-curated classification of biological entities and relationships (e.g., the Gene Ontology). A defining feature of the "Software 2.0" and black-box data paradigms is that models often bypass the need for rigid, predefined biological taxonimies or ontologies. Instead of relying on human-annotated labels, modern neural architectures learn their own implicit representations and latent spaces directly from raw, uncurated experimental readouts, allowing the data itself to dictate the structural relationships.

View Article Online
DOI: 10.1039/D5SC01189F

instructions, resulting in a black-box system whose rules are encoded implicitly in weights rather than interpretable code. In the ImageNet example, traditional computer vision approaches ("Software 1.0") attempted to derive an "equation of a cat" via hand-crafted features distinguishing cats from dogs. Modern deep-learning based approaches ("Software 2.0") instead present many labeled examples to a neural network, which learns the distinction on its own.

Fig. 1: Schematic overview of the Software 2.0 paradigm introduced with black-box algorithms.

Over the past decade, the abandonment of handcrafted inductive biases and features towards ever more general learning methods has been a predominant trend, often referred to as "The Bitter Lesson": simple systems that scale with available computational power will eventually outperform more complex systems relying on human knowledge[17]. One could perhaps see black-box data as the natural continuation of this trend beyond algorithms, by removing handcrafted human biases also from the data generation process.

Protein structure prediction methods are a prime example of this trend. Early works on protein folding exploited first-principle physics-based models. AlphaFold1 (2018), DeepMind's first foray into the field, still combined three separate deep learning models to predict distances and confidence, and then folded proteins with predicted distances given "handcrafted" Multiple Sequence Alignments (MSAs)[18] . AlphaFold2 (2021) was a single end-to-end trained model with strong domain-specific inductive biases, including a geometric frame-representation and roto-translation-invariant and chirally-aware Frame-Aligned Point Error (FAPE), geometry-inspired triangle operations, and a roto-translation Invariant Point Attention (IPA)[19]. AlphaFold3 (AF3, 2024) removed most inductive biases (IPA, frame representation, FAPE, and invariant featurization) in favor of data augmentation to learn the required invariances with a general Transformer-inspired architecture[20]. Yet, the further expected simplification has not

View Article Online
DOI: 10.1039/D6SC01189F

yet happened: AF3 still notoriously relies on MSA and a geometry-inspired architecture. More "general" methods that remove MSA and/or triangle operations (ESMFold, SimpleFold, ESM3) perform significantly worse[21–23]. Hence, while this has been a trend, it is not yet a *fait accompli*.

Black-Box AI methods require AI-first data sources

One of the unique aspects of the PDB is that it provided not only a large, clean dataset covering many folds occurring in nature, but also a clear scientific question with a benchmark (CASP competition) allowing to measure progress. This is far from being the case in many other life science domains, such as single-cell biology, protein-ligand complexes[10], protein-protein complexes[11], disease biology, human phenotypes, and other biological subdomains[24,25]. Not only is there less homogenous and complete data, but in some cases, even the question one tries to answer from the data is not well defined.

ML models in biology have traditionally relied on data produced as a *byproduct* of scientific inquiry. Repositories like ChEMBL[26], PDB[27], PRIDE[28], GEO[29], OpenTargets[30] aggregate and standardize this data, but they are not purpose-generated for ML. This leads to poor standardization, limited scale, and systematic biases. Unlike the diverse naturally occurring protein structures in the PDB, protein-drug complexes, for example, are often derived from drug discovery campaigns biased toward a narrow set of medically relevant targets. Freitas et al. found that of 11,016 high-quality protein-ligand structures spanning over 500 families, more than 45% came from just ten families of enzymes and receptors[31]. This lack of diversity contributes to ML performance on protein-ligand complexes significantly trailing protein monomers[20].

Since existing data sources are optimized for human-driven discovery, they have tradeoffs often suboptimal for ML. In traditional drug discovery, for example, deep coverage of individual targets is critical, whereas ML methods likely require diverse targets[32]. Humans generally prefer small, low-noise datasets for deductive reasoning, while ML is noise-tolerant but sample-inefficient, benefiting more from large noisy datasets. Rolnick et al. demonstrated that models achieve over 90% accuracy on MNIST even with 100 mislabeled examples for every true label, but a minimum absolute number of correct labels is critical[33]. The adage

View Article Online
DOI: 10.1039/D6SC01189F

"quality over quantity" is inverted in the black-box era: "quantity *is* quality," as single-cell pioneer Aviv Regev puts it[34].

Current foundation models bear this out. Large language models are pretrained on massive, diverse, noisy web-crawled corpora (C4, CommonCrawl[35]), then fine-tuned on small high-quality datasets via reinforcement learning from human feedback[36]. Intentional noise injection is also broadly useful in ML for improving generalization[37]. Denoising diffusion models[38] take this further, training explicitly on data synthetically noised to varying levels. This also allows naturally noisy experimental data to be integrated at higher-noise training steps, where inherent measurement noise becomes indistinguishable from synthetic noise[39].

Another problem with "data as a byproduct of scientific inquiry" is the lack of *true negatives*. Partly a consequence of the academic incentive and publication model, unsuccessful experiments are rarely reported, also known as the "file drawer problem"[40]. Yet such data is critical for AI[41].

Above all, the cost of generating datasets such as the PDB is likely in the billions,[14] making a fundamental paradigm shift on biological data generation all but necessary.

Black-Box Data: a paradigm of AI-first, scalable data sources

In recent years, we observe the emergence of black-box data sources that focus directly on ML consumption. The leitmotif is moving away from tradeoffs optimized for human consumption – high signal-to-noise protocols, expensive low-noise readouts, limited diversity and human interpretability – enables orders of magnitude improvements in throughput. We can see a clear analogy to the emergence of black-box algorithms – trading off human interpretability in favor of scalability optimized for ML. In the most extreme version of this paradigm, such datasets can become completely unintelligible to humans, hence our term "black-box data"[5].

Computationally, the defining characteristic of black-box data is the complexity of the inverse problem separating the raw measurement from the quantity of interest. In white-box data, this mapping is simple, often identity or a known analytical transformation. In black-box data, reconstruction requires algorithms with high representational complexity (whether classical optimization or learned mappings) and strong priors drawn from domain knowledge or adjacent

View Article Online
DOI: 10.1039/D6SC01189F

examples. The distinction is not binary, illustrated in Figure 2. The degree of "black-boxness" corresponds to how much algorithmic machinery, and how strong the priors, must sit between measurement and useful output. At one end is simply the DL-based denoising of assay outputs (i.e., output data structure is fully preserved), e.g. Direct-to-Biology (D2B)[42]. At the other are measurements used purely as pretraining signal, where the relationship to downstream outputs is learned implicitly and not fully known explicitly (e.g., chemical probing reactivity as latent supervision signal for RNA structure prediction[43]). Hence "black-box data" is defined by emphasizing trade-offs that optimize for ML rather than the lack of interpretability – similar to state-of-the-art "black-box models," i.e., neural networks that often still contain interpretable elements such as attention matrices or equivariant layers[44].

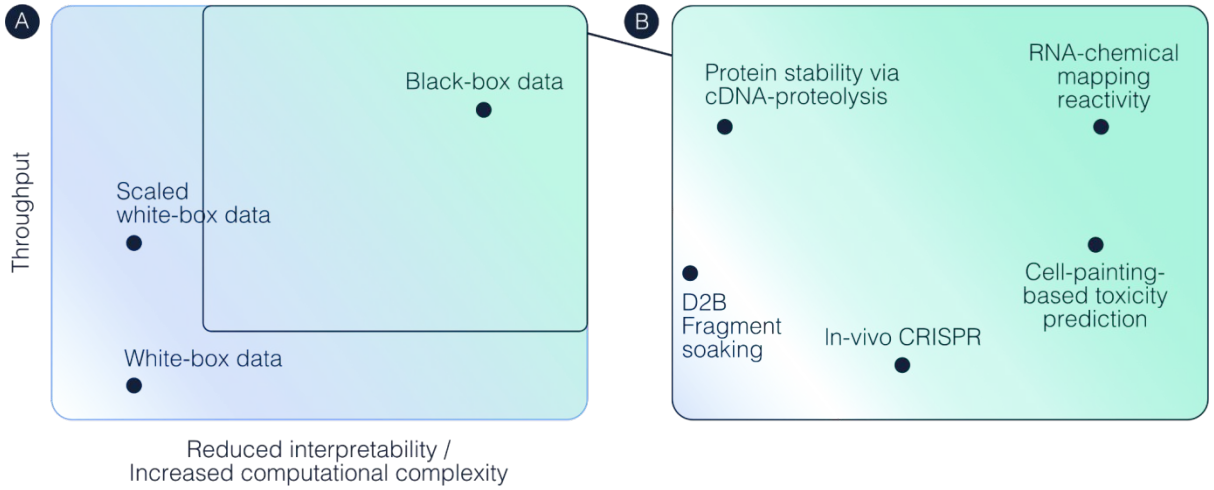

Figure 2: White-box versus black-box data. (A) Black-box data trades interpretability for throughput via increased computational processing complexity. (B) Zoomed in black-box data examples (non-exhaustive).

We note that "black-box" refers to the relationship between measurement and quantity of interest from the perspective of unaided human interpretation, not to the rigor of data generation. As with black-box algorithms, which require careful architectural and training choices despite producing opaque internal representations, black-box data demands deep domain expertise to hypothesize which modalities carry sufficient (possibly latent) signal and to design the corresponding reconstruction machinery.

Historical examples of black-box data

Our notion of "black-box data" in fact predates the recent dominance of black-box methods and was successfully applied in the past with white-box methods. Perhaps the most extreme example of black-box data is from the mathematical theory of compressed sensing,[45] providing

View Article Online
DOI: 10.1039/D5SC01189F

a theoretical guarantee of the ability to recover sparse signals from a few incoherent random projections. The data in compressed sensing intentionally looks like noise illegible to a human. The measurement process (represented by a random sensing matrix) is designed together with the reconstruction algorithm (based on solving an optimization problem) and requires a good understanding of the underlying problem (e.g. the assumption that data has a sparse representation in some basis, and the sensing matrix is incoherent with it). Compressed sensing techniques have been used in the medical imaging domain to design faster MRI acquisition protocols[46]. Multiple-access protocols, such as CDMA (Code-Division Multiple Access) and OFDM (Orthogonal Frequency-Division Multiplexing), are further examples, critical in mobile communications such as 5G. They allow transmitting information from multiple users over a single channel by combining the different users' signals via a convolution scheme. This reduces latency and increases throughput by avoiding time-dividing the channel (TDMA, Time-Division Multiple Access). The combined signal is completely uninterpretable without the deconvolution algorithm[47].

Within biology, the strategy of trading interpretability for scale is not new either. The Human Genome Project succeeded in part due to "whole-genome shotgun sequencing" (WGS) pioneered by Craig Venter and Celera Genomics[48]. Rather than assembling the genome sequentially, it is fragmented into millions of short pieces that are reassembled using alignment methods. The individual reads are too short to carry useful information, but in conjunction with alignment they can reconstruct (nearly) the whole genome. Shifting away from sequential assembly is what ultimately made the completion of the human genome possible.

Modern Examples and a Taxonomy of Black-box data generation tricks

While not new conceptually, we now increasingly see the emergence of new experimental "black-box data" sources developed in conjunction with ML. Excitingly, such black-box methods have already been employed across the whole lifecycle of basic biological research and drug discovery and development, as depicted in Figure 3.

View Article Online
DOI: 10.1039/D6SC01189F

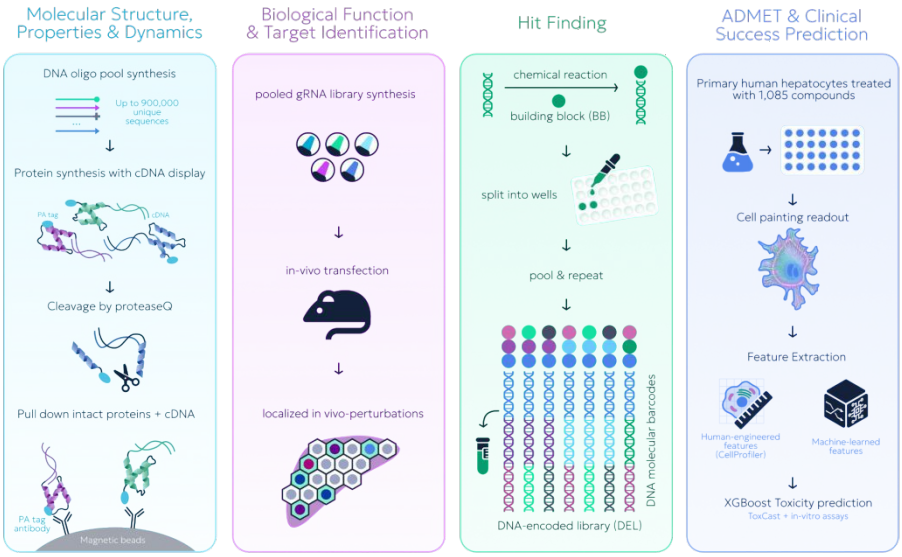

Fig. 3: Black-box data paradigms have been applied to the whole drug discovery spectrum. From left to right: high-throughput protein stability via cDNA proteolysis (Main text)[49], target identification through in-vivo CRISPR (Appendix)[50], DNA-encoded libraries for hit finding (Appendix)[51], liver toxicity prediction using Cell Painting readouts (Main text)[52].

Despite divergent objectives and modalities, black-box data approaches share recurring structural themes or "tricks". The examples listed below and in the Appendix can be organized into a seven-trick taxonomy (Figure 4). Many methods combine multiple tricks, often synergistically increasing throughput.

View Article Online
DOI: 10.1039/D6SC01189F

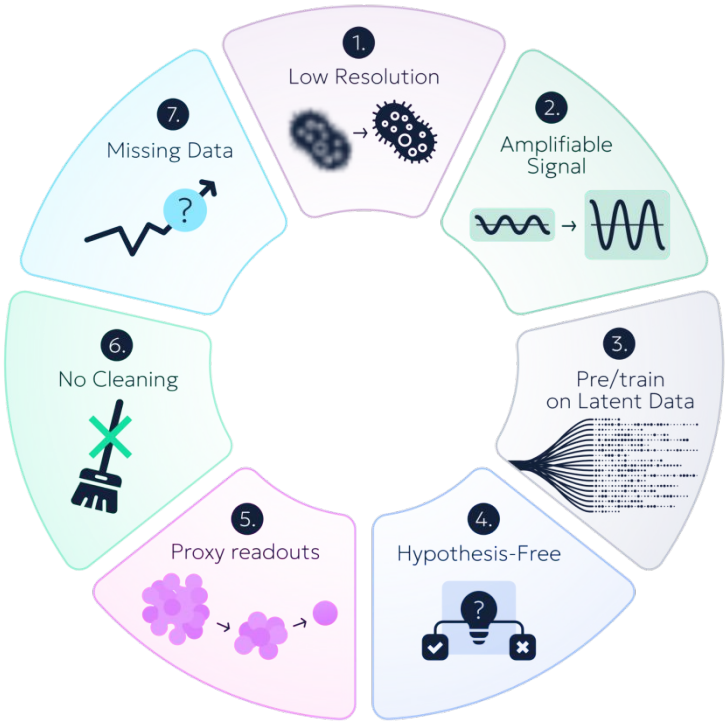

Fig. 4: Common tricks observed in black-box data approaches to shift trade-offs towards ML use.

Name	Primary trick	Primary Application Area	Short description	Location
Cross-linking Mass Spectrometry	Low Resolution	Molecular Structure, Properties & Dynamics	Whole-proteome structure determination via covalent distance markers	Main text
CMap	Low Resolution	Biological Function & Target Identification	Perturbomics via landmark gene arrays and imputation (L1000)	Appendix
cDNA proteolysis	Amplifiable Signal	Molecular Structure, Properties & Dynamics	Mega-scale protein folding stability measurements via sequencing	Main text
Human Domainome 1	Amplifiable Signal	Molecular Structure, Properties & Dynamics	Yeast protein complementation assay to link domain stability to growth read via sequencing readout	Appendix
DNA-encoded Library	Amplifiable Signal	Molecular Structure, Properties & Dynamics	Combinatorial library construction with DNA tags for massive-scale small molecule screens with sequencing readout	Appendix
Chemical Mapping	Latent Data	Molecular Structure, Properties & Dynamics	Mutational profiling of RNA after secondary structure-dependent reagent exposure for structure prediction via reactivity pre-training	Main text
ESMFold	Latent Data	Molecular Structure, Properties & Dynamics	Sequence-only pretraining extracts co-evolutionary motifs benefiting downstream structure prediction	Appendix
Sup35-aggregation screen	Hypothesis-Free	Molecular Structure, Properties & Dynamics	Massive random peptide screening via yeast Sup35 aggregation measured through ade1 reporter to train aggregation predictors	Main text
Cell Painting to predict toxicity	Proxy Readout	ADMET & Clinical Success Prediction	In-vivo (and in-vitro) toxicity prediction at high-throughput via ML on in-vitro Cell Painting outputs from hepatocytes post compound exposure	Main text

L1000 toxicity translation	Proxy Readout	ADMET & Clinical Success Prediction	In-vivo rat kidney toxicity prediction at high-throughput via L1000 human cell line profiling and ML	Appendix
Organoid toxicity translation	Proxy Readout	ADMET & Clinical Success Prediction	Human in-vivo GIT toxicity prediction at high-throughput via human ileal organoids and ML	Appendix
In-vivo CRISPR	Proxy Readout	ADMET & Clinical Success Prediction	Pooled, localized in-vivo CRISPR for high-throughput in-vivo target identification	Appendix
CRM Fragment Screening	No Cleaning	Hit Finding	Crystallographic fragment screening of Crude Reaction Mixtures denoised via SAR model for high-throughput hit-to-lead/lead optimization	Main text
Dyna-1	Missing Data	Molecular Structure, Properties & Dynamics	Reinterpret systematic absences in solution NMR tables as dynamics to train protein dynamics DL model Dyna-1	Main text
Diffuse scattering	Missing Data	Molecular Structure, Properties & Dynamics	Reinterpret diffuse scattering signal around bragg-peaks as dynamics to train protein-dynamics DL models	Appendix
AlphaMissense	Missing Data	Biological Function & Target Identification	Reinterpret missing variants as purifying selection to weakly label pathogenic variants and train DL variant effect predictor AlphaMissense	Appendix

Table 1: Black-box data examples mentioned in this article. Primary trick denotes the theme that is primarily highlighted in this article to provide full example coverage over tricks. Methods frequently combine many tricks, all of which are often critical.

Trick 1: Leverage low resolution readouts

Resolution and throughput are often inversely related. Low resolution manifests in two forms: sparse sampling (measuring a subset of features, as in microarrays capturing landmark transcripts rather than full transcriptomes[53]) or increased noise (losing high-frequency information, as in low-resolution Cryo-EM maps). Deep learning is particularly well-suited to both cases because learned priors can reconstruct missing or corrupted information from partial observations. Examples include cross-linking mass spectrometry (below), mutate-and-map chemical probing, and NMR-derived distance restraints (both covered later).

Example: Sparse distance restraints combined with structural foundation models unlock atomic-accuracy structures at a fraction of the cost

One of the central tenets of protein biology is that structure defines function[54]. However, as mentioned previously, current methods for structure elucidation are cost prohibitive without advances in data generation[14]. Structural proteomics provides one pathway to address this by switching to a readout that enables characterizing complex mixtures of many proteins in parallel (Mass Spectrometry). Cross-Linking Mass Spectrometry (XLMS) is a type of structural proteomics which allows to obtain sparse distance restraints at a whole proteome

View Article Online
DOI: 10.1039/D5SC01189F

scale[55]. By incubating cells or lysates with covalent probes called Cross-Linkers, residues that are at a distance no further than the maximum length of the extended crosslinker are connected covalently. Through these covalent markers, along with specialized spectral search methods[56], cross-linked peptides can be detected using MS, yielding a potentially whole-proteome, sparse residue contact map. Traditionally, this was predominantly employed to validate or guide biophysics approaches with orthogonal structural data (HDX-MS, NMR, templates) [57]. However, with deep learning-based structure prediction methods, models can use these structural restraints at inference time to dramatically increase success rates in structures prediction[58,59]. Recently, a biotech startup, Proxima Bio, has massively scaled this data source and unveiled their foundation model, Neo-1, making use of this data. They demonstrated that they can leverage their XLMS-based platform to predict novel protein interfaces that cannot be predicted by current models alone, and dramatically increase accuracy of small-molecule induced protein complex prediction[60].

Trick 2: Link quantity of interest to amplifiable signal

Increasing signal-to-noise ratio is a central problem in any assay. Including signal amplification steps can maintain high signal while increasing throughput. This has made two general readout categories "privileged": sequencing readouts, allowing for amplification via PCR; and positive selection assays where the quantity is tied to growth. Examples of the former include DNA-encoded libraries (appendix), mRNA-display[61], cDNA-linked proteolysis as in the MEGAScale experiment (covered below) and many more. Examples of the latter include yeast-two-hybrid screens[62] and the Sup35 and aPCA aggregation and stability assays covered later.

Example: MEGAscale proteolysis to train protein dynamics foundation models

Beyond structure, understanding how sequence and structural variation affect protein function is critical for protein engineering, therapeutic design, and disease mechanism. An example of addressing this data-need is cDNA display proteolysis, developed by the Rocklin lab [49]. In the MEGAscale assay, large DNA pooled libraries are produced with a puromycin linker, yielding protein–cDNA fusions captured on resin, making the proteins easily identifiable via the cDNA. The pool is exposed to titrations of proteases such as trypsin and chymotrypsin and after washing away cleaved proteins, intact proteins can be sequenced and counted. For each protein, the decline in counts across titrations yields a K_{50} value which is converted to folding free energies. Unlike direct structure determination, this gives indirect structural data via folding

View Article Online
DOI: 10.1039/D5SC01189F

free energies from a massively scalable assay, enabling over 776,000 measurements across hundreds of domains and dense mutational scans.

This data was used by a team at Microsoft Research to train BioEmu, a generative model that emulates protein equilibrium ensembles[63]. To obtain paired structural data, short molecular dynamics (MD) simulations for over 22,000 proteins from the dataset were generated to produce ensembles of folded and unfolded structures with experimental ΔG values used to reweight the simulation, creating a training set where the ratio of folded to unfolded structures reflects the measured stability. They further developed Property-Prediction Fine-Tuning (PPFT) to train the model directly to produce a distribution matching the experimental ΔG values without expensive paired simulation data..

Trick 3: Pre-train models on large abundant data with related but latent information

Black-box methods learn transferable representations rather than fixed algorithms. This is significant because it enables pretraining on abundant data from adjacent domains and fine-tuning for data-limited tasks without an *explicit* known relationship between pretraining and target tasks. RibonanzaNet2 and Dyna-1, covered in the following, as well as AlphaMissense and ESMFold (Appendix) all pretrain on data containing latent but not directly mappable task-relevant information.

Example: Chemical mapping data for RNA structure prediction and design

Structure prediction of nucleic acids currently significantly lags the performance of proteins. At CASP16, all top methods were expert predictors, with ML methods significantly behind[64]. Data scarcity is a core contributor: the PDB contained over 227,000 protein structures in 2025[27], but only 14,750 DNA/RNA-protein structures. And the 6,500 RNA structures are biased towards a few families (tRNAs, riboswitches, and ribozymes)[65]. A promising black-box remedy for this is *chemical mapping*[43] where reagents preferentially mutate flexible and/or unpaired nucleotides to yield reactivity values when sequenced under mutational profiling (MaP). This provides indirect RNA structural data that is many orders of magnitude more economic than traditional, direct structure determination (Cryo-EM/crystallography). Two popular reagents are dimethyl sulfate (DMS) for N^1-adenine and N^3-cytosine methylation for solvent-exposed nucleotides, and 2-*aminopyridine*-3-carboxylic acid imidazolide (2A3) for conformationally dynamic nucleotides. Multiple extensions exist: Mutate-and-map (M2)[66] systematic substitutions change reactivity of nucleotide pairs, creating correlated off-diagonal

View Article Online
DOI: 10.1039/D6SC01189F

signals providing a contact map. MOHCA-seq[67] and KARR-seq[68] correlate with distograms. Recent efforts[69] have scaled this data source significantly. As part of the 2025 Stanford 3D RNA Folding Kaggle challenge, 40M sequences chemical mapping profiles were produced and used to pre-train a range of different models, including an encoder predicting sequence-wise chemical reactivity. Combined with a diffusion structure prediction head trained on RNA structures, structure prediction performance dramatically improved compared to a model that was not pretrained on this task[70,71]. Representations learned from these tasks were also leveraged in a recent model, RNAPro, to surpass AF3 in RNA structure prediction[72]. The practical therapeutic relevance of this data source for RNA design was further demonstrated by Joshi et al. which combined a RNA language model (gRNAde) filtered by the first version of RibonanzaNet to yield high success rates surpassing previous models[73].

Trick 4: Hypothesis-free assays – unbiased, multiplexed and highly pooled readouts

Hypothesis-driven fields like drug discovery test limited, targeted hypotheses with low-throughput, low-noise, interpretable readouts. Unbiased "hypothesis-free" approaches sacrifice these properties but provide the diversity ML models need. Examples include the aforementioned structural proteomics analyzing whole-cell lysates or Sup35-based aggregation screens testing random 20-mer peptides rather than curated sets that we discuss below. Multiplexed and pooled readouts typically enable this scale.

Example: Massively scaled random peptide screening to train aggregation predictors

Thompson et al. used a hypothesis-free approach to predict protein aggregation[74]. Instead of a biased library, they constructed a pooled library of 100,000 random 20-mer peptides, each cloned in-frame upstream of the yeast prion nucleation domain of Sup35. In this system, peptides that nucleate or promote amyloid-like assembly of the Sup35 fusion trigger nonsense suppression of an ade1 reporter, causing growth on adenine-deficient media. Over several growth cycles fitness differences are amplified and barcodes are quantified by sequencing. They then developed a DL model (CANYA) with this data, outperforming all existing aggregation predictors trained on smaller, human-curated datasets.

View Article Online
DOI: 10.1039/D6SC01189F

Trick 5: Use proxy readouts that are less costly and higher throughput

Proxy readouts are ubiquitous in science: biomarkers, cell cultures, or model organisms. In our context, the use of proxies is a special case of Trick 3 (pretraining on latent information), where the readout derives from a proxy system sharing relevant biology but without direct translation to the target. Black-box methods excel here because they can learn the proxy-to-target mapping. Examples include Cell Painting to predict hepatotoxicity, ileal organoids to predict gastrointestinal toxicity (Appendix), in-vivo localized CRISPR to enable pooled in-vivo target discovery (Appendix).

Example: High-throughput in-vitro readouts to predict in-vivo toxicity across species

Cell Painting[75] has historically been extensively used as a biological proxy-readout in conjunction with black-box methods, perhaps most prominently by the startup Recursion Pharmaceuticals. In Cell Painting, cells are imaged after staining with six dyes highlighting distinct morphological features. Embeddings are typically extracted via hand-crafted pipelines (CellProfiler) or learned models. For example, in a seminal 2024 publication, Recursion leveraged black-box embeddings of Cell Painting data through their proprietary model to discover a hitherto unknown undesired genomic off-target pattern in CRISPR-Cas9 screens.[76]

Cell Painting was recently also applied in the context of toxicity prediction. Next to potency and affinity, developing drugs involves optimizing many other properties such as Absorption, Distribution, Metabolism and Toxicity (ADMET). Since these are often measured in-vivo, throughput is limited. Data sources which alleviate this bottleneck are needed, with an early example being pooling multiple compounds into a single animal via Cassette Dosing[77].

To address these throughput limitations, in a recent preprint by a team from MIT, the Broad Institute and the startup Axiom Bio, authors exposed primary human hepatocytes to 1085 compounds with previously known in-vivo hepatotoxicity readouts from ToxCast[78]. They assessed whether Cell Painting is simultaneously predictive of hundreds of *in vitro* toxicity readouts, including assays for metabolic activity and membrane damage measured by the authors and 412 cytotoxicity and mode-of-action endpoints measured by the ToxCast screening effort[52]. They extracted various features from the images, including hand-crafted CellProfiler

View Article Online
DOI: 10.1039/D3SC01189F

features and "black-box" embeddings (CellPainting CNN and DINOv2), and used the features to predict assay activity using XGBoost.

In a similar spirit, Gardiner et al. developed a black-box inspired cross-species toxicity translation: they train a ML model on the molecular structure and their associated in-vitro human cell line landmark gene array profiles to predict in-vivo rat kidney toxicity measured via blood urea nitrogen (combining both the "proxy" and "low resolution theme" via the L1000 assay also leveraged in CMap, see appendix)[79]. Researchers at Celgene employed a similar idea – in this case leveraging Machine Learning and human ileal organoid models to recapitulate human clinical in-vivo Gastrointestinal toxicity (GIT) with 90% accuracy[80].

Trick 6: Omit cleaning or filtering of readouts

Purification and isolation are often rate-limiting: protein production, small molecule synthesis, and sample preparation are all bottlenecked by purification. Skipping purification yields noisier readouts, connecting this trick to Trick 1 (low-resolution readouts) through shared reliance on DL noise tolerance and priors. The distinction: Trick 1 changes the readout modality (e.g., Cryo-EM to XLMS), while Trick 6 removes a protocol step within the same modality. Direct-to-Biology (D2B, below) and DNA-encoded libraries (appendix) exemplify this approach. Trick 6 also intersects with Trick 4 (hypothesis-free assays), which often measure unpurified, unbiased samples by design.

Example: Crude Reaction Mixtures unlock high-throughput lead optimization

Nowadays, many thousands of DNA sequences or proteins can be synthesized and tested on the order of weeks[81]. Yet for small molecules, custom synthesis is significantly slower, often requiring months, years, or even decades in extreme cases to establish synthetic pathways for just a single new molecule[82], dramatically slowing hit-to-lead campaigns in drug discovery. The large synthesis quantities required (due to material losses during purification) and the slow speed of purification are dominant bottlenecks in synthesis[83]. In D2B this is addressed by testing Crude Reaction Mixtures (CRM) directly without purification, allowing well-based parallel nanoscale synthesis with dramatic throughput increase at the cost of increased false negatives/positives due to impurities[42]. Again, ML is used to compensate for this. An example is McCorkindale et al., who predicted potency of each molecule, allowing to rescue mixtures that are incorrectly labelled as weak binders in the assay but labeled positive by the model[84].

A promising combination has been with fragment screening. As opposed to traditional high-throughput screening, where often billions of compounds are tested for binding, in fragment screening a smaller set of diverse (e.g. few thousand) of low-molecular weight fragments (120-250 Dalton) are screened[85]. These can be merged, linked or elaborated into complete and high affinity binders, allowing to explore a wide range of chemical space with a small "basis set" of tested fragments. Since initial fragments are often of lower affinity, highly sensitive readouts, such as crystallographic soaking are needed[86]. Grosjean et al. employed this to the bromodomain of PHIP(2)[87]. CRMs of 957 analogs of a fragment hit yielded 22 binders. A simple Structure-Activity-Relationship algorithm built on the 22 binders could denoise the results by rescuing false negatives, identifying 26 additional, mislabeled binders. They further used it in virtual screening, identifying 9 binders. Recently, the OpenBind consortium was formed to systematically scale this data source for Machine Learning purposes[88].

D2B is also extensively used by many larger companies and startups[89]. Kimia Therapeutics employs it in conjunction with Chemotype Evolution[90] where "generation of molecules" are produced starting from a target-binding bait fragment with a reactive handle that is combinatorially coupled to fragments with further reactive handles, enabling iterative "activity-based" evolution. They demonstrated this for a selective covalent KRAS-G12C inhibitor series[91]. Octant is another example coupling D2B to multiplexed reporter assays[92–94].

Trick 7: Leverage signal from missing, or weakly-labeled data

Missing or noisy data is often systematic rather than random. While insufficient for human interpretation, the structure implicit in the missing data can provide valuable weak supervision for ML. Examples: AlphaMissense[95] treats unobserved variants as pathogenic (purifying selection), diffuse X-ray scattering[96] extracts dynamics from traditionally discarded scattering noise (all in appendix), and missing NMR peaks can indicate conformational exchange rather than experimental failure (below). Trick 7 differs from Trick 1 since here the perceived "noise" is the primary source of signal and not noise after all. This is related to Trick 3 as the patterns in the noise may be primarily latent.

View Article Online
DOI: 10.1039/D6SC01189F

Example: Missing NMR peaks is supervision to learn protein dynamics

Experimental data on protein dynamics is hard to obtain at scale. Models are often exclusively trained on a limited amount of simulated data obtained at high computational cost. Wayment-Steele, El Nesr et al. developed a method to extract microsecond-to-millisecond (μs–ms) protein dynamics from missing NMR peak assignments in the Biological Magnetic Resonance Data Bank (BMRB) and trained Dyna-1, a deep learning model predicting these dynamics, to demonstrate the data's potential[97]. The main realization was that "missing" NMR peaks (systematic absences) from chemical shift datasets can indicate μs–ms dynamics. In the regime where the interconversion rate is comparable to the angular frequency separation of exchanging states, transverse relaxation acquires an exchange contribution, producing line broadening and, in the limit, loss of peaks in standard heteronuclear experiments. Solution NMR assignment tables in the BMRB often contain these absences which were traditionally interpreted as experimental artifacts. To extract signals from this noisy weak supervision, they built on a frozen pre-trained language model, training a prediction head on data extracted from ~10,000 proteins from the BMRB (two orders of magnitude larger than existing datasets) to produce Dyna-1. It accurately predicts μs–ms dynamics when evaluated against high quality curated datasets (RelaxDB, RelaxDB-CPMG). This demonstrates another critical component of black-box data approaches - the requirement for high-quality "white-box" (often smaller sized) benchmark data.

Future Outlook: The Scientific Process in the Black-Box Age

We believe that black-box data provides a viable path to overcome data limitations that bottleneck the advancement of biological applications of AI. As the field is increasingly driven through model-led discoveries, much of the way data is generated will change to adapt to ML as the primary consumer. This will drive a fundamental shift in the types of data created and there will be critical considerations to make this transition work.

The generation of black-box data should not be viewed solely as a static, feed-forward process. With lab-in-the-loop and active learning approaches, instead of passively mining existing datasets, models can evaluate their own uncertainty to dictate the next batch of massively multiplexed experiments. This closed-loop experimental design will ensure that high-throughput, noisy assays are deployed where the model needs them most, maximizing data

View Article Online
DOI: 10.1039/D6SC01189F

efficiency and accelerating the discovery process in a way that traditional, hypothesis-driven screening cannot.

Since black-box datasets sacrifice human interpretability for scale, they might be fundamentally unintelligible without algorithmic deconvolution. Standardizing raw readouts, precise experimental metadata, and noise profiles across different laboratories is thus an important hurdle. We believe that more generally, the field requires a shift toward "AI-first" data repositories moving away from the traditional scientific publishing route—where only cleaned, processed conclusions are shared—toward a pipeline where raw, unpurified readouts are standardized and fed directly into foundation models.

Nascent consortium efforts provide an early blueprint for this infrastructure. For example, the OpenBind consortium was formed to systematically scale and standardize crude reaction mixture fragment screening data for ML purposes. Similarly, the Diffuse Project aims to systematically collect and standardize previously discarded X-ray scattering noise into a shared repository for learning protein dynamics.

The importance of human intuition remains undiminished in the black-box data era. Even though the required signal-to-noise of black-box data is reduced, and the signal can be indirect and latent, there still needs to be sufficient signal for the models to extract. Hypothesizing and testing which black-box data methods contain sufficient signal will need to be driven by domain expertise, as will be evaluating the resulting models.

White-box and noise-free data will be needed to verify and benchmark black-box methods. While the learning process of these methods is noise tolerant, their evaluation is less so[98]. Being black-box makes this even more critical as we cannot always understand, but only assess these methods for their accuracy, making evaluation a critical fail-safe. In practice, we expect white-box and black-box data to exist in a symbiotic relationship: small curated white-box datasets can be used to construct initial priors, calibrate reconstruction algorithms, validate model predictions, and it will often be desirable to “post-train” models trained on more noisy “black-box” data with noise-minimized high-quality white-box data.

This lack of interpretability is a common criticism of black-box data and algorithms. The overarching concern is that black-box approaches usher in a “post-theory” science where we

View Article Online
DOI: 10.1039/D6SC01189F

can simulate, predict, yet not understand future scientific phenomena[99]. While black-box methods are not *directly* interpretable, there exists nonetheless a plethora of tools from the field of Mechanistic Interpretability (MI) to extract knowledge from these models such as sparse autoencoders[100]. A biological example is the categorical Jacobian, applied to language-model based folding methods, to understand the motif-driven recall process that underlies the success of the current generation of folding methods[101]. In another case, teams from Prima Mente and Goodfire used MI on a large foundation model, *Pleiades*, pretrained on large amounts of patient genomic and ctDNA, to extract the finding that ctDNA fragment length could serve as diagnostic marker for Alzheimer's from the weights of the black-box model[102]. These point to a future where there might be an inversion of the scientific method from *understand, encode, and then simulate* – a Software 1.0 paradigm – to *encode, simulate, understand* (via MI) – a Software 2.0 approach in which logic is not an *input* into the scientific discovery process but its *output*. Hence mechanistic interpretability will become critical to extract knowledge about scientific processes from the weights of neural networks that simulate them – standing on the shoulders of giants increasingly made of silicon.

Acknowledgements

We thank Abhishaike Mahajan, Imran Haque, Hannah Wayment-Steele, Jess Ewald, Chaitanya Joshi, Ian Quigley and Martin Borch Jensen for thoughtful feedback.

Luca Naef is the Chief Technology Officer and Co-Founder of Proxima, a Biotech and Frontier AI research lab combining high-throughput data generation and foundation models to tackle the rational discovery of proximity modulators such as molecular glues and PROTACs. He previously helped develop Machine Learning models across top biopharma companies as part of McKinsey's QuantumBlack and has worked in multiple Biotech and Software companies. He holds a BSc and MSc from ETH Zurich, has conducted research across Stanford, ETH, Tokyo Institute of Technology, University of New South Wales. He has extensively published on novel Deep Learning models, datasets and evaluations for Biology and Chemistry.

View Article Online
DOI: 10.1039/D6SC01189F

Michael Bronstein is the DeepMind Professor of AI at the University of Oxford and Founding Scientific Director of the Aithyra institute in Vienna. Previously, he was Head of Graph Learning Research at Twitter and a professor at Imperial College London, and held visiting appointments at Stanford, MIT, and Harvard. He received his PhD from the Technion in 2007. Michael's main research focus is theoretical and computational methods in Geometric Deep Learning and their applications to biochemistry and structural biology. His work has been recognised with the EPSRC Turing AI World-Leading Research Fellowship, the Royal Society Wolfson Research Merit Award, the Royal Academy of Engineering Silver Medal, and multiple ERC grants. He is a Member of Academia Europaea and a Fellow of the IEEE, IAPR, and ELLIS. Michael founded multiple startups including Invision (acquired by Intel in 2012) and Fabula AI (acquired by Twitter in 2019). He currently serves as Chief Scientist-in-Residence at Proxima Bio and is on the scientific advisory board of Recursion Pharmaceuticals and Relation Therapeutics.

References

(1) *Mastering the game of Go with deep neural networks and tree search | Nature*. https://www.nature.com/articles/nature16961 (accessed 2026-01-03).
(2) *Grandmaster level in StarCraft II using multi-agent reinforcement learning | Nature*. https://www.nature.com/articles/s41586-019-1724-z (accessed 2026-01-03).
(3) Castelvecchi, D. DeepMind and OpenAI Models Solve Maths Problems at Level of Top Students. *Nature* **2025**, *644* (8075), 20–20. https://doi.org/10.1038/d41586-025-02343-x.
(4) Pacesa, M.; Nickel, L.; Schellhaas, C.; Schmidt, J.; Pyatova, E.; Kissling, L.; Barendse, P.; Choudhury, J.; Kapoor, S.; Alcaraz-Serna, A.; Cho, Y.; Ghamary, K. H.; Vinué, L.; Yachnin, B. J.; Wollacott, A. M.; Buckley, S.; Westphal, A. H.; Lindhoud, S.; Georgeon, S.; Goverde, C. A.; Hatzopoulos, G. N.; Gönczy, P.; Muller, Y. D.; Schwank, G.; Swarts, D. C.; Vecchio, A. J.; Schneider, B. L.; Ovchinnikov, S.; Correia, B. E. One-Shot Design of Functional Protein Binders with BindCraft. *Nature* **2025**, 1–10. https://doi.org/10.1038/s41586-025-09429-6.
(5) Naef, L.; Bronstein, M. *The Road to Biology 2.0 Will Pass Through Black-Box Data*. TDS Archive. https://medium.com/data-science/the-road-to-biology-2-0-will-pass-through-black-box-data-bbd00fabf959 (accessed 2025-09-24).
(6) Gao, M.; Skolnick, J. Structural Space of Protein-Protein Interfaces Is Degenerate, Close to Complete, and Highly Connected. *Proc. Natl. Acad. Sci. U. S. A.* **2010**, *107* (52), 22517–22522. https://doi.org/10.1073/pnas.1012820107.
(7) Chen, S.-J.; Hassan, M.; Jernigan, R. L.; Jia, K.; Kihara, D.; Kloczkowski, A.; Kotelnikov, S.; Kozakov, D.; Liang, J.; Liwo, A.; Matysiak, S.; Meller, J.; Micheletti, C.; Mitchell, J. C.; Mondal, S.; Nussinov, R.; Okazaki, K.; Padhorny, D.; Skolnick, J.; Sosnick, T. R.; Stan, G.; Vakser, I.; Zou, X.; Rose, G. D. Protein Folds vs. Protein Folding: Differing Questions, Different Challenges. *Proc. Natl. Acad. Sci.* **2023**, *120* (1), e2214423119. https://doi.org/10.1073/pnas.2214423119.

View Article Online
DOI: 10.1039/D6SC01189F

(8) Mackenzie, C. O.; Zhou, J.; Grigoryan, G. Tertiary Alphabet for the Observable Protein Structural Universe. *Proc. Natl. Acad. Sci.* **2016**, *113* (47), E7438–E7447. https://doi.org/10.1073/pnas.1607178113.
(9) Yeo, J.; Han, Y.; Bordin, N.; Lau, A. M.; Kandathil, S. M.; Kim, H.; Karin, E. L.; Mirdita, M.; Jones, D. T.; Orengo, C.; Steinegger, M. Metagenomic-Scale Analysis of the Predicted Protein Structure Universe. bioRxiv April 26, 2025, p 2025.04.23.650224. https://doi.org/10.1101/2025.04.23.650224.
(10) Durairaj, J.; Adeshina, Y.; Cao, Z.; Zhang, X.; Oleinikovas, V.; Duignan, T.; McClure, Z.; Robin, X.; Kovtun, D.; Rossi, E.; Zhou, G.; Veccham, S.; Isert, C.; Peng, Y.; Sundareson, P.; Akdel, M.; Corso, G.; Stärk, H.; Carpenter, Z.; Bronstein, M.; Kucukbenli, E.; Schwede, T.; Naef, L. PLINDER: The Protein-Ligand Interactions Dataset and Evaluation Resource. bioRxiv July 17, 2024, p 2024.07.17.603955. https://doi.org/10.1101/2024.07.17.603955.
(11) Kovtun, D.; Akdel, M.; Goncearenco, A.; Zhou, G.; Holt, G.; Baugher, D.; Lin, D.; Adeshina, Y.; Castiglione, T.; Wang, X.; Marquet, C.; McPartlon, M.; Geffner, T.; Corso, G.; Stärk, H.; Carpenter, Z.; Kucukbenli, E.; Bronstein, M.; Naef, L. PINDER: The Protein Interaction Dataset and Evaluation Resource. bioRxiv July 21, 2024, p 2024.07.17.603980. https://doi.org/10.1101/2024.07.17.603980.
(12) Škrinjar, P.; Eberhardt, J.; Durairaj, J.; Schwede, T. Have Protein-Ligand Co-Folding Methods Moved beyond Memorisation? bioRxiv February 7, 2025, p 2025.02.03.636309. https://doi.org/10.1101/2025.02.03.636309.
(13) Erika DeBenedictis. *How do we make the next PDB? Introducing Align's Data Platform*. NeurIPS Machine Learning in Structural Biology Workshop. https://slideslive.com/embed/presentation/39031188?js_embed_version=3&embed_init_token=eyJhbGciOiJIUzI1NiJ9.eyJpYXQiOjE3NjE4MDAxNDksImV4cCI6MTc2MTkyOTc0OSwidSI6eyJ1dWlkIjoiYmExYTViNzYtZjM1Zi00NThlLTljYTAtOGY3OWI3ZDcyZDk2IiwiaSI6bnVsbCwiZSI6bnVsbCwibSI6ZmFsc2V9LCJkIjoibmV1cmlwcy5jYyJ9.emkI0Ur-Pg-TPByWfqZPE406jDNCy55AFvlhHlf3q_w (accessed 2025-10-29).
(14) *Protein Data Bank Marks 50 Years Helping to Unlock Mysteries of Human Disease | Rutgers University*. https://www.rutgers.edu/news/protein-data-bank-marks-50-years-lifesaving-scientific-collaboration-and-coordination (accessed 2025-09-24).
(15) Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. In *Advances in Neural Information Processing Systems*; Curran Associates, Inc., 2012; Vol. 25.
(16) Karpathy, A. *Software 2.0*. Medium. https://karpathy.medium.com/software-2-0-a64152b37c35 (accessed 2025-09-24).
(17) Sutton, R. *The Bitter Lesson*. The Bitter Lesson. https://www.cs.utexas.edu/~eunsol/courses/data/bitter_lesson.pdf (accessed 2025-09-24).
(18) Senior, A. W.; Evans, R.; Jumper, J.; Kirkpatrick, J.; Sifre, L.; Green, T.; Qin, C.; Žídek, A.; Nelson, A. W. R.; Bridgland, A.; Penedones, H.; Petersen, S.; Simonyan, K.; Crossan, S.; Kohli, P.; Jones, D. T.; Silver, D.; Kavukcuoglu, K.; Hassabis, D. Protein Structure Prediction Using Multiple Deep Neural Networks in the 13th Critical Assessment of Protein Structure Prediction (CASP13). *Proteins Struct. Funct. Bioinforma.* **2019**, *87* (12), 1141–1148. https://doi.org/10.1002/prot.25834.
(19) Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; Bridgland, A.; Meyer, C.; Kohl, S. A. A.; Ballard, A. J.; Cowie, A.; Romera-Paredes, B.; Nikolov, S.; Jain, R.; Adler, J.; Back, T.; Petersen, S.; Reiman, D.; Clancy, E.; Zielinski, M.; Steinegger, M.;

Pacholska, M.; Berghammer, T.; Bodenstein, S.; Silver, D.; Vinyals, O.; Senior, A. W.; Kavukcuoglu, K.; Kohli, P.; Hassabis, D. Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* **2021**, *596* (7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2.

(20) Abramson, J.; Adler, J.; Dunger, J.; Evans, R.; Green, T.; Pritzel, A.; Ronneberger, O.; Willmore, L.; Ballard, A. J.; Bambrick, J.; Bodenstein, S. W.; Evans, D. A.; Hung, C.-C.; O'Neill, M.; Reiman, D.; Tunyasuvunakool, K.; Wu, Z.; Žemgulytė, A.; Arvaniti, E.; Beattie, C.; Bertolli, O.; Bridgland, A.; Cherepanov, A.; Congreve, M.; Cowen-Rivers, A. I.; Cowie, A.; Figurnov, M.; Fuchs, F. B.; Gladman, H.; Jain, R.; Khan, Y. A.; Low, C. M. R.; Perlin, K.; Potapenko, A.; Savy, P.; Singh, S.; Stecula, A.; Thillaisundaram, A.; Tong, C.; Yakneen, S.; Zhong, E. D.; Zielinski, M.; Žídek, A.; Bapst, V.; Kohli, P.; Jaderberg, M.; Hassabis, D.; Jumper, J. M. Accurate Structure Prediction of Biomolecular Interactions with AlphaFold 3. *Nature* **2024**, 1–3. https://doi.org/10.1038/s41586-024-07487-w.

(21) Lin, Z.; Akin, H.; Rao, R.; Hie, B.; Zhu, Z.; Lu, W.; Smetanin, N.; Verkuil, R.; Kabeli, O.; Shmueli, Y.; Costa, A. dos S.; Fazel-Zarandi, M.; Sercu, T.; Candido, S.; Rives, A. Evolutionary-Scale Prediction of Atomic Level Protein Structure with a Language Model. bioRxiv October 31, 2022, p 2022.07.20.500902. https://doi.org/10.1101/2022.07.20.500902.

(22) Wang, Y.; Lu, J.; Jaitly, N.; Susskind, J.; Bautista, M. A. SimpleFold: Folding Proteins Is Simpler than You Think. arXiv September 23, 2025. https://doi.org/10.48550/arXiv.2509.18480.

(23) Hayes, T.; Rao, R.; Akin, H.; Sofroniew, N. J.; Oktay, D.; Lin, Z.; Verkuil, R.; Tran, V. Q.; Deaton, J.; Wiggert, M.; Badkundri, R.; Shafkat, I.; Gong, J.; Derry, A.; Molina, R. S.; Thomas, N.; Khan, Y.; Mishra, C.; Kim, C.; Bartie, L. J.; Nemeth, M.; Hsu, P. D.; Sercu, T.; Candido, S.; Rives, A. Simulating 500 Million Years of Evolution with a Language Model. bioRxiv July 2, 2024, p 2024.07.01.600583. https://doi.org/10.1101/2024.07.01.600583.

(24) Antonsson, S. E.; Melsted, P. Batch Correction Methods Used in Single Cell RNA-Sequencing Analyses Are Often Poorly Calibrated. bioRxiv March 21, 2024, p 2024.03.19.585562. https://doi.org/10.1101/2024.03.19.585562.

(25) Spinner, A.; DeBenedictis, E.; Hudson, C. M. Scaling and Data Saturation in Protein Language Models. arXiv July 29, 2025. https://doi.org/10.48550/arXiv.2507.22210.

(26) Zdrazil, B. Fifteen Years of ChEMBL and Its Role in Cheminformatics and Drug Discovery. *J. Cheminformatics* **2025**, *17* (1), 32. https://doi.org/10.1186/s13321-025-00963-z.

(27) Burley, S. K.; Bhatt, R.; Bhikadiya, C.; Bi, C.; Biester, A.; Biswas, P.; Bittrich, S.; Blaumann, S.; Brown, R.; Chao, H.; Chithari, V. R.; Craig, P. A.; Crichlow, G. V.; Duarte, J. M.; Dutta, S.; Feng, Z.; Flatt, J. W.; Ghosh, S.; Goodsell, D. S.; Green, R. K.; Guranovic, V.; Henry, J.; Hudson, B. P.; Joy, M.; Kaelber, J. T.; Khokhriakov, I.; Lai, J.-S.; Lawson, C. L.; Liang, Y.; Myers-Turnbull, D.; Peisach, E.; Persikova, I.; Piehl, D. W.; Pingale, A.; Rose, Y.; Sagendorf, J.; Sali, A.; Segura, J.; Sekharan, M.; Shao, C.; Smith, J.; Trumbull, M.; Vallat, B.; Voigt, M.; Webb, B.; Whetstone, S.; Wu-Wu, A.; Xing, T.; Young, J. Y.; Zalevsky, A.; Zardecki, C. Updated Resources for Exploring Experimentally-Determined PDB Structures and Computed Structure Models at the RCSB Protein Data Bank. *Nucleic Acids Res.* **2025**, *53* (D1), D564–D574. https://doi.org/10.1093/nar/gkae1091.

(28) Martens, L.; Hermjakob, H.; Jones, P.; Adamski, M.; Taylor, C.; States, D.; Gevaert, K.; Vandekerckhove, J.; Apweiler, R. PRIDE: The Proteomics Identifications

View Article Online
DOI: 10.1039/D6SC01189F

Database. *Proteomics* **2005**, *5* (13), 3537–3545. https://doi.org/10.1002/pmic.200401303.
(29) Clough, E.; Barrett, T. The Gene Expression Omnibus Database. *Methods Mol. Biol. Clifton NJ* **2016**, *1418*, 93–110. https://doi.org/10.1007/978-1-4939-3578-9_5.
(30) Koscielny, G.; An, P.; Carvalho-Silva, D.; Cham, J. A.; Fumis, L.; Gasparyan, R.; Hasan, S.; Karamanis, N.; Maguire, M.; Papa, E.; Pierleoni, A.; Pignatelli, M.; Platt, T.; Rowland, F.; Wankar, P.; Bento, A. P.; Burdett, T.; Fabregat, A.; Forbes, S.; Gaulton, A.; Gonzalez, C. Y.; Hermjakob, H.; Hersey, A.; Jupe, S.; Kafkas, Ş.; Keays, M.; Leroy, C.; Lopez, F.-J.; Magarinos, M. P.; Malone, J.; McEntyre, J.; Munoz-Pomer Fuentes, A.; O'Donovan, C.; Papatheodorou, I.; Parkinson, H.; Palka, B.; Paschall, J.; Petryszak, R.; Pratanwanich, N.; Sarntivijal, S.; Saunders, G.; Sidiropoulos, K.; Smith, T.; Sondka, Z.; Stegle, O.; Tang, Y. A.; Turner, E.; Vaughan, B.; Vrousgou, O.; Watkins, X.; Martin, M.-J.; Sanseau, P.; Vamathevan, J.; Birney, E.; Barrett, J.; Dunham, I. Open Targets: A Platform for Therapeutic Target Identification and Validation. *Nucleic Acids Res.* **2017**, *45* (Database issue), D985–D994. https://doi.org/10.1093/nar/gkw1055.
(31) Ferreira de Freitas, R.; Schapira, M. A Systematic Analysis of Atomic Protein–Ligand Interactions in the PDB. *MedChemComm* **2017**, *8* (10), 1970–1981. https://doi.org/10.1039/C7MD00381A.
(32) Ahdritz, G.; Bouatta, N.; Kadyan, S.; Xia, Q.; Gerecke, W.; O'Donnell, T. J.; Berenberg, D.; Fisk, I.; Zanichelli, N.; Zhang, B.; Nowaczynski, A.; Wang, B.; Stepniewska-Dziubinska, M. M.; Zhang, S.; Ojewole, A.; Guney, M. E.; Biderman, S.; Watkins, A. M.; Ra, S.; Lorenzo, P. R.; Nivon, L.; Weitzner, B.; Ban, Y.-E. A.; Sorger, P. K.; Mostaque, E.; Zhang, Z.; Bonneau, R.; AlQuraishi, M. OpenFold: Retraining AlphaFold2 Yields New Insights into Its Learning Mechanisms and Capacity for Generalization. bioRxiv November 22, 2022, p 2022.11.20.517210. https://doi.org/10.1101/2022.11.20.517210.
(33) Rolnick, D.; Veit, A.; Belongie, S.; Shavit, N. Deep Learning Is Robust to Massive Label Noise. arXiv February 26, 2018. https://doi.org/10.48550/arXiv.1705.10694.
(34) *When Quantity Becomes Quality with Aviv Regev*. Andreessen Horowitz. https://a16z.com/podcast/when-quantity-becomes-quality-with-aviv-regev/ (accessed 2025-10-27).
(35) Dodge, J.; Sap, M.; Marasović, A.; Agnew, W.; Ilharco, G.; Groeneveld, D.; Mitchell, M.; Gardner, M. Documenting Large Webtext Corpora: A Case Study on the Colossal Clean Crawled Corpus. arXiv September 30, 2021. https://doi.org/10.48550/arXiv.2104.08758.
(36) Lambert, N. Reinforcement Learning from Human Feedback. arXiv June 11, 2025. https://doi.org/10.48550/arXiv.2504.12501.
(37) Joshi, C. K.; Fu, X.; Liao, Y.-L.; Gharakhanyan, V.; Miller, B. K.; Sriram, A.; Ulissi, Z. W. All-Atom Diffusion Transformers: Unified Generative Modelling of Molecules and Materials. arXiv May 22, 2025. https://doi.org/10.48550/arXiv.2503.03965.
(38) Ho, J.; Jain, A.; Abbeel, P. Denoising Diffusion Probabilistic Models. arXiv December 16, 2020. https://doi.org/10.48550/arXiv.2006.11239.
(39) Daras, G.; Shah, K.; Dagan, Y.; Gollakota, A.; Dimakis, A. G.; Klivans, A. Ambient Diffusion: Learning Clean Distributions from Corrupted Data. arXiv May 30, 2023. https://doi.org/10.48550/arXiv.2305.19256.

(40) Rosenthal, R. The File Drawer Problem and Tolerance for Null Results. *Psychol. Bull.* **1979**, *86* (3), 638–641. https://doi.org/10.1037/0033-2909.86.3.638.
(41) Maloney, M. P.; Coley, C. W.; Genheden, S.; Carson, N.; Helquist, P.; Norrby, P.-O.; Wiest, O. Negative Data in Data Sets for Machine Learning Training. *J. Org. Chem.* **2023**, *88* (9), 5239–5241. https://doi.org/10.1021/acs.joc.3c00844.
(42) Thomas, R. P.; Heap, R. E.; Zappacosta, F.; Grant, E. K.; Pogány, P.; Besley, S.; Fallon, D. J.; Hann, M. M.; House, D.; Tomkinson, N. C. O.; Bush, J. T. A Direct-to-Biology High-Throughput Chemistry Approach to Reactive Fragment Screening. *Chem. Sci.* **2021**, *12* (36), 12098–12106. https://doi.org/10.1039/D1SC03551G.
(43) Cheng, C. Y.; Kladwang, W.; Yesselman, J. D.; Das, R. RNA Structure Inference through Chemical Mapping after Accidental or Intentional Mutations. *Proc. Natl. Acad. Sci.* **2017**, *114* (37), 9876–9881. https://doi.org/10.1073/pnas.1619897114.
(44) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł. ukasz; Polosukhin, I. Attention Is All You Need. In *Advances in Neural Information Processing Systems*; Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R., Eds.; Curran Associates, Inc., 2017; Vol. 30.
(45) Candes, E.; Romberg, J.; Tao, T. Stable Signal Recovery from Incomplete and Inaccurate Measurements. arXiv December 7, 2005. https://doi.org/10.48550/arXiv.math/0503066.
(46) Lustig, M.; Donoho, D.; Pauly, J. M. Sparse MRI: The Application of Compressed Sensing for Rapid MR Imaging. *Magn. Reson. Med.* **2007**, *58* (6), 1182–1195. https://doi.org/10.1002/mrm.21391.
(47) Stüber, G. L. *Principles of Mobile Communication*; Springer International Publishing: Cham, 2017. https://doi.org/10.1007/978-3-319-55615-4.
(48) Venter, J. C.; Adams, M. D.; Sutton, G. G.; Kerlavage, A. R.; Smith, H. O.; Hunkapiller, M. Shotgun Sequencing of the Human Genome. *Science* **1998**, *280* (5369), 1540–1542. https://doi.org/10.1126/science.280.5369.1540.
(49) Tsuboyama, K.; Dauparas, J.; Chen, J.; Laine, E.; Mohseni Behbahani, Y.; Weinstein, J. J.; Mangan, N. M.; Ovchinnikov, S.; Rocklin, G. J. Mega-Scale Experimental Analysis of Protein Folding Stability in Biology and Design. *Nature* **2023**, *620* (7973), 434–444. https://doi.org/10.1038/s41586-023-06328-6.
(50) Santinha, A. J.; Strano, A.; Platt, R. J. Methods and Applications of in Vivo CRISPR Screening. *Nat. Rev. Genet.* **2025**, *26* (10), 702–718. https://doi.org/10.1038/s41576-025-00873-8.
(51) McCloskey, K.; Sigel, E. A.; Kearnes, S.; Xue, L.; Tian, X.; Moccia, D.; Gikunju, D.; Bazzaz, S.; Chan, B.; Clark, M. A.; Cuozzo, J. W.; Guié, M.-A.; Guilinger, J. P.; Huguet, C.; Hupp, C. D.; Keefe, A. D.; Mulhern, C. J.; Zhang, Y.; Riley, P. Machine Learning on DNA-Encoded Libraries: A New Paradigm for Hit Finding. *J. Med. Chem.* **2020**, *63* (16), 8857–8866. https://doi.org/10.1021/acs.jmedchem.0c00452.
(52) Ewald, J. D.; Titterton, K. L.; Bäuerle, A.; Beatson, A.; Boiko, D. A.; Cabrera, Á. A.; Cheah, J.; Cimini, B. A.; Gorissen, B.; Jones, T.; Karczewski, K. J.; Rouquie, D.; Seal, S.; Weisbart, E.; White, B.; Carpenter, A. E.; Singh, S. Cell Painting for Cytotoxicity and Mode-of-Action Analysis in Primary Human Hepatocytes. bioRxiv January 24, 2025, p 2025.01.22.634152. https://doi.org/10.1101/2025.01.22.634152.
(53) Subramanian, A.; Narayan, R.; Corsello, S. M.; Peck, D. D.; Natoli, T. E.; Lu, X.; Gould, J.; Davis, J. F.; Tubelli, A. A.; Asiedu, J. K.; Lahr, D. L.; Hirschman, J. E.; Liu, Z.; Donahue, M.; Julian, B.; Khan, M.; Wadden, D.; Smith, I. C.; Lam, D.; Liberzon, A.; Toder, C.; Bagul, M.; Orzechowski, M.; Enache, O. M.; Piccioni, F.; Johnson, S. A.; Lyons, N. J.; Berger, A. H.; Shamji, A. F.; Brooks, A. N.; Vrcic, A.;

Flynn, C.; Rosains, J.; Takeda, D. Y.; Hu, R.; Davison, D.; Lamb, J.; Ardlie, K.; Hogstrom, L.; Greenside, P.; Gray, N. S.; Clemons, P. A.; Silver, S.; Wu, X.; Zhao, W.-N.; Read-Button, W.; Wu, X.; Haggarty, S. J.; Ronco, L. V.; Boehm, J. S.; Schreiber, S. L.; Doench, J. G.; Bittker, J. A.; Root, D. E.; Wong, B.; Golub, T. R. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. *Cell* **2017**, *171* (6), 1437-1452.e17. https://doi.org/10.1016/j.cell.2017.10.049.

(54) Konieczny, L.; Roterman-Konieczna, I.; Spólnik, P. The Structure and Function of Living Organisms. In *Systems Biology: Functional Strategies of Living Organisms [Internet]. 2nd edition*; Springer, 2023. https://doi.org/10.1007/978-3-031-31557-2_1.

(55) O'Reilly, F. J.; Rappsilber, J. Cross-Linking Mass Spectrometry: Methods and Applications in Structural, Molecular and Systems Biology. *Nat. Struct. Mol. Biol.* **2018**, *25* (11), 1000–1008. https://doi.org/10.1038/s41594-018-0147-0.

(56) Clasen, M. A.; Ruwolt, M.; Wang, C.; Ruta, J.; Bogdanow, B.; Kurt, L. U.; Zhang, Z.; Wang, S.; Gozzo, F. C.; Chen, T.; Carvalho, P. C.; Lima, D. B.; Liu, F. Proteome-Scale Recombinant Standards and a Robust High-Speed Search Engine to Advance Cross-Linking MS-Based Interactomics. *Nat. Methods* **2024**, *21* (12), 2327–2335. https://doi.org/10.1038/s41592-024-02478-1.

(57) Bush, W. S. Introduction to Bioinformatics and Computational Biology. In *Proceedings of the fourteenth international conference on Genetic and evolutionary computation conference companion - GECCO Companion '12*; ACM Press: Philadelphia, Pennsylvania, USA, 2012; p 1141. https://doi.org/10.1145/2330784.2330935.

(58) Stahl, K.; Graziadei, A.; Dau, T.; Brock, O.; Rappsilber, J. Protein Structure Prediction with In-Cell Photo-Crosslinking Mass Spectrometry and Deep Learning. *Nat. Biotechnol.* **2023**, *41* (12), 1810–1819. https://doi.org/10.1038/s41587-023-01704-z.

(59) Gilep, K.; Obarska-Kosinska, A.; Kosinski, J. Improving AlphaFold 3 Structural Modeling by Incorporating Explicit Crosslinks. bioRxiv May 30, 2025, p 2024.12.03.626671. https://doi.org/10.1101/2024.12.03.626671.

(60) Clemens Isert; Michael Pun; Emanuele Rossi; Doug Tischer; Mehmet Akdel; Daniel Kovtun; Marco Pegoraro; Thomas Duignan; Alex Zhang; Vladas Oleinikovas; Graham Holt; Yusuf Adeshina; Patrick Kunzmann; Arjun Ramesh; Douglas Wu; Alex Goncearenco; Lidor Foguel; Dana Felker; Davide Sabbadin; Vivian Lam; Matthias Grass; Zach Carpenter; Michael Bronstein; Luca Naef. *NEO-1 - Decode and Design the Structure of Life*. NEO-1 - Decode and Design the Structure of Life. https://www.proximabio.com/neo-1 (accessed 2025-09-24).

(61) Wilson, D. S.; Keefe, A. D.; Szostak, J. W. The Use of mRNA Display to Select High-Affinity Protein-Binding Peptides. *Proc. Natl. Acad. Sci.* **2001**, *98* (7), 3750–3755. https://doi.org/10.1073/pnas.061028198.

(62) Brückner, A.; Polge, C.; Lentze, N.; Auerbach, D.; Schlattner, U. Yeast Two-Hybrid, a Powerful Tool for Systems Biology. *Int. J. Mol. Sci.* **2009**, *10* (6), 2763–2788. https://doi.org/10.3390/ijms10062763.

(63) Lewis, S.; Hempel, T.; Jiménez-Luna, J.; Gastegger, M.; Xie, Y.; Foong, A. Y. K.; Satorras, V. G.; Abdin, O.; Veeling, B. S.; Zaporozhets, I.; Chen, Y.; Yang, S.; Schneuing, A.; Nigam, J.; Barbero, F.; Stimper, V.; Campbell, A.; Yim, J.; Lienen, M.; Shi, Y.; Zheng, S.; Schulz, H.; Munir, U.; Clementi, C.; Noé, F. Scalable Emulation of Protein Equilibrium Ensembles with Generative Deep Learning. bioRxiv December 5, 2024, p 2024.12.05.626885. https://doi.org/10.1101/2024.12.05.626885.

(64) Kretsch, R. C.; Hummer, A. M.; He, S.; Yuan, R.; Zhang, J.; Karagianes, T.; Cong, Q.; Kryshtafovych, A.; Das, R. Assessment of Nucleic Acid Structure Prediction in CASP16. bioRxiv May 10, 2025, p 2025.05.06.652459. https://doi.org/10.1101/2025.05.06.652459.
(65) Laine, E.; Grudinin, S.; Klypa, R.; Beauchêne, I. C. de. Navigating Protein–Nucleic Acid Sequence-Structure Landscapes with Deep Learning. *Curr. Opin. Struct. Biol.* **2025**, *95*, 103162. https://doi.org/10.1016/j.sbi.2025.103162.
(66) Cordero, P.; Kladwang, W.; VanLang, C. C.; Das, R. The Mutate-and-Map Protocol for Inferring Base Pairs in Structured RNA. *Methods Mol. Biol. Clifton NJ* **2014**, *1086*, 53–77. https://doi.org/10.1007/978-1-62703-667-2_4.
(67) Cheng, C. Y.; Chou, F.-C.; Kladwang, W.; Tian, S.; Cordero, P.; Das, R. Consistent Global Structures of Complex RNA States through Multidimensional Chemical Mapping. *eLife* **2015**, *4*, e07600. https://doi.org/10.7554/eLife.07600.
(68) Wu, T.; Cheng, A. Y.; Zhang, Y.; Xu, J.; Wu, J.; Wen, L.; Li, X.; Liu, B.; Dou, X.; Wang, P.; Zhang, L.; Fei, J.; Li, J.; Ouyang, Z.; He, C. KARR-Seq Reveals Cellular Higher-Order RNA Structures and RNA–RNA Interactions. *Nat. Biotechnol.* **2024**, *42* (12), 1909–1920. https://doi.org/10.1038/s41587-023-02109-8.
(69) He, S.; Huang, R.; Townley, J.; Kretsch, R. C.; Karagianes, T. G.; Cox, D. B. T.; Blair, H.; Penzar, D.; Vyaltsev, V.; Aristova, E.; Zinkevich, A.; Bakulin, A.; Sohn, H.; Krstevski, D.; Fukui, T.; Tatematsu, F.; Uchida, Y.; Jang, D.; Lee, J. S.; Shieh, R.; Ma, T.; Martynov, E.; Shugaev, M. V.; Bukhari, H. S. T.; Fujikawa, K.; Onodera, K.; Henkel, C.; Ron, S.; Romano, J.; Nicol, J. J.; Nye, G. P.; Wu, Y.; Choe, C.; Reade, W.; Participants, E.; Das, R. Ribonanza: Deep Learning of RNA Structure through Dual Crowdsourcing. bioRxiv February 27, 2024, p 2024.02.24.581671. https://doi.org/10.1101/2024.02.24.581671.
(70) He, S.; organizers, C.; experimentalists, C. R.; consortium, R.-P.; team, V.; Kretsch, R.; Hummer, A.; Favor, A.; Reade, W.; Demkin, M.; Das, R.; others. Stanford RNA 3D Folding, 2025. https://kaggle.com/competitions/stanford-rna-3d-folding.
(71) *RibonanzaNet2 alpha release*. https://www.kaggle.com/competitions/stanford-rna-3d-folding/discussion/571704 (accessed 2025-12-14).
(72) Lee, Y.; He, S.; Oda, T.; Rao, G. J.; Kim, Y.; Kim, R.; Kim, H.; Heng, C. K.; Kowerko, D.; Li, H.; Nguyen, H.; Sampathkumar, A.; Gómez, R. E.; Chen, M.; Yoshizawa, A.; Kuraishi, S.; Ogawa, K.; Zou, S.; Paullier, A.; Zhao, B.; Chen, H.-L.; Hsu, T.-A.; Hirano, T.; Chiu, W.; Gezelle, J. G.; Haack, D.; Hong, Y.; Jadhav, S.; Koirala, D.; Kretsch, R. C.; Lewicka, A.; Li, S.; Marcia, M.; Piccirilli, J.; Rudolfs, B.; Srivastava, Y.; Steckelberg, A.-L.; Su, Z.; Toor, N.; Wang, L.; Yang, Z.; Zhang, K.; Zou, J.; Baker, D.; Chen, S.-J.; Demkin, M.; Favor, A.; Hummer, A. M.; Joshi, C. K.; Kryshtafovych, A.; Küçükbenli, E.; Miao, Z.; Moult, J.; Munley, C.; Reade, W.; Viel, T.; Westhof, E.; Zhang, S.; Das, R. Template-Based RNA Structure Prediction Advanced through a Blind Code Competition. bioRxiv December 30, 2025, p 2025.12.30.696949. https://doi.org/10.64898/2025.12.30.696949.
(73) Joshi, C. K.; Gianni, E.; Kwok, S. L. Y.; Mathis, S. V.; Liò, P.; Holliger, P. Generative Inverse Design of RNA Structure and Function with gRNAde. bioRxiv December 1, 2025, p 2025.11.29.691298. https://doi.org/10.1101/2025.11.29.691298.
(74) Thompson, M.; Martín, M.; Olmo, T. S.; Rajesh, C.; Koo, P. K.; Bolognesi, B.; Lehner, B. Massive Experimental Quantification Allows Interpretable Deep Learning of Protein Aggregation. *Sci. Adv.* **2025**, *11* (18), eadt5111. https://doi.org/10.1126/sciadv.adt5111.
(75) Bray, M.-A.; Singh, S.; Han, H.; Davis, C. T.; Borgeson, B.; Hartland, C.; Kost-Alimova, M.; Gustafsdottir, S. M.; Gibson, C. C.; Carpenter, A. E. Cell Painting,

a High-Content Image-Based Assay for Morphological Profiling Using Multiplexed Fluorescent Dyes. *Nat. Protoc.* **2016**, *11* (9), 1757–1774. https://doi.org/10.1038/nprot.2016.105.
(76) Lazar, N. H.; Celik, S.; Chen, L.; Fay, M. M.; Irish, J. C.; Jensen, J.; Tillinghast, C. A.; Urbanik, J.; Bone, W. P.; Gibson, C. C.; Haque, I. S. High-Resolution Genome-Wide Mapping of Chromosome-Arm-Scale Truncations Induced by CRISPR–Cas9 Editing. *Nat. Genet.* **2024**, *56* (7), 1482–1493. https://doi.org/10.1038/s41588-024-01758-y.
(77) White, R. E.; Manitpisitkul, P. Pharmacokinetic Theory of Cassette Dosing in Drug Discovery Screening. *Drug Metab. Dispos. Biol. Fate Chem.* **2001**, *29* (7), 957–966.
(78) Richard, A. M.; Judson, R. S.; Houck, K. A.; Grulke, C. M.; Volarath, P.; Thillainadarajah, I.; Yang, C.; Rathman, J.; Martin, M. T.; Wambaugh, J. F.; Knudsen, T. B.; Kancherla, J.; Mansouri, K.; Patlewicz, G.; Williams, A. J.; Little, S. B.; Crofton, K. M.; Thomas, R. S. ToxCast Chemical Landscape: Paving the Road to 21st Century Toxicology. *Chem. Res. Toxicol.* **2016**, *29* (8), 1225–1251. https://doi.org/10.1021/acs.chemrestox.6b00135.
(79) Gardiner, L.-J.; Carrieri, A. P.; Wilshaw, J.; Checkley, S.; Pyzer-Knapp, E. O.; Krishna, R. Using Human in Vitro Transcriptome Analysis to Build Trustworthy Machine Learning Models for Prediction of Animal Drug Toxicity. *Sci. Rep.* **2020**, *10* (1), 9522. https://doi.org/10.1038/s41598-020-66481-0.
(80) Belair, D. G.; Visconti, R. J.; Hong, M.; Marella, M.; Peters, M. F.; Scott, C. W.; Kolaja, K. L. Human Ileal Organoid Model Recapitulates Clinical Incidence of Diarrhea Associated with Small Molecule Drugs. *Toxicol. In Vitro* **2020**, *68*, 104928. https://doi.org/10.1016/j.tiv.2020.104928.
(81) Ma, Y.; Zhang, Z.; Jia, B.; Yuan, Y. Automated High-Throughput DNA Synthesis and Assembly. *Heliyon* **2024**, *10* (6), e26967. https://doi.org/10.1016/j.heliyon.2024.e26967.
(82) Zhang, A. J.; Burgess, K. Total Syntheses of Vancomycin. *Angew. Chem. Int. Ed Engl.* **1999**, *38* (5), 634–636. https://doi.org/10.1002/(SICI)1521-3773(19990301)38:5%253C634::AID-ANIE634%253E3.0.CO;2-G.
(83) Jones, M.; Goodyear, R. L. High-Throughput Purification in Drug Discovery: Scaling New Heights of Productivity. *ACS Med. Chem. Lett.* **2023**, *14* (7), 916–919. https://doi.org/10.1021/acsmedchemlett.3c00073.
(84) McCorkindale, W.; Filep, M.; London, N.; A. Lee, A.; King-Smith, E. Deconvoluting Low Yield from Weak Potency in Direct-to-Biology Workflows with Machine Learning. *RSC Med. Chem.* **2024**, *15* (3), 1015–1021. https://doi.org/10.1039/D3MD00719G.
(85) Rees, D. C.; Congreve, M.; Murray, C. W.; Carr, R. Fragment-Based Lead Discovery. *Nat. Rev. Drug Discov.* **2004**, *3* (8), 660–672. https://doi.org/10.1038/nrd1467.
(86) Fearon, D.; Powell, A.; Douangamath, A.; Dias, A.; Tomlinson, C. W. E.; Balcomb, B. H.; Aschenbrenner, J. C.; Aimon, A.; Barker, I. A.; Bertram, F.; Brandão-Neto, J.; Coe, P. A.; Collins, P.; Dunnett, L. E.; Fairhead, M.; Gildea, R. J.; Golding, M.; Gorrie-Stone, T.; Hathaway, P. V.; Koekemoer, L.; Krojer, T.; Lithgo, R. M.; Maclean, E. M.; Marples, P. G.; Mikolajek, H.; Ni, X.; Nidamarthi, K. H. V.; O'Donnell, G.; Skyner, R.; Talon, R.; Thompson, W.; Watt, G.; Wild, C. F.; Williams, M. A.; Winokan, M.; Wright, N. D.; Winter, G.; Shotton, E. J.; von Delft, F. Accelerating Drug Discovery With High-Throughput Crystallographic Fragment

Screening and Structural Enablement. *Appl. Res.* **2025**, *4* (1), e202400192. https://doi.org/10.1002/appl.202400192.
(87) Grosjean, H.; Fieseler, K.; Sanchez-Garcia, R.; Thompson, W.; Deane, C. M.; Delft, F. von; Biggin, P. C. Structure-Activity-Relationships Can Be Directly Extracted from High-Throughput Crystallographic Evaluation of Fragment Elaborations in Crude Reaction Mixtures. ChemRxiv May 28, 2025. https://doi.org/10.26434/chemrxiv-2025-bg2ll.
(88) *OpenBind - - Diamond Light Source*. https://www.diamond.ac.uk/Science/Collaborations/openbind.html (accessed 2025-11-08).
(89) Hendrick, C. E.; Jorgensen, J. R.; Chaudhry, C.; Strambeanu, I. I.; Brazeau, J.-F.; Schiffer, J.; Shi, Z.; Venable, J. D.; Wolkenberg, S. E. Direct-to-Biology Accelerates PROTAC Synthesis and the Evaluation of Linker Effects on Permeability and Degradation. *ACS Med. Chem. Lett.* **2022**, *13* (7), 1182–1190. https://doi.org/10.1021/acsmedchemlett.2c00124.
(90) *Science and Tech*. https://www.kimiatx.com/science-tech (accessed 2025-10-28).
(91) Lanman, B. A.; Allen, J. R.; Allen, J. G.; Amegadzie, A. K.; Ashton, K. S.; Booker, S. K.; Chen, J. J.; Chen, N.; Frohn, M. J.; Goodman, G.; Kopecky, D. J.; Liu, L.; Lopez, P.; Low, J. D.; Ma, V.; Minatti, A. E.; Nguyen, T. T.; Nishimura, N.; Pickrell, A. J.; Reed, A. B.; Shin, Y.; Siegmund, A. C.; Tamayo, N. A.; Tegley, C. M.; Walton, M. C.; Wang, H.-L.; Wurz, R. P.; Xue, M.; Yang, K. C.; Achanta, P.; Bartberger, M. D.; Canon, J.; Hollis, L. S.; McCarter, J. D.; Mohr, C.; Rex, K.; Saiki, A. Y.; San Miguel, T.; Volak, L. P.; Wang, K. H.; Whittington, D. A.; Zech, S. G.; Lipford, J. R.; Cee, V. J. Discovery of a Covalent Inhibitor of KRASG12C (AMG 510) for the Treatment of Solid Tumors. *J. Med. Chem.* **2020**, *63* (1), 52–65. https://doi.org/10.1021/acs.jmedchem.9b01180.
(92) Zahm, A. M.; Owens, W. S.; Himes, S. R.; Rondem, K. E.; Fallon, B. S.; Gormick, A. N.; Bloom, J. S.; Kosuri, S.; Chan, H.; English, J. G. Discovery and Validation of Context-Dependent Synthetic Mammalian Promoters. *bioRxiv* **2023**, 2023.05.11.539703. https://doi.org/10.1101/2023.05.11.539703.
(93) *Octant Blog: Miniaturizing Drug Discovery*. https://www.octant.bio/blog-posts/miniaturizing-drug-discovery-using-high-throughput-chemistry (accessed 2025-10-02).
(94) *Octant Blog: Harnessing Multiplexing for Cellular Assay and Reporter Development*. https://www.octant.bio/blog-posts/harnessing-multiplexing-for-cellular-assay-and-reporter-development (accessed 2025-10-02).
(95) Cheng, J.; Novati, G.; Pan, J.; Bycroft, C.; Žemgulytė, A.; Applebaum, T.; Pritzel, A.; Wong, L. H.; Zielinski, M.; Sargeant, T.; Schneider, R. G.; Senior, A. W.; Jumper, J.; Hassabis, D.; Kohli, P.; Avsec, Ž. Accurate Proteome-Wide Missense Variant Effect Prediction with AlphaMissense. *Science* **2023**, *381* (6664), eadg7492. https://doi.org/10.1126/science.adg7492.
(96) Meisburger, S. P.; Case, D. A.; Ando, N. Diffuse X-Ray Scattering from Correlated Motions in a Protein Crystal. *Nat. Commun.* **2020**, *11* (1), 1271. https://doi.org/10.1038/s41467-020-14933-6.
(97) Wayment-Steele, H. K.; Nesr, G. E.; Hettiarachchi, R.; Kariyawasam, H.; Ovchinnikov, S.; Kern, D. Learning Millisecond Protein Dynamics from What Is Missing in NMR Spectra. bioRxiv March 19, 2025, p 2025.03.19.642801. https://doi.org/10.1101/2025.03.19.642801.

(98) Northcutt, C. G.; Athalye, A.; Mueller, J. Pervasive Label Errors in Test Sets Destabilize Machine Learning Benchmarks. arXiv November 7, 2021. https://doi.org/10.48550/arXiv.2103.14749.
(99) Wellnitz, J.; Ahmad, S.; Begale, N.; Joseph, J.; Zeng, H.; Bolotokova, A.; Dong, A.; Reza, S.; Ghiabi, P.; Elisa, G.; Cheng, X.; Tu, G.; Li, X.; Liu, J.; Dou, D.; Li, J.; Harding, R. J.; Edwards, A. M.; Haibe-Kains, B.; Halabelian, L.; Tropsha, A.; Couñago, R. Enabling Open Machine Learning of DNA Encoded Library Selections to Accelerate the Discovery of Small Molecule Protein Binders. ChemRxiv October 17, 2024. https://doi.org/10.26434/chemrxiv-2024-xd385.
(100) Gujral, O.; Bafna, M.; Alm, E.; Berger, B. Sparse Autoencoders Uncover Biologically Interpretable Features in Protein Language Model Representations. *Proc. Natl. Acad. Sci.* **2025**, *122* (34), e2506316122. https://doi.org/10.1073/pnas.2506316122.
(101) Zhang, Z.; Wayment-Steele, H. K.; Brixi, G.; Wang, H.; Peraro, M. D.; Kern, D.; Ovchinnikov, S. Protein Language Models Learn Evolutionary Statistics of Interacting Sequence Motifs. bioRxiv January 31, 2024, p 2024.01.30.577970. https://doi.org/10.1101/2024.01.30.577970.
(102) Wang, N.; Fang, C.; Bissell, M.; Jain, A.; Balsam, D.; 2026. *Using Interpretability to Identify a Novel Class of Alzheimer's Biomarkers*. https://www.goodfire.ai/research/interpretability-for-alzheimers-detection (accessed 2026-01-30).
(103) Göbel, U.; Sander, C.; Schneider, R.; Valencia, A. Correlated Mutations and Residue Contacts in Proteins. *Proteins* **1994**, *18* (4), 309–317. https://doi.org/10.1002/prot.340180402.
(104) Meisburger, S. P.; Case, D. A.; Ando, N. Robust Total X-Ray Scattering Workflow to Study Correlated Motion of Proteins in Crystals. *Nat. Commun.* **2023**, *14* (1), 1228. https://doi.org/10.1038/s41467-023-36734-3.
(105) Chou, S. *From systems operators to systems architects*. Seemay's blog. https://seemay.substack.com/p/from-systems-operators-to-systems (accessed 2025-09-24).
(106) Beltran, A.; Jiang, X.; Shen, Y.; Lehner, B. Site-Saturation Mutagenesis of 500 Human Protein Domains. *Nature* **2025**, *637* (8047), 885–894. https://doi.org/10.1038/s41586-024-08370-4.
(107) Meier, J.; Rao, R.; Verkuil, R.; Liu, J.; Sercu, T.; Rives, A. Language Models Enable Zero-Shot Prediction of the Effects of Mutations on Protein Function. bioRxiv July 10, 2021, p 2021.07.09.450648. https://doi.org/10.1101/2021.07.09.450648.
(108) Peck, D.; Crawford, E. D.; Ross, K. N.; Stegmaier, K.; Golub, T. R.; Lamb, J. A Method for High-Throughput Gene Expression Signature Analysis. *Genome Biol.* **2006**, *7* (7), R61. https://doi.org/10.1186/gb-2006-7-7-r61.
(109) Emmerich, C. H.; Gamboa, L. M.; Hofmann, M. C. J.; Bonin-Andresen, M.; Arbach, O.; Schendel, P.; Gerlach, B.; Hempel, K.; Bespalov, A.; Dirnagl, U.; Parnham, M. J. Improving Target Assessment in Biomedical Research: The GOT-IT Recommendations. *Nat. Rev. Drug Discov.* **2021**, *20* (1), 64–81. https://doi.org/10.1038/s41573-020-0087-3.
(110) McNamee, L. M.; Walsh, M. J.; Ledley, F. D. Timelines of Translational Science: From Technology Initiation to FDA Approval. *PLoS ONE* **2017**, *12* (5), e0177371. https://doi.org/10.1371/journal.pone.0177371.
(111) Scannell, J. W.; Bosley, J.; Hickman, J. A.; Dawson, G. R.; Truebel, H.; Ferreira, G. S.; Richards, D.; Treherne, J. M. Predictive Validity in Drug Discovery:

What It Is, Why It Matters and How to Improve It. *Nat. Rev. Drug Discov.* **2022**, *21* (12), 915–931. https://doi.org/10.1038/s41573-022-00552-x.
(112) Geeleher, P.; Cox, N. J.; Huang, R. S. Clinical Drug Response Can Be Predicted Using Baseline Gene Expression Levels and in Vitro Drug Sensitivity in Cell Lines. *Genome Biol.* **2014**, *15* (3), R47. https://doi.org/10.1186/gb-2014-15-3-r47.
(113) *How to build the virtual cell with artificial intelligence: Priorities and opportunities: Cell*. https://www.cell.com/cell/fulltext/S0092-8674(24)01332-1 (accessed 2025-10-29).
(114) Jansen, R.; Embden, J. D. A. van; Gaastra, W.; Schouls, L. M. Identification of Genes That Are Associated with DNA Repeats in Prokaryotes. *Mol. Microbiol.* **2002**, *43* (6), 1565–1575. https://doi.org/10.1046/j.1365-2958.2002.02839.x.
(115) Park, B.-S.; Lee, M.; Kim, J.; Kim, T. Perturbomics: CRISPR–Cas Screening-Based Functional Genomics Approach for Drug Target Discovery. *Exp. Mol. Med.* **2025**, *57* (7), 1443–1454. https://doi.org/10.1038/s12276-025-01487-0.
(116) Vazquez, F.; Sellers, W. R. Are CRISPR Screens Providing the Next Generation of Therapeutic Targets? *Cancer Res.* **2021**, *81* (23), 5806–5809. https://doi.org/10.1158/0008-5472.CAN-21-1784.
(117) *Gordian Biotechnology Keystone 2025 - Heart - Martin - YouTube*. https://www.youtube.com/watch?v=MVhaSfktffw (accessed 2025-09-25).
(118) Dang, C. V.; Reddy, E. P.; Shokat, K. M.; Soucek, L. Drugging the "undruggable" Cancer Targets. *Nat. Rev. Cancer* **2017**, *17* (8), 502–508. https://doi.org/10.1038/nrc.2017.36.
(119) Gironda-Martínez, A.; Donckele, E. J.; Samain, F.; Neri, D. DNA-Encoded Chemical Libraries: A Comprehensive Review with Succesful Stories and Future Challenges. *ACS Pharmacol. Transl. Sci.* **2021**, *4* (4), 1265–1279. https://doi.org/10.1021/acsptsci.1c00118.
(120) Reiher, C. A.; Schuman, D. P.; Simmons, N.; Wolkenberg, S. E. Trends in Hit-to-Lead Optimization Following DNA-Encoded Library Screens. *ACS Med. Chem. Lett.* **2021**, *12* (3), 343–350. https://doi.org/10.1021/acsmedchemlett.0c00615.
(121) Iqbal, S.; Jiang, W.; Hansen, E.; Aristotelous, T.; Liu, S.; Reidenbach, A.; Raffier, C.; Leed, A.; Chen, C.; Chung, L.; Sigel, E.; Burgin, A.; Gould, S.; Soutter, H. H. Evaluation of DNA Encoded Library and Machine Learning Model Combinations for Hit Discovery. *Npj Drug Discov.* **2025**, *2* (1), 5. https://doi.org/10.1038/s44386-025-00007-4.
(122) Kleinsasser, M.; Quigley, I. *Good binding data is all you need*. Leash. https://leashbio.substack.com/p/good-binding-data-is-all-you-need (accessed 2025-11-10).

View Article Online
DOI: 10.1039/D6SC01189F

Appendix

Below we present an expanded example list of black-box data sources that have been combined with ML. We group them by the corresponding application area.

Structural Biology

Large-scale sequence pretraining extracts structure-relevant co-evolutionary motifs

The use of large-scale sequence pretraining to infer protein structure, exemplified by ESMFold[21], can also be seen as an application of the black-box data paradigm. Generating sequence data is significantly cheaper and higher throughput than determining experimental 3D structure. A single protein sequence on its own is "black-box" in the context of structure - it does not directly measure structural information. However, large collections of coevolved sequences can contain implicit structural information, where residues in contact are more likely to coevolve. This data is traditionally extracted via MSA which aligns sequences as rows in a table where each column contains the observed evolutionary variance for a given residue position. Columns showing covariance are more likely to be in contact[103]. Rather than explicitly extracting this via MSA, language models can learn a similar covariance pattern implicitly from large, unaligned sequence databases[101]. For example, ESM2 is trained on hundreds of millions of unaligned protein sequences. Due to this, models often show emergent structure prediction performance. ESM2 is able to predict contacts with a simple linear head calibrated on as little as 20 structures[21]. These language models can then also be used to replace MSAs in folding models (such as ESMFold or SimpleFold) that show strong folding performance.

Haze and halos around Bragg peaks are not experimental noise but contain diffuse scattering information encoding protein dynamics

Current structural biology pipelines based on X-ray crystallography obtain exact structural measurements, and measures of experimental variance (such as B-factors), from electron density maps that are extracted from Bragg diffraction peaks[104]. However, often there is an additional signal in these diffraction patterns not from pure periodic arrangements of crystals, called "diffuse scattering", giving information about dynamics[96]. This signal manifests as haze or halos around Bragg peaks and is typically discarded as noise in traditional structural analysis.

View Article Online
DOI: 10.1039/D6SC01189F

It arises from correlated atomic motions and deviations from perfect periodicity in macromolecular crystals. "The Diffuse Project" was recently proposed as a systematic effort to collect this previously discarded signal at a large scale and make it available for machine learning models[105]. They propose a shared representation trained on Bragg, diffuse, and molecular dynamics data. They are also building a standardized *Diffuse Data Bank* with raw frames, processed diffuse volumes, paired PDBs, and processing scripts - laying the foundation for pretraining machine-learning models on experimental dynamics at scale.

Protein biophysical properties

Massively scaled protein stability mutagenesis data via enzyme-complementation

A related approach to the experiment by the Rocklin lab was published by the Lehner lab[106]. They conducted a large-scale site-saturation mutagenesis of human domains to create the "Human Domainome 1". They designed a yeast library with 1.23 million single–amino-acid substitutions spanning 1248 protein domains. Variant stability was read out with an abundance Protein Fragment Complementation Assay (aPCA): each domain variant was fused to a fragment of dihydrofolate reductase (DHFR), such that in-cell domain stability controlled the concentration of DHF. This drove yeast growth at a rate linearly proportional to enzyme abundance correlating variant stability with sequencing readcounts. This yielded 563,534 variant–abundance measurements across 522 domains, correlating strongly with independent biophysical measurements and the MEGAscale experiment. They combined the data with the protein language model ESM1v[107] correlating measured stability scores with predicted fitness scores, identifying mutations whose effect on fitness was not explained by stability. Outlier mutations were strongly enriched in functional sites like DNA/protein-binding interfaces, yielding additional structural readouts when combined with Deep Learning models.

Biological function

View Article Online
DOI: 10.1039/D5SC01189F

Purifying selection unlocks unsupervised variant effect prediction by AlphaMissense through weak labelling

Understanding biological function, in the context of the complex cellular, tissue and whole organism environment is critical for understanding disease and finding effective treatments. This can require high-content readouts in cell lines or even ideally in-vivo clinical contexts that are difficult to scale versus more isolated biophysical readouts. Additionally, it is often difficult to reproduce disease cellular states in-vitro. Both AlphaMissense and the CMap effort try to address the paucity of data in this data. The AlphaMissense team applied a related insight to Wayment-Steele, El Nesr et al. (extracting signal from noise, Trick 7) even earlier to predict the biological pathogenicity of protein variants[95]. The central observation in their case is that purifying selection eliminates harmful variants from populations. Therefore, the absence of a particular amino acid at a specific residue position across a large set of homologous sequences indicates a non-tolerated substitution rather than missing data. Conversely, substitutions frequently observed in natural sequence variation are likely benign. Rather than relying on relatively small, clinically annotated variant datasets such as ClinVar, this allows one to construct weakly labelled data from observed and unobserved variants at much larger scale. The result is a continuous pathogenicity score that reflects the extent to which a given substitution deviates from tolerated evolutionary patterns. AlphaMissense is a modified version of AF2[19] initially trained on a combination of structure prediction and MSA masked-language modeling. This already provides a variant likelihood task that can be seen as a second "black-box" data source. The weakly-labeled variant data is then used in a second training stage to then train a variant-effect prediction head.

Sparse landmark genes can be used to approximate full transcriptomic response to perturbation at increased scale

The Connectivity Map (CMap) is an older example of increasing throughput by sparsifying the readout (Trick 1), in this case for large perturbation-response RNA expression datasets[53]. It contains gene expression signatures from human cells exposed to 42,080 perturbations (19,811 small molecule compounds, 18,493 shRNAs, 3,462 cDNAs, and 314 biologics). Its high throughput stems from the sparse L1000 assay, which measures a pre-defined reduced set of landmark genes per sample via an array instead of performing bulk or single-cell RNA

View Article Online
DOI: 10.1039/D0SC01189F

sequencing. This significantly reduces cost and increases throughput (1.3 million profiles) – 978 "landmark" transcripts are measured (selected to span diverse pathways and be representative for co-expressed genes) – in addition to 80 invariant controls. This is done via ligation-mediated amplification (LMA) and bead-based fluorescent detection[108]. In CMap, cells from a standardized nine-line panel are exposed to small molecules or genetic perturbations over multiple concentrations and short exposure times. Transcriptome-wide values for the 11,350 non-measured genes are inferred by regression trained on external RNA-sequencing compendia, with 9,196 of 11,350 inferred genes (81%) achieving correlations in the top 95th percentile relative to their RNA-sequencing counterparts, indicating adequate performance.

Target identification

CRISPR combined with ML can unlock scalable in-vivo target identification

Target identification is one of the most critical and costly steps in drug discovery[109]. It frequently requires decades of fundamental biological research[110]. One of the key limitations is that most experiments can only ethically be conducted in humans once the disease is somewhat understood and treatments have been shown effective in animals. The discovery of targets often has to be done in-vitro or in organisms with short reproductive cycles to reduce time. The predictive validity of these model systems (especially cell lines) is not ideal[111]. Many black-box data methods hence focus on inferring human effects based on in-vitro or model systems[112]. Using cheap readouts, one can train models (e.g. virtual cells[113]) to predict the effects of different genetic perturbations or drug candidates on these biological systems much more cheaply than collecting human data.

Clustered Regularly Interspaced Short Palindromic Repeats[114] (CRISPR) is a naturally occurring adaptive immune system in bacteria that has been repurposed as a programmable genome-editing technology and promises to help address the target discovery problem. Its key components are a Cas effector protein, most commonly Cas9, which acts as a nuclease, and a guide RNA (gRNA) that directs Cas9 to a specific genomic locus through base-pairing with the target DNA. CRISPR has revolutionized the scope and scale at which we can conduct genetic interventions to elucidate the biological function of proteins. Nowadays, CRISPR is heavily used in functional genomics applications to understand how gene-level perturbations, often obtained in high-throughput, pooled fashion, affect phenotypes[115]. This has already led to the discovery of novel clinical targets[116]. Traditionally, these screens have been conducted

View Article Online
DOI: 10.1039/D5SC01189F

most frequently in cell lines due to their cost and scalability. In-vivo CRISPR has recently emerged allowing for local genetic perturbations in the complete biological context of a living organism in a pooled fashion to test hundreds or thousands of target hypotheses[50]. Perturbations are generally done at low transfection multiples such that most cells only obtain a single perturbation and perturbed cells are isolated by non-perturbed tissue to localize perturbation effects, acting as a proxy readout for the full organism (Trick 5). The perturbagen is tracked through barcodes. Phenotypic readouts e.g. single-cell transcriptomics, spatial profiling, imaging or others, enable the reconstruction of genotype–phenotype relationships within a physiologically relevant milieu. The startup Gordian Biotechnology is further leveraging perturbation-local single-cell sequencing readouts to predict in-vivo disease responses for a given perturbation using paired disease-response and sequencing readouts[117].

Hit Finding

DNA-encoded libraries

Once a disease-related target is identified, molecules that engage with the target are needed. For small molecules in particular, this step has historically been a key bottleneck, with many well validated targets still being considered undruggable, although this list is becoming successively smaller[118]. Hit Finding has traditionally often had a strong High-Throughput Screening (HTS) focus. Increasingly, these methods are also developed specifically with Machine Learning in mind. DNA-encoded libraries (DELs) are large pooled combinatorial collections (10^6–10^9 molecules) in which each member is covalently linked to a DNA "barcode" encoding its identity[119]. They are assembled via a split-and-pool approach, which combines a library of barcode-linked fragments into a large combinatorial library of multi-fragment barcoded small molecules. Pooled screens can be done via affinity selection against a protein target and hits are identified via sequencing, applying Trick 2. Often, however, DEL-derived molecules are large or require further local search to identify more lead-like molecules. Reiher et al., for example, report around 400–1100 Da for DEL hits versus roughly 200–700 Da for hits from the Janssen HTS library[120]. Additionally, the count-based readouts from DNA-encoded library screens can be noisy due to truncates, unspecific binding, low read depth, and other biases. To combat this, one of the central ideas of black-box data is again applied here: a model is trained on the noisy data to extract the signal, which can then be used on "clean"

View Article Online
DOI: 10.1039/D6SC01189F

purchasable chemical libraries. This idea was applied in multiple studies across diverse targets below.

McCloskey[51] et al. screened three targets (soluble epoxide hydrolase, estrogen receptor-α, c-KIT), trained classifiers purely on DEL selections (graph convolutional models trained on 355,804, 74,741, and 50,186 positive training examples for sEH, ERα, and c-KIT), virtually screened 88 million catalog compounds, tested 2,000, and obtained a 30% hit rate at 30 μM with sub-10 nM actives for every target.

Iqbal et al.[121] targeted CK1α/δ with three differently sized libraries: MS (10 million), DD (11 million), HG (1 billion). They trained five models on orthosteric-binding site enriched outputs and screened a 140,000-member test set. This yielded 808 compounds tested via SPR with 80 hits (9.9%), including two nanomolar (187 nM, 69.6 nM).

Wellnitz et al.[99] screened WDR91 via HitGen OpenDEL (3 billion) and applied the resulting models on Enamine REAL (37 billion members). They nominated 50 molecules, tested 48, and confirmed 7 hits (14.5%, 2.7–21 μM) and generated co-crystal structures with WDR91.

The startup Leash Bio has scaled DEL against many targets and recently unveiled *Hermes*, a model achieving competitive binding prediction accuracy on non-DEL data despite exclusively being trained on DEL readouts[122].

No primary research results, software or code have been included and no new data were generated or analysed as part of this Perspective paper.

View Article Online
DOI: 10.1039/D6SC01189F