Recent advances and future trends for protein–small molecule interaction predictions with protein language models

Alexander Kroll, Yvan Rousset

Article - Version of Record

Suggested Citation:
Kroll, A., & Rousset, Y. (2025). Recent advances and future trends for protein–small molecule interaction predictions with protein language models. Current Opinion in Structural Biology, 93, Article 103070. https://doi.org/10.1016/j.sbi.2025.103070

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Structural Biology

Recent advances and future trends for protein–small molecule interaction predictions with protein language models

Alexander Kroll and Yvan Rousset

In recent years, the application of natural language models to protein amino acid sequences, referred to as protein language models (PLMs), has demonstrated a significant potential for uncovering hidden patterns related to protein structure, function, and stability. The critical functions of proteins in biological processes often arise through interactions with small molecules; central examples are enzymes, receptors, and transporters. Understanding these interactions is particularly important for drug design, for bioengineering, and for understanding cellular metabolism. In this review, we present state-of-the-art PLMs and explore how they can be integrated with small molecule information to predict protein-small molecule interactions. We present several such prediction tasks and discuss current limitations and potential areas for improvement.

Address
Heinrich-Heine-University, Universitätsstraße 1, Düsseldorf, 40225, NRW, Germany

Corresponding author: Kroll, Alexander (alexander.kroll@hhu.de)

Current Opinion in Structural Biology 2025, **93**:103070

This review comes from a themed issue on **Sequences and Topology 2025**

Edited by **Arne Elofsson** and **Rachel Kolodny**

For a complete overview see the Issue and the Editorial

Available online xxx

https://doi.org/10.1016/j.sbi.2025.103070

Introduction

The determination of protein–small molecule interactions is important in many scientific and industrial fields; for example, it is important for pharmaceutical and biotechnological research because the activities of most enzymes and drugs depend on protein–small molecule interactions [1,2]. Unfortunately, it is typically time-consuming and costly to determine such interactions experimentally. Accurate machine learning (ML) prediction models can greatly accelerate this process by predicting various aspects of protein–small molecule interactions, such as specific properties of protein–small molecule combinations or by identifying promising candidate protein–small molecule pairs that are likely to interact. Thereby, prediction models can significantly reduce the large number of potential protein–small molecule interaction pairs to a feasible number of pairs that can be tested experimentally.

Predicting protein–small molecule interactions with machine learning models requires the generation of informative and meaningful numerical representations of the proteins and small molecules. The state-of-the-art general purpose method for numerical representations of proteins are protein language models (PLMs), which are deep learning models originally developed for natural language text. Similar to natural language, where the arrangement of words follows grammatical rules and must result in a meaningful sentence, protein sequences follow specific constraints on the arrangement of amino acids to result in a functional protein. Therefore, the same methodology that is used in natural language processing (NLP) can be successfully applied to protein sequences.

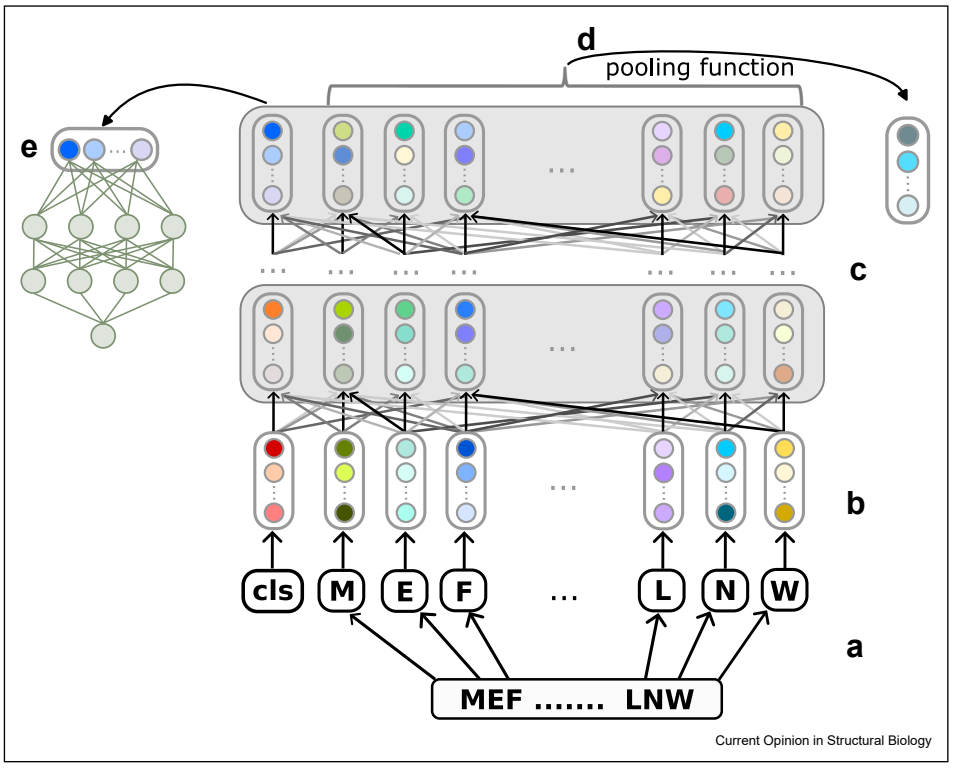
In this review, we discuss and present the various methods and applications that have recently been used for predicting protein–small molecule interactions using PLMs. We begin with an introduction to PLMs, followed by a discussion of numerical representation techniques for small molecules. We then examine different ways in which these numerical representations can be combined to develop protein–small molecule interaction models. Finally, we present several areas of application, followed by a discussion of current limitations and potential ways to improve predictive capabilities.

Protein–small molecule interaction models

Protein language models

The state-of-the-art general purpose method for numerical representations of protein sequences is using PLMs, in particular protein transformer encoders [3]. These models process protein sequences by partitioning them into smaller subsequences called tokens. The most common method is amino acid-level tokenization (Figure 1a), but alternative strategies exist and have been explored [4]. Each token is initially represented by

Figure 1



Forward process of a protein language model. (a) A protein sequence is divided in its individual amino acids. An additional classification representation, cls, can be added if the model is to be trained for a specific protein prediction task. **(b)** Each aminoacid is mapped to an initial vector representation that encodes the type of the amino acid and its position in the input sequence. Different colors represent different numeric values. **(c)** Multiple transformer encoder layers are applied to update all representations by incorporating information from other representations. Different darknesses of the arrows indicate that different amounts of information from different representations are used. **(d)** After the update steps, a learnable or predefined pooling function can be applied to convert all updated amino acid representations into a single vector that can be used as the whole protein representation. **(e)** Alternatively, the classification representation can be used as input to a feedforward network to predict the function or property of the protein.

a separate numeric vector representing its type and position within the sequence (Figure 1b). The objective of the encoder is to improve all representations by incorporating information from other amino acids within the sequence (Figure 1c). The specific way in which the amino acids are updated and what information is drawn from other amino acids is determined by update functions that are learned during the training phase of the model.

The most common training task for PLMs is to randomly mask a fraction of the amino acids in a protein sequence and train the model to predict the type of those amino acids using information from the unmasked amino acids. Most models use a default masking rate of ∼15% but optimal rates can vary; for example, larger models tend to perform better with higher masking rates [5]. The recently developed ESM-3 model introduced a noise schedule that varies the masking rate during pretraining [6].

Meta AI's ESM models [7,8], especially the ESM-2 series with models ranging from 8 million to 15 billion trainable parameters, are the most widely used PLMs. The ESM-2 models were trained on a dataset with 65 million different protein sequences. In addition, Elnaggar et al. [9] developed two notable and also widely used PLMs: ProtBERT-BFD, with 420 million parameters trained on 2.1 billion protein sequences, and ProtT5, with 3 billion parameters trained on 45 million proteins. All of the above models were trained using the masking strategy described above, i.e., predicting the type of amino acids that were masked in the input sequence.

PLMs trained in this manner can compute updated vector representations of each amino acid in a given input protein sequence. To represent the complete protein, it is desirable to compute a single vector that summarizes this amino acid-specific information. A common way to achieve this is to apply a pooling function to all updated amino acid representations (Figure 1d). This can be a predefined function such as the element-wise average of all representations, which results in a loss of information but still captures important structural and functional protein characteristics [9].

Alternatively, a pretrained PLM can be further trained for a specific downstream prediction task, a process known as fine-tuning [10]. During this process, the model learns an appropriate pooling function, or the model is trained to store all relevant information in an additional vector that represents the whole protein, the so-called classification representation (Figure 1e). After fine-tuning, the resulting representations can serve as task-specific protein embeddings. While fine-tuning large PLMs typically improves performance [10], it is computationally expensive and memory-intensive. Parameter-efficient fine-tuning methods, such as LoRA, mitigate these requirements by updating only a smaller fraction of parameters and have shown promising results in both NLP and protein language modeling [11].

The recent advances in protein structure prediction with models such as AlphaFold 2, RoseTTAFold, and ESMFold have enabled highly accurate structure predictions from protein sequences for many proteins [12–14]. Building on these developments, several approaches now incorporate 3D structural information as input to protein language models. For example, DeepFRI and ESM-GearNet have integrated graph neural networks (GNNs) to capture amino acid connectivity [15,16]. These GNNs process protein sequences similar to standard PLMs: amino acids are tokenized, but information exchange between tokens is restricted to amino acids that are spatially close in the 3D protein structure. This helps the model focus on relevant amino acids when updating its representations, leading to slight performance improvements in some protein prediction tasks [15].

More recently, the ESM-3 model [6] has introduced a new approach by tokenizing both the protein sequence and its 3D structure in the model input, allowing it to generate representations for both sequence- and structure-based tasks. These representations can be extracted and used for downstream protein prediction tasks [17]. While ESM-3 could be fine-tuned for applications such as protein–small molecule interaction prediction, no such studies have been published to date.

Numerical representations of small molecules

Numerical representations of small molecules can be derived using neural networks or expert-designed methods. This subsection provides a brief overview of the most common approaches. One common approach is training GNNs, which represent small molecules as graphs, where the atoms of the molecule are interpreted as graph nodes and the bonds as edges [18,19]. Each atom and bond is encoded as a numerical vector that is iteratively updated by the GNN based on neighboring atom and bond information (Figure 2a). A second way to learn small molecule representations with neural networks is through transformer encoders, which can process small molecule SMILES strings that encode the small molecule structure including its stereochemistry (Figure 2b) [20,21].

Similar to PLMs, both GNNs and small molecule transformers can be pretrained by masking parts of the input and predicting the identity of what is masked by extracting information from the remaining input. Alternatively, the models can be trained to predict easily computable molecular descriptors, or, given sufficient training data, for a specific prediction task of interest. After training, a single numerical vector can be extracted for each small molecule by applying a pooling function, such as the element-wise average, to all updated numerical representations (Figures 2a and b).

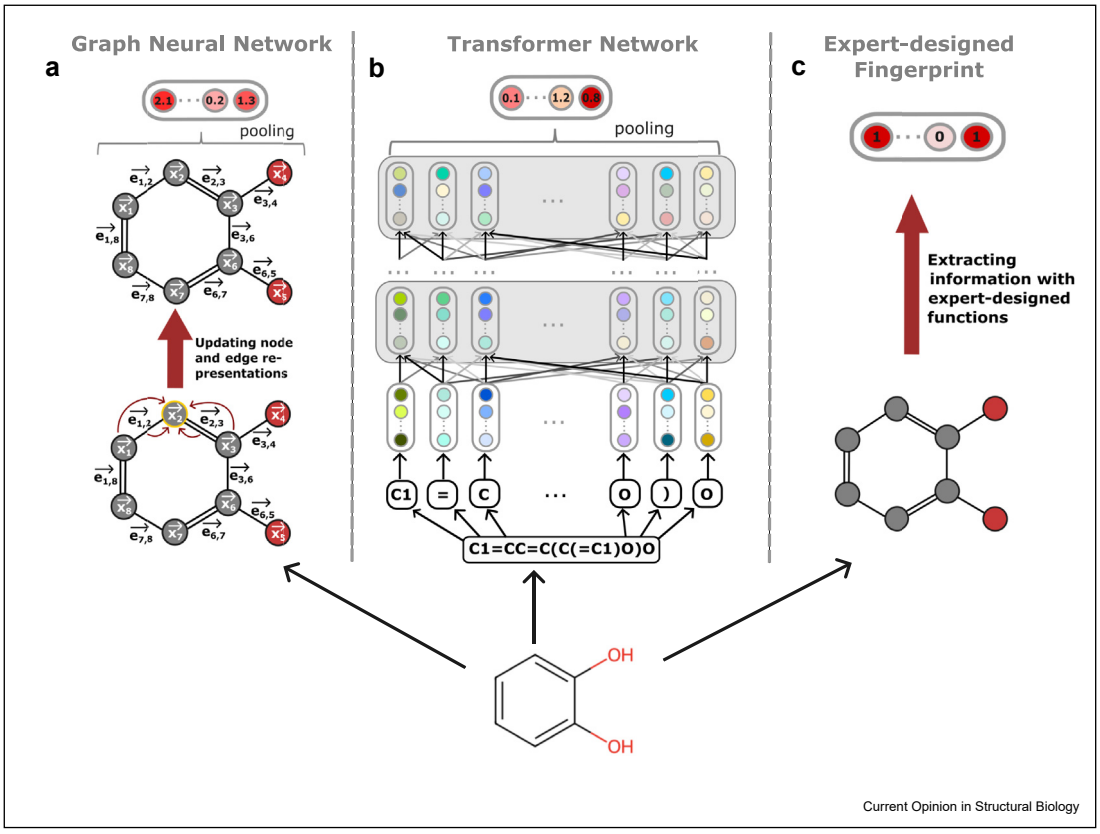
Alternatively, expert-designed methods can be used to encode small molecule information without ML. These methods are typically based on graph representations of the molecule and produce binary vectors that encode the presence or absence of specific substructures (Figure 2c) [22,23]. While ML-based methods generally provide better numerical representations [24], they also require more computational resources because they involve training models with millions of parameters on large datasets.

Protein–small molecule interaction predictions with PLMs

Protein–small molecule interaction prediction models fall into two categories. Approaches in the first category use pretrained deep learning models or expert-designed fingerprints to generate representations of proteins and small molecules. These representations are concatenated into an input vector for a separate ML prediction model, typically a small feedforward neural network or a gradient boosting decision tree (Figure 3a), with the latter often performing slightly better for such prediction tasks [25,26]. During training, the prediction model extracts relevant information from the fixed input vectors. Although the representations are not fine-tuned, they work well with limited data because they often capture essential features, and fine-tuning with small datasets is typically less effective and can easily lead to overfitting [27].

The second category involves end-to-end training of a deep learning model to achieve two goals simultaneously: (i) generating task-specific molecule representations and (ii) providing predictions (Figures 3b and c). “End-to-end training” refers to the simultaneous adjustment of model parameters for tasks (i) and (ii) during the same process. The generation of task-specific representations is often based on further parameter tuning of a pretrained deep learning model (see Section 2.1). Unlike the first approach, this method extracts more task-relevant information from the molecules

Figure 2



Approaches for generating numerical representations of small molecules. (a) To process small molecules with graph neural networks, the small molecule is interpreted as a graph with atoms represented as nodes of the graph and bonds represented as edges. Each node and edge is represented by a numeric vector. All node vector representations are iteratively updated by extracting information from neighboring bonds and atoms. A pooling function is applied to all updated node vectors to obtain a single graph representation. **(b)** A small molecule is represented by a SMILES string that encodes the molecular structure. The SMILES string is subdivided, and each subpart is represented by a numeric vector. All vectors are passed through a small molecule transformer encoder to update the representations. A pooling function is applied to all updated token vectors to obtain a single small molecule representation. **(c)** The small molecule is represented as a graph, and expert-designed functions are applied to extract structural information. A binary vector, called a molecular fingerprint, stores the extracted information.

but requires larger datasets and more computational resources.

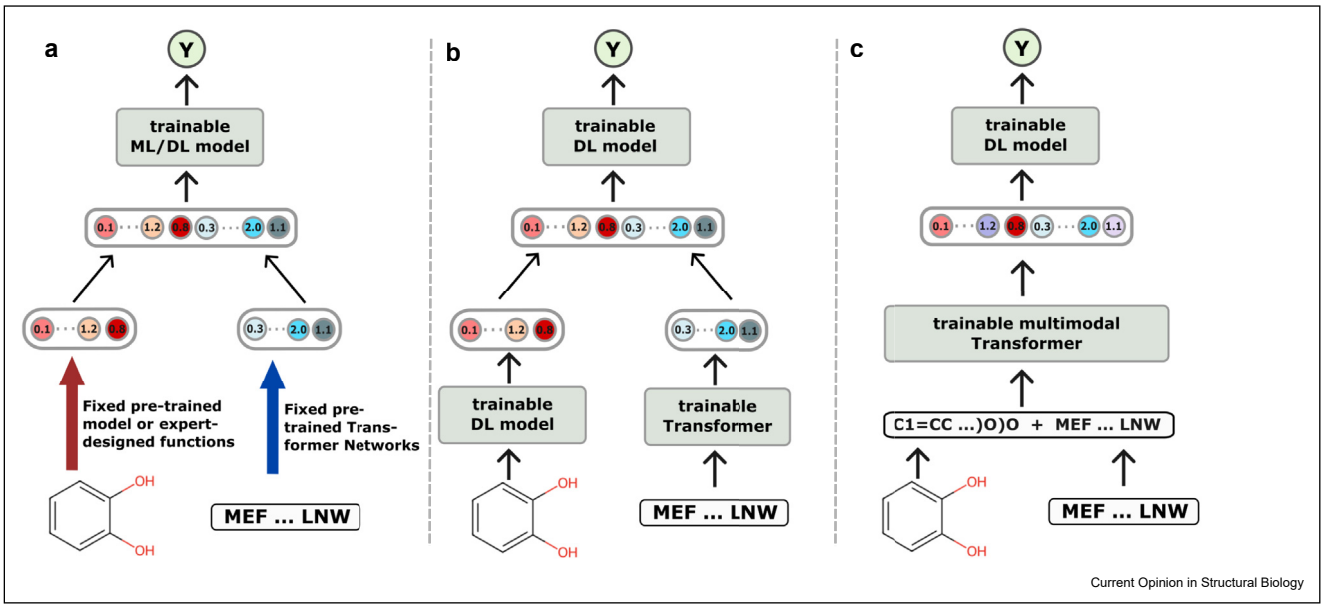
Traditionally, end-to-end models first consist of two separate but parallel blocks, each responsible for generating the protein and small molecule vector representations, respectively. Only after the numerical vector representations are generated, another deep learning block uses both representations to provide a prediction for the protein–small molecule interaction of interest (Figure 3b). However, generating numerical representations for the protein and the small molecule separately, without considering their interactions, is likely to result in suboptimal representations. Recently, ProSmith [28] and ESM-AA [29] have proposed to combine both the protein sequence and the small molecule structural information in the input of a single multimodal transformer network to generate a joint numerical representation (Figure 3c). This allows a better exchange of information during the representation generation process, and it allows capturing complex relationships and interactions between the two different types of molecules.

Alternative approaches for protein–small molecule interaction modeling

Although this review focuses on the adaptation of PLMs for predicting protein–small molecule interactions, in this subsection we provide a brief overview of the alternative computational approaches, docking, co-folding, and molecular dynamics (MD) simulations, for modeling these interactions.

Docking methods predict how a small molecule binds within the binding site of a protein by using sampling algorithms and scoring functions to identify the lowest energy conformation [30]. Docking is widely used in structure-based virtual screening for drug discovery

Figure 3



Protein–small molecule interaction predictions can be achieved by different approaches. (a) Small molecule representations are computed using an expert-designed method or a pretrained deep learning model, a GNN, or a transformer network. A protein representation is computed using a pretrained transformer network. Both representations are concatenated to produce a single vector containing both small molecule and protein information. This vector is used as input to a deep learning or machine learning model, such as a gradient boosting model. This model is trained for a protein–small molecule interaction prediction task. **(b)** Small molecule representations are generated using a deep learning model with trainable parameters. In parallel, protein representations are computed using a transformer network with trainable parameters. The resulting small molecule and protein vectors are concatenated and fed to a deep learning model, such as a feedforward neural network, which outputs a prediction. All trainable parameters are adjusted in the same training process. **(c)** A small molecule SMILES string and the protein amino acid sequence are fed as input to the same transformer network with trainable parameters. This model can account for the interactions of the protein and small molecule while generating a common numerical representation. The resulting vector is used as input to a trainable deep learning model to provide a prediction. All trainable parameters are adjusted in the same training process. GNN, graph neural network.

[31], helping to identify potential drug candidates. Its accuracy depends on the scoring function used to estimate binding affinity and typically requires prior knowledge of the binding site.

Co-folding methods predict protein–ligand complexes by integrating protein folding with ligand docking, often using deep learning. Recent transformer-based models such as AlphaFold 3 (AF3) citeabramson2024 and RoseTTAFold All-Atom (RFAA) [32] are the most prominent models in this area. Both methods build on previous protein structure predictors (AlphaFold 2 and RoseTTAFold) and extend them to predict biomolecular complexes, including proteins bound to small molecules, nucleic acids, and ions. AF3 and RFAA use end-to-end learning and large structural datasets to model interactions. Co-folding methods can provide accurate predictions, but they require extensive training data and struggle to predict affinities for unseen ligands [33]. Furthermore, while the models can provide confidence scores that correlate with the quality of the binding pose, they are primarily optimized for structure prediction and cannot be easily fine-tuned for downstream protein tasks.

MD simulations provide a time-resolved view of atomic interactions within protein–ligand complexes, capturing conformational dynamics and binding kinetics [34]. They use force fields to compute atomic interactions and integrate equations of motion to model molecular trajectories. While valuable for refining docking predictions, MD simulations are computationally expensive, particularly for high-throughput screening or slow binding processes [35].

In contrast, traditional PLMs do not require a protein's 3D structure and can be easily fine-tuned to predict not only whether binding occurs but also different types of interactions, such as inhibition, activation, or catalytic activity. However, the black box nature of PLMs limits the interpretability of the underlying binding mechanisms.

Protein–small molecule interaction prediction tasks

Predicting enzyme kinetic parameters

Protein–small molecule interaction models are essential for predicting enzyme kinetic parameters such as turnover numbers k_{cat} and Michaelis constants K_M, which

define an enzyme's catalytic rate and affinity for its substrate(s), respectively. Knowledge of these parameters is important for characterizing the catalytic properties of enzymes and for parameterizing genome-scale metabolic models. Traditionally, missing kinetic parameters have been estimated using data from closely related enzymes with measured kinetic parameters, but recently developed ML models have demonstrated superior performance [36].

TurNuP and EITLEM are the state-of-the-art for k_{cat} prediction [37,36]. TurNuP uses protein embeddings from the pretrained ESM-1b model and expert-designed small molecule fingerprints (Figure 3a) [36], while EITLEM uses transfer learning to learn from related tasks and fine-tunes a protein transformer network (Figure 3b). For enzymes with less than 40 % sequence identity compared with all training enzymes, a naive homology-based inference, i.e. averaging over the k_{cat} values of the most similar training enzymes, results in predicting only 2 % of the variance in k_{cat} values [36]. In contrast, TurNuP and EITLEM can explain about a third of the variance for those enzymes [37,36].

Although DLKcat [38] reports higher overall performance metrics than TurNuP and EITLEM, the model generalizes poorly to unseen enzymes. It has been shown that DLKcat performs worse than a simple homology-based approach for enzymes with less than 60 % sequence identity compared with all training enzymes [39].

For K_M prediction, the current state-of-the-art models EITLEM and ProSmith$_{KM}$ [28,37] achieve a coefficient of determination R^2 greater than 0.5. R^2 measures the proportion of the variance in the observed values that is explained by the predictions and thus these models can predict more than half of the variance in K_M values. UniKP shows slightly lower performance and uncertain generalizability to unseen enzymes [40].

The enzyme specificity constant k_{cat}/K_M is a valuable but less frequently predicted kinetic parameter, likely due to limited training data compared with k_{cat} and K_M [37,41]. Predicting this constant, as done in UniKP and EITLEM [40,37], has several advantages: k_{cat}/K_M can be measured directly under certain conditions, typically with higher accuracy than K_M measurements, which are often estimated by curve fitting. More reliable input data improves model performance, allowing k_{cat}/K_M models to explain more variance in observations with less training data [37,40].

Small molecule scope of proteins

Substrate scope of enzymes

Determining substrates for enzymes is critical for pharmaceutical research and bioengineering, including drugs, food, and biofuel production [42]. The largest protein database, UniProt, has high-quality annotations including the substrate(s) for only 1 % of the 36 million enzymes it stores [43]. Recent advances in PLMs have led to the development of enzyme substrate prediction models, helping to identify whether a small molecule is a substrate for a given enzyme [44,29,28,45].

Enzyme–substrate prediction models can be specific to a protein family [46] or general to all enzymes. General models, which are typically more accurate [44], are ideally trained on all available experimentally validated enzyme–substrate pairs. The enzyme substrate prediction (ESP) model [44], the first general enzyme–substrate prediction model, uses a fine-tuned version of the ESM-1b model to generate protein representations and a GNN to generate task-specific small molecule representations. Both vectors are concatenated and input to a gradient boosting binary classifier (Figure 3a), achieving a prediction accuracy of 91.5 %. This accuracy was improved to 92.3 % by ESM-AA [29] and to 94.2 % by ProSmith [28]. ESM-AA and ProSmith use multimodal transformers to facilitate the exchange of relevant information between the protein and small molecules during computation of their numerical representations (Figure 3c). Recently, FusionESP further improved prediction accuracy to 94.8 % on the same test set by integrating contrastive learning to generate more discriminative substrate and non-substrate representations [45]. The performance of all approaches drops for substrates not seen during model training. For example, ProSmith's Matthews correlation coefficient (MCC) drops from 0.85 for training known substrates to 0.29 for unseen substrates.

Current enzyme–substrate prediction models achieve a true positive rate of ~80% and a false positive rate of ~5% [28,44]. For an enzyme of unknown function, if 200 potential substrate candidates are tested, of which only one is a true substrate, the models will incorrectly identify about 10 molecules as substrates and correctly identify the true substrate 80 % of the time.

Substrate scope of transporters

Transport proteins make up only ~10% of all cellular proteins [47] and are even less well studied than enzymes in terms of structure and function. Recent transporter–substrate prediction models assess the likelihood whether a small molecule is a substrate for a given transporter [48,49]. The SPOT model [48], similar in approach to the enzyme–substrate prediction model ESP achieves a recall of 83.1 % and a precision of 88.0 % on an independent test set. In contrast, a naive homology-based approach, which assigns substrates based on the three most similar transporters with known functions, yields a recall of 80.9 % and a much lower precision of 56.8 % on the same test set [48].

Drug–target interactions

Identifying small molecule interactions with target proteins is a key challenge in drug discovery, which has traditionally relied on inefficient and costly high-throughput screening. Advances in AI are reducing costs by accelerating the identification of drug candidates, improving predictions of efficacy and safety, and enabling drug repurposing. Even small gains in predictive accuracy can save tens to hundreds of millions of dollars per drug by reducing late-stage failures [50]. Exscientia, for example, reported an 80 % cost reduction and 70 % faster development process using AI [51].

Machine learning (ML)-based drug–target interaction (DTI) prediction focuses on predicting binding affinity, inhibition, and other key interactions to guide drug design. Many advances and novel prediction approaches have been developed and published in the last two years. For example, ConPLex uses contrastive learning on PLM-derived embeddings to distinguish true drug–target interactions from non-binding compounds [52]. NHGNN-DTA and PGraphDTA combine PLM-based sequence data with protein structure data, using graph neural networks or protein contact maps to refine affinity predictions [53,54]. DTI-LM and MIFAM-DTI, on the other hand, integrate ESM protein embeddings with small molecule embeddings using graph attention networks (GATs) [55,56]. All of these approaches generate separate embeddings for proteins and small molecules (Figure 3b) but some allow information flow between the two types of molecules.

The recent trends highlight the importance of direct information exchange between proteins and small molecules. ProSmith and ESM-AA [28,29] fine-tune multimodal transformers to process both types of molecules in a common input sequence (Figure 3c), allowing easy and direct information exchange between the two types of molecules.

Predicting variant effects

Bioengineering of proteins to improve protein properties or to obtain novel functions is a key task in bioindustry. PLMs have been used to predict protein variants with desired properties [57,6,58]; for example, DLKcat [38] and UniKP [40] integrate small molecule information with PLMs to predict kinetic parameters of enzyme variants. However, it has been shown that DLKcat primarily estimates the average k_{cat} value of training mutants for the same enzyme rather than accurately predicting mutation effects [39]. When evaluating predictions beyond this average, a negative correlation emerges, highlighting the limitations of current PLM-based models in capturing mutation effects on enzyme kinetics.

On the other hand, sequence alignment-based methods such as GEMME [59] and SIFT [60] can predict mutation effects and outperform some PLM-based approaches [61]. GEMME has not yet been used to predict mutation effects on enzyme kinetics but used for the related task of identifying functionally critical residues and those essential for thermodynamic stability [61]. The recent efforts to combine GEMME with PLMs have further improved prediction performance [62] and may also offer a promising way to improve predictions of mutation effects on kinetic parameters.

Discussion

The development of successful prediction models for protein–small molecule interactions depends primarily on two key components: training datasets and model architecture, including molecule representation methods. While over 200 million proteins have been sequenced [43], which can successfully be used for unsupervised pretraining of PLMs, for downstream prediction tasks, current models are often limited by insufficient training data, both in terms of quantity and quality [63]. For example, Bar-Even et al. [64] found that up to 20 % of entries in BRENDA [41], the main resource for experimental K_M and k_{cat} values, differ from their reference papers, likely due to copying errors and misinterpretation of units. Beyond such obvious errors, measurements of kinetic parameters can be highly variable. In large kinetic databases, k_{cat} values for the same enzyme–reaction pair often differ severalfold between measurements from different labs [36].

To accurately assess the performance of a model, it is important to evaluate its ability to generalize to unseen proteins. This requires careful test set construction, and simply randomly splitting the data can be misleading, especially when datasets contain many related proteins, such as enzyme variants. Including proteins in the test set that are highly similar to those in the training set can inflate performance metrics while masking poor predictive power for truly novel proteins as demonstrated by some models performing worse than baseline comparisons when tested on dissimilar sequences [38,39]. Therefore, to ensure a fair assessment of generalization, test sets should ideally consist of proteins with low maximum sequence similarity (e.g. below 20–40 %) to any protein used during training.

An important choice in the construction of protein–small molecule interaction prediction models is the method of protein encoding. While the trend has favored ever larger models—exemplified by the growth within the ESM series from ESM-1b's 650 million parameters to ESM-2's 15 billion and ESM-3's 98 billion—recent evidence suggests that performance gains may be plateauing. Studies comparing models with a wide range of parameters, including the ESM-2 family, have found that medium-sized models perform as well as their much larger counterparts on many biological

benchmarks. For example, for predicting mutational fitness effects, an ESM-2 model with 150 million parameters showed slightly better performance than a 20-fold larger ESM-2 variant with 3 billion parameters [65]. Coupled with the significant computational cost of training and deploying ultra-large models, which often leads researchers to opt for smaller, more practical versions, these observations suggest that simply increasing model size without changing architecture or training techniques may yield diminishing returns.

Beyond predicting interaction likelihoods or kinetic parameters of protein–small molecule pairs, modeling the accurate three-dimensional structure of protein–small molecule interactions remains a critical challenge. Approaches such as docking provide structure-based predictions but often depend on known binding sites and reliable scoring functions [31]. More recently, co-folding techniques, exemplified by AlphaFold 3 [66] and RoseTTAFold All-Atom [32], have emerged as powerful tools capable of generating high-resolution geometric predictions (poses) for protein–ligand complexes, demonstrating high accuracy, particularly for systems similar to those in their training data [33]. However, their ability to generalize to truly novel ligand or pocket types can be limited by their reliance on learned patterns, and they can be computationally intensive. In contrast, sequence-based PLM approaches, while typically lacking detailed geometric precision, have greater ability to predict interaction likelihood or strength and show stronger generalization, especially in low-data scenarios or for novel entities, making them suitable for large-scale screening [28].

The choice between co-folding approaches and PLM-based methods often depends on the research goal: co-folding can provide highly accurate structure prediction for known systems, PLMs allow a large-scale screening and better generalization capabilities. Increasingly, the field is moving toward hybrid models that integrate the strengths of both, such as using PLM embeddings within structure-aware graph networks [15,16] or enhancing co-folding models with sequence-level insights, aiming to bridge the gap between structural detail and predictive generalization [67].

While protein–small molecule interaction prediction methods that incorporate 3D structure—often by encoding amino acid connectivity through graph neural networks—show only marginal performance gains over sequence-based approaches [15,16,68], this may be because sequence-based PLMs already capture essential protein structural information at the amino acid level. More complementary approaches also incorporate the 3D atomic structure of the protein, showing greater performance gains over alternative methods [69]. These methods work well even when no experimental protein structure data are available, but only protein structure prediction methods can be used [69]. This suggests that a future trend may be to integrate methods that encode the 3D atomic structure of proteins.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank Martin J. Lercher for insightful discussions and useful comments. The authors are funded by the European Union through a grant (ERC AdG "MechSys"—Project ID 101055141) awarded to Martin J. Lercher.

Data availability

No data was used for the research described in the article.

References

Papers of particular interest, published within the period of review, have been highlighted as:

* of special interest
** of outstanding interest

1. Adrio JL, Demain AL: **Microbial enzymes: tools for biotechnological processes**. *Biomolecules* 2014, **4**:117–139.
2. Bull SC, Doig AJ: **Properties of protein drug target classes**. *PLoS One* 2015, **10**, e0117955.
3. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I: **Attention is all you need**. *Adv Neural Inf Process Syst* 2017, **30**.
4. Pourmirzaei M, Esmaili F, Pourmirzaei M, Wang D, Xu D: **Prot2token: a multi-task framework for protein language processing using autoregressive language modeling**. *bioRxiv* 2024:2024. 05.
5. Wettig A, Gao T, Zhong Z, Chen D: **Should you mask 15% in masked language modeling?** *arXiv preprint* 2022. arXiv: 2202.08005.
6. Hayes T, Rao R, Akin H, Sofroniew NJ, Oktay D, Lin Z, Verkuil R, Tran VQ, Deaton J, Wiggert M, *et al.*: **Simulating 500 million years of evolution with a language model**. *Science* 2025, eads0018.
7. Rives A, Meier J, Sercu T, Goyal S, Lin Z, Liu J, Guo D, Ott M, Zitnick CL, Ma J, *et al.*: **Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences**. *Proc Natl Acad Sci USA* 2021, **118**, e2016239118.
8. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, *et al.*: **Evolutionary-scale prediction of atomic-level protein structure with a language model**. *Science* 2023, **379**:1123–1130.
9. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, Bhowmik D, Rost B: **Prottrans: toward understanding the language of life through self-supervised learning**. *IEEE Trans Pattern Anal Mach Intell* 2022, **44**:7112–7127.
10. Schmirler R, Heinzinger M, Rost B: **Fine-tuning protein language models boosts predictions across diverse tasks**. *Nat Commun* 2024, **15**:7407.
11. Sledzieski S, Kshirsagar M, Baek M, Dodhia R, Lavista Ferres J, Berger B: **Democratizing protein language models with**

parameter-efficient fine-tuning. *Proc Natl Acad Sci USA* 2024, **121**, e2405840121.

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

13. Baek M, DiMaio F, Anishchenko I, Dauparas J, Ovchinnikov S, Lee GR, Wang J, Cong Q, Kinch LN, Schaeffer RD, Millán C, Park H, Adams C, Glassman CR, DeGiovanni A, Pereira JH, Rodrigues AV, van Dijk AA, Ebrecht AC, Opperman DJ, Sagmeister T, Buhlheller C, Pavkov-Keller T, Rathinaswamy MK, Dalwadi U, Yip CK, Burke JE, Garcia KC, Grishin NV, Adams PD, Read RJ, Baker D: **Accurate prediction of protein structures and interactions using a three-track neural network**. *Science* 2021, **373**:871–876, https://doi.org/10.1126/science.abj8754.

14. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, dos Santos Costa A, Fazel-Zarandi M, Sercu T, Candido S, Rives A: **Evolutionary-scale prediction of atomic-level protein structure with a language model**. *Science* 2023, **379**:1123–1130, https://doi.org/10.1126/science.ade2574.

15. Zhang Z, Wang C, Xu M, Chenthamarakshan V, Lozano A, Das P, Tang J: **A systematic study of joint representation learning on protein sequences and structures**. *arXiv preprint* 2023. arXiv:2303.06275.

16. Gligorijević V, Renfrew PD, Kosciolek T, Leman JK, Berenberg D, Vatanen T, Chandler C, Taylor BC, Fisk IM, Vlamakis H, *et al.*: **Structure-based protein function prediction using graph convolutional networks**. *Nat Commun* 2021, **12**:3168.

17. Zhang D, Zeng Y, Hong X, Xu J: **Leveraging multimodal protein representations to predict protein melting temperatures**. *arXiv preprint arXiv:2412.04526* 2024.

18. Wang Y, Li Z, Barati Farimani A: **Graph neural networks for molecules**. In *Int. J. Mol. Sci.* Springer; 2023:21–66.

19. Heid E, Greenman KP, Chung Y, Li S-C, Graff DE, Vermeire FH, Wu H, Green WH, McGill CJ: **Chemprop: a machine learning package for chemical property prediction**. *J Chem Inf Model* 2023, **64**:9–17.

20. Ahmad W, Simon E, Chithrananda S, Grand G, Ramsundar B: **Chemberta-2: towards chemical foundation models**. *arXiv preprint* 2022. arXiv:2209.01712.

21. Ross J, Belgodere B, Chenthamarakshan V, Padhi I, Mroueh Y, Das P: **Large-scale chemical language representations capture molecular structure and properties**. *Nat Mach Intell* 2022, **4**:1256–1264.

22. Rogers D, Hahn M: **Extended-connectivity fingerprints**. *J Chem Inf Model* 2010, **50**:742–754.

23. Durant JL, Leland BA, Henry DR, Nourse JG: **Reoptimization of mdl keys for use in drug discovery**. *J. Chem. Inf. Comput.* 2002, **42**:1273–1280.

24. Yang K, Swanson K, Jin W, Coley C, Eiden P, Gao H, Guzman-Perez A, Hopper T, Kelley B, Mathea M, *et al.*: **Analyzing learned molecular representations for property prediction**. *J Chem Inf Model* 2019, **59**:3370–3388.

25. Grinsztajn L, Oyallon E, Varoquaux G: **Why do tree-based models still outperform deep learning on typical tabular data?** *Adv Neural Inf Process Syst* 2022, **35**:507–520.

26. Kroll A, Engqvist MK, Heckmann D, Lercher MJ: **Deep learning allows genome-scale prediction of michaelis constants from structural features**. *PLoS Biol* 2021, **19**, e3001402.

27. Lin Y, Ma X, Chu X, Jin Y, Yang Z, Wang Y, Mei H: **Lora dropout as a sparsity regularizer for overfitting control**. *arXiv preprint* 2024. arXiv:2404.09610.

28. ** Kroll A, Ranjan S, Lercher MJ: **A multimodal transformer network for protein-small molecule interactions enhances predictions of kinase inhibition and enzyme-substrate relationships**. *PLoS Comput Biol* 2024, **20**, e1012100.
ProSmith is the first multimodal transformer network for processing proteins and small molecules in the same input sequence.

29. ** Zheng K, Long S, Lu T, Yang J, Dai X, Zhang M, Nie Z, Ma W-Y, Zhou H: **Esm all-atom: multi-scale protein language model for unified molecular modeling**. *arXiv preprint* 2024. arXiv:2403.12995.
ESM-AA is a multimodal transformer network for processing proteins and small molecules in the same input.

30. Meng X-Y, Zhang H-X, Mezei M, Cui M: **Molecular docking: a powerful approach for structure-based drug discovery**. *Curr Comput Aided Drug Des* 2011, **7**:146–157.

31. Grinter SZ, Zou X: **Challenges, applications, and recent advances of protein-ligand docking in structure-based drug design**. *Molecules* 2014, **19**:10150–10176.

32. * Krishna R, Wang J, Ahern W, Sturmfels P, Venkatesh P, Kalvet I, Lee GR, Morey-Burrows FS, Anishchenko I, Humphreys IR, *et al.*: **Generalized biomolecular modeling and design with rosettafold all-atom**. *Science* 2024, **384**, eadl2528.
RosettaFold All-Atom is an extension of RoseTTAFold enabling all-atom modeling of protein-ligand complexes, capable of generating detailed structural predictions.

33. Škrinjar P, Eberhardt J, Durairaj J, Schwede T: **Have protein-ligand co-folding methods moved beyond memorisation?** *bioRxiv* 2025:2025. 02.

34. Durrant JD, McCammon JA: **Molecular dynamics simulations and drug discovery**. *BMC Biol* 2011, **9**:1–9.

35. Shoichet BK: **Virtual screening of chemical libraries**. *Nature* 2004, **432**:862–865.

36. Kroll A, Rousset Y, Hu X-P, Liebrand NA, Lercher MJ: **Turnover number predictions for kinetically uncharacterized enzymes using machine and deep learning**. *Nat Commun* 2023, **14**:4139, https://doi.org/10.1038/s41467-023-39840-4.

37. ** Shen X, Cui Z, Long J, Zhang S, Chen B, Tan T: **Eitlem-kinetics: a deep-learning framework for kinetic parameter prediction of mutant enzymes**. *Chem Catal* 2024, **4**.
EITLEM improves the prediction of enzyme kinetic parameters by fine-tuning a PLM via transfer learning.

38. Li F, Yuan L, Lu H, Li G, Chen Y, Engqvist MKM, Kerkhoven EJ, Nielsen J: **Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction**. *Nat Catal* 2022, **5**:662–672, https://doi.org/10.1038/s41929-022-00798-z.

39. * Kroll A, Lercher MJ: **Dlkcat cannot predict meaningful kcat values for mutants and unfamiliar enzymes**. *Biol. Methods Protoc.* 2024, bpae061.
This study discusses the limitations of protein small molecule models when test sets are not properly generated.

40. Yu H, Deng H, He J, Keasling JD, Luo X: **Unikp: a unified framework for the prediction of enzyme kinetic parameters**. *Nat Commun* 2023, **14**:8211, https://doi.org/10.1038/s41467-023-44113-1.

41. Placzek S, Schomburg I, Chang A, Jeske L, Ulbrich M, Tillack J, Schomburg D: **Brenda in 2017: new perspectives and new tools in brenda**. *Nucleic Acids Res* 2016, gkw952.

42. Kell DB, Swainston N, Pir P, Oliver SG: **Membrane transporter engineering in industrial biotechnology and whole cell biocatalysis**. *Trends Biotechnol* 2015, **33**:237–246.

43. Consortium TU: **UniProt: the universal protein knowledgebase in 2021**. *Nucleic Acids Res* 2020, **49**:D480–D489, https://doi.org/10.1093/nar/gkaa1100.

44. Kroll A, Ranjan S, Engqvist MK, Lercher MJ: **A general model to predict small molecule substrates of enzymes based on machine and deep learning**. *Nat Commun* 2023, **14**:2787.

45. ** Du Z, Fu W, Guo X, Caragea D, Li Y: **Fusionesp: improved enzyme-substrate pair prediction by fusing protein and chemical knowledge**. *bioRxiv* 2024:2024. 08.
FusionESP achieves state-of-the-art performance for enzyme–substrate pair prediction via contrastive learning.

46. Goldman S, Das R, Yang KK, Coley CW: **Machine learning modeling of family wide enzyme-substrate specificity screens**. *PLoS Comput Biol* 2022, **18**, e1009853.

47. Quick M, Javitch JA: **Monitoring the function of membrane transport proteins in detergent-solubilized form**. *Proc Natl Acad Sci USA* 2007, **104**:3603–3608.

48. Kroll A, Niebuhr N, Butler G, Lercher MJ: **Spot: a machine learning model that predicts specific substrates for transport proteins**. *PLoS Biol* 2024, **22**, e3002807.

49. Ataei S, Butler G: **Predicting the specific substrate for transmembrane transport proteins using bert language model**. In *2022 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*; 2022:1–8.

50. Scannell JW, Bosley J, Hickman JA, Dawson GR, Truebel H, Ferreira GS, Richards D, Treherne JM: **Predictive validity in drug discovery: what it is, why it matters and how to improve it**. *Nat Rev Drug Discov* 2022, **21**:915–931, https://doi.org/10.1038/s41573-022-00552-x.

51. **Amazon web services, Exscientia uses generative AI on AWS to advance**. *Drug Discovery* 2025. URL, https://aws.amazon.com/de/solutions/case-studies/exscientia-generative-ai/. Accessed 3 March 2025; 2025.

52. Singh R, Sledzieski S, Bryson B, Cowen L, Berger B: **Contrastive learning in protein language space predicts interactions between drugs and protein targets**. *Proc Natl Acad Sci USA* 2023, **120**, e2220778120, https://doi.org/10.1073/pnas.2220778120.

53. Bal R, Xiao Y, Wang W: **Pgraphdta: improving drug target interaction prediction using protein language models and contact maps**. *arXiv preprint* 2023. arXiv:2310.04017.

54. * He H, Chen G, Chen CY-C: **Nhgnn-dta: a node-adaptive hybrid graph neural network for interpretable drug–target binding affinity prediction**. *Bioinformatics* 2023, **39**, btad355.
NHGNN-DTA combines protein sequence and protein structure information for drug-target affinity prediction.

55. * Ahmed KT, Ansari MI, Zhang W: **Dti-lm: language model powered drug–target interaction prediction**. *Bioinformatics* 2024, **40**:btae533, https://doi.org/10.1093/bioinformatics/btae533.
DTI-LM integrates PLM embeddings with small molecule embeddings using graph attention networks.

56. * Li J, Sun L, Liu L, Li Z: **Mifam-dti: a drug-target interactions predicting model based on multi-source information fusion and attention mechanism**. *Front Genet* 2024, **15**, https://doi.org/10.3389/fgene.2024.1381997.
MIFAM-DTI integrates PLM embeddings with small molecule embeddings using graph attention networks.

57. Nijkamp E, Ruffolo JA, Weinstein EN, Naik N, Madani A: **Progen2: exploring the boundaries of protein language models**. *Cell Syst* 2023, **14**:968–978.e3, https://doi.org/10.1016/j.cels.2023.10.002.

58. 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**. *Adv Neural Inf Process Syst* 2021, **34**:29287–29303.

59. Laine E, Karami Y, Carbone A: **Gemme: a simple and fast global epistatic model predicting mutational effects**. *Mol Biol Evol* 2019, **36**:2604–2619.

60. Ng PC, Henikoff S: **Sift: predicting amino acid changes that affect protein function**. *Nucleic Acids Res* 2003, **31**:3812–3814.

61. Carbone A, Laine E, Lombardi G: **Unlocking protein evolution insights: efficient and interpretable mutational effect predictions with gemme**. *bioRxiv* 2024:2024. 08.

62. Marquet C, Schlensok J, Abakarova M, Rost B, Laine E: **Expert-guided protein language models enable accurate and blazingly fast fitness prediction**. *Bioinformatics* 2024, **40**, btae621.

63. Fan FJ, Shi Y: **Effects of data quality and quantity on deep learning for protein-ligand binding affinity prediction**. *Bioorg Med Chem* 2022, **72**, 117003.

64. Bar-Even A, Noor E, Savir Y, Liebermeister W, Davidi D, Tawfik DS, Milo R: **The moderately efficient enzyme: evolutionary and physicochemical trends shaping enzyme parameters**. *Biochemistry* 2011, **50**:4402–4410.

65. Vieira LC, Handojo ML, Wilke CO: **Scaling down for efficiency: medium-sized protein language models perform well at transfer learning on realistic datasets**. *bioRxiv* 2025:2024. 11.

66. * Abramson J, Adler J, Dunger J, Evans R, Green T, Pritzel A, Ronneberger O, Willmore L, Ballard AJ, Bambrick J, *et al.*: **Accurate structure prediction of biomolecular interactions with alphafold 3**. *Nature* 2024, **630**:493–500.
AlphaFold 3 extends AlphaFold 2 to predict protein-ligand and biomolecular complex structures with high accuracy, but is limited in generalization to novel ligands or protein binding pockets.

67. Bryant P, Kelkar A, Guljas A, Clementi C, Noé F: **Structure prediction of protein-ligand complexes from sequence information with umol**. *Nat Commun* 2024, **15**:4536.

68. Meng L, Wang X: **Tawfn: a deep learning framework for protein function prediction**. *Bioinformatics* 2024, **40**, btae571.

69. Yuan Q, Tian C, Song Y, Ou P, Zhu M, Zhao H, Yang Y: **Gpsfun: geometry-aware protein sequence function predictions with language models**. *Nucleic Acids Res* 2024, **52**:W248–W255.