Check for updates

A comprehensive review of emerging approaches in machine learning for *de novo* PROTAC design

Yossra Gharbi [ID] and Rocío Mercado [ID] *

Targeted protein degradation (TPD) is a rapidly growing field in modern drug discovery that aims to regulate the intracellular levels of proteins by harnessing the cell's innate degradation pathways to selectively target and degrade disease-related proteins. This strategy creates new opportunities for therapeutic intervention in cases where occupancy-based inhibitors have not been successful. Proteolysis-targeting chimeras (PROTACs) are at the heart of TPD strategies, which leverage the ubiquitin–proteasome system for the selective targeting and proteasomal degradation of pathogenic proteins. This unique mechanism can be particularly useful for dealing with proteins that were once deemed "undruggable" using conventional small-molecule drugs. PROTACs are hetero-bifunctional molecules consisting of two ligands, connected by a chemical linker. As the field evolves, it becomes increasingly apparent that traditional methodologies for designing such complex molecules have limitations. This has led to the use of machine learning (ML) and generative modeling to improve and accelerate the development process. In this review, we aim to provide a thorough exploration of the impact of ML on *de novo* PROTAC design – an aspect of molecular design that has not been comprehensively reviewed despite its significance. Initially, we delve into the distinct characteristics of PROTAC linker design, underscoring the complexities required to create effective bifunctional molecules capable of TPD. We then examine how ML in the context of fragment-based drug design (FBDD), honed in the realm of small-molecule drug discovery, is paving the way for PROTAC linker design. Our review provides a critical evaluation of the limitations inherent in applying this method to the complex field of PROTAC development. Moreover, we review existing ML works applied to PROTAC design, highlighting pioneering efforts and, importantly, the limitations these studies face. By offering insights into the current state of PROTAC development and the integral role of ML in PROTAC design, we aim to provide valuable perspectives for biologists, chemists, and ML practitioners alike in their pursuit of better design strategies for this new modality.

Department of Computer Science and Engineering, Section for Data Science and AI, Chalmers University of Technology, Chalmersplatsen 4, 412 96 Gothenburg, Sweden. E-mail: yossra@chalmers.se; rocio.mercado@chalmers.se

Yossra Gharbi is a PhD student in the Data Science and AI division in the AI Laboratory for Molecular Engineering at Chalmers University of Technology. There, she focuses on the development of generative models to engineer proteolysis-targeting chimeras (PROTACs) for drug discovery. Her research explores the application of machine learning to streamline the design process of new drugs, and accelerate the development of next-generation therapeutics.

Yossra Gharbi

Rocío Mercado is an assistant professor in the Data Science and AI division at Chalmers University of Technology. She leads the AI Laboratory for Molecular Engineering in the Department of Computer Science and Engineering, where she and her team seek to bridge methods from machine learning, chemistry, and the life sciences to engineer molecular systems for therapeutic applications and sustainable materials, focusing on new AI method development. She received her PhD in Chemistry from UC Berkeley in 2018, and her BS in Chemistry from Caltech in 2013.

Rocío Mercado

1 Introduction

Targeted protein degradation (TPD) is a novel therapeutic approach with attractive potential to eliminate disease-causing proteins from within cells.[1–4] Traditional drug development strategies have focused on inhibiting the activity of such proteins, but TPD goes a step further by removing or reducing protein levels from the cell. This is particularly useful for targeting proteins that are difficult to inhibit with small molecules or biologics, often due to the absence of well-defined binding sites; these are frequently referred to as "undruggable" targets, and they can be challenging to target due to their structure, location, and/or function.[5,6] Proteolysis-targeting chimeras (PROTACs) are hetero-bifunctional molecules engineered to bind simultaneously to an E3 ligase, a key enzyme involved in the process of tagging proteins for degradation,[7] and the protein of interest (POI) that is targeted for degradation (Fig. 1a).[8–10] A PROTAC molecule brings the E3 ligase and the POI into close proximity, facilitating the formation of a ternary complex consisting of the E3 ligase system, the PROTAC, and the POI to induce POI ubiquitination and its subsequent degradation by the proteasome (Fig. 1b).[11]

1.1 Milestone events in PROTAC technology

PROTACs were first reported in 2001, when the first fully synthesized PROTAC, named Protac-1, was developed by Crews, Deshaies, and co-workers.[12] This provided an *in vitro* proof-of-concept, which proved the feasibility of designing molecules with the potential to selectively target and degrade cellular proteins by hijacking the ubiquitin proteasome system (UPS). Protac-1 was specifically designed to target the methionyl aminopeptidase 2 (MetAP-2) protein, which plays a role in angiogenesis and various other pathologies, including cancer. Protac-1 was designed to target MetAP-2, as ovalicin and fumagillin do, but with the added mechanism of promoting its degradation. The binding of Protac-1 to MetAP-2 led to the tethering of MetAP-2 to beta-transducin repeat-containing protein (β-TrCP), functioning as an E3 ubiquitin ligase responsible for ubiquitination of MetAP-2. The effectiveness of Protac-1 in facilitating the ubiquitination of MetAP-2 was demonstrated using extracts from unfertilized *Xenopus laevis* eggs, a common model organism in biomedical research.[13] These extracts provided a controlled environment rich in cellular machinery that mimics the conditions inside a living cell needed for ubiquitination, protein degradation, and
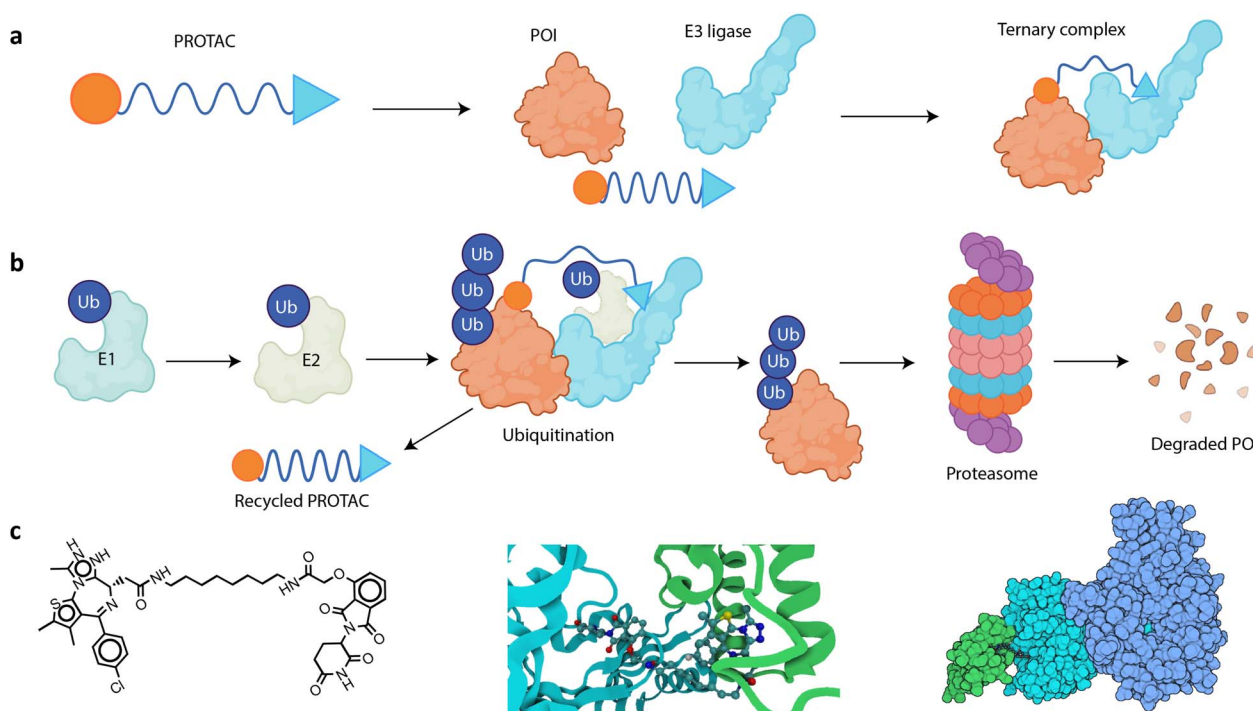

Fig. 1 (a) A PROTAC is a hetero-bifunctional molecule, consisting of a ligand (blue triangle) that recruits an E3 ubiquitin ligase, a warhead (orange circle) that binds to the POI, and a linker (blue curve) that connects the two binding moieties. The PROTAC functions by simultaneously binding to the POI and the E3 ligase, thus bringing them into close proximity and inducing the formation of a ternary complex. (b) The PROTAC MoA begins with an E1–ubiquitin-activating enzyme that activates ubiquitin (Ub) in an ATP-dependent manner. This activated Ub is then transferred to an E2–Ub-conjugating enzyme. Subsequently, a PROTAC simultaneously binds to the POI and an E3 ubiquitin ligase, bringing them into close proximity. This facilitates the transfer of Ub from the E2 enzyme to the POI, catalyzed by the E3 ligase. The polyubiquitinated POI is then recognized and degraded by the proteasome into smaller peptides, and the PROTAC is released back into the cellular environment where it can be reused, initiating the process again with another instance of the same POI. (c) Visual representations of dBET6 and its respective ternary complex: left – a 2D skeletal formula of the PROTAC molecule dBET6; middle – a close-up of the dBET6 degrader's three-dimensional (3D) structure in complex with CRBN and BRD4 (PDBID:6BOY), emphasizing the importance of the PROTAC's spatial orientation in forming a good ternary complex; and right – a space filling model for the same complex, involving BRD4, CRBN, DNA damage-binding protein 1 (DDB1), and dBET6. Color key: BRD4 (green), CRBN (cyan), and DDB1 (dark blue).

observing the interaction between Protac-1 and MetAP-2. Results showed that MetAP-2 selectively binds the angiogenesis inhibitor ovalicin moiety of Protac-1, and that Protac-1 can mediate the ubiquitination of MetAP-2 by β-TrCP, leading to its degradation.[12]

In 2003, the same group synthesized a PROTAC using estradiol, a form of estrogen, as part of its structure.[14] This PROTAC was designed to target and promote the destruction of the estrogen receptor alpha (ERα), which, when activated by estrogen, can promote the growth of some breast cancers.[15] It has been shown that the estradiol-based PROTAC could effectively enforce the ubiquitination and subsequent degradation of the α isoform of ER in vitro.[14] Similarly, they created a PROTAC that incorporates dihydrotestosterone (DHT) to target and degrade the androgen receptor (AR). When activated by androgens like DHT, the AR can stimulate the growth of prostate cancer cells.[16] The DHT-based PROTAC has shown efficacy in promoting the rapid ubiquitination and proteasome-dependent degradation of AR in cellular tests.[14] These PROTACs served as proof that they are a promising modality for selectively degrading key proteins involved in cancer, opening up potential treatment benefits by TPD in hormone-responsive cancers.[17,18]

While first-generation PROTACs were capable of degrading target proteins, they suffered from poor cell permeability and chemical stability stemming from their high molecular weight.[19] They generally exhibited low potency using micromolar concentrations, which is less desirable than the nanomolar concentrations used for more potent drugs, indicating that higher doses are required to exhibit efficacy.[1] Notably, early PROTACs were peptide-based and commonly used β-TrCP or Von Hippel-Lindau (VHL) as E3 ligases. One significant drawback of peptide-based therapeutics is their high molecular weight, which affects their ability to cross cell membranes. This poor permeability is a critical limitation because it means that even if a PROTAC is theoretically effective, its inability to enter cells renders it ineffective in practice.[1] These limitations promoted the need to develop second-generation PROTACs, motivating a transition from peptide-based to small-molecule PROTACs. The use of small molecules expanded the range of potentially targetable proteins by taking advantage of a more extensive array of E3 ligases beyond β-TrCP and VHL, such as mouse double minute 2 homologue (MDM2), inhibitors of apoptosis proteins (IAPs), and cereblon (CRBN).[19] In 2008, the Crews lab developed the first small-molecule PROTAC that could degrade a target protein within cells, in this case, targeting AR.[20] This PROTAC was composed of nutlin-3A, a ligand for MDM2, and a non-steroidal androgen receptor ligand (SARM) for AR, connected by a polyethylene glycol (PEG) linker.[20] The SARM-nutlin PROTAC induced the degradation of AR in a proteasome-dependent manner with enhanced cell penetration in vitro.

Since the first PROTAC was reported in the literature, the field of PROTACs has experienced remarkable growth[3,21] and has led to the design of compounds with improved drug-like properties, demonstrating effectiveness both in vitro and in vivo.[22–25] In 2013, the first in vivo success of PROTACs occurred with the development of phosphoPROTACs. PhosphoPROTACs are a particular

form of PROTACs that exploit phosphorylation-dependent binding interactions.[26] This modification was made to improve the selective targeting of proteins involved in signaling pathways. These compounds were able to inhibit tumor growth in mouse models. This was a major breakthrough, as it proved that PROTACs could be used not only in cell-based assays but also in living organisms to exert therapeutic effects.

In 2019, the first PROTACs to enter clinical trials were ARV-110 (ref. 27 and 28) and ARV-471,[29] which target AR and ER, respectively. ARV-110 was tested in a heavily pre-treated population with metastatic castration-resistant prostate cancer (mCRPC). Results from a phase I trial showed that ARV-110 could reduce the levels of AR in cancer cells by at least 95%, which is a significant reduction that hampers the cancer cell's ability to grow and survive. Notably, its effectiveness in ENZ-resistant models offers a potential treatment option for patients who no longer respond to ENZ, addressing a critical gap in prostate cancer therapy. ARV-110 advanced to phase II clinical trials in 2020 based on initial phase I data that demonstrated the drug's good oral availability, safety, and tolerability in patients.[30] On the other hand, ARV-471 is designed for oral administration in patients with hormone receptor-positive (HR+) and HER2-negative metastatic breast cancer. In a phase I clinical study involving breast cancer patients who had undergone multiple prior treatments, ARV-471 significantly reduced the expression level of ER in tumor tissues of patients. It was also reported that ARV-471 is well tolerated across all tested doses (30–700 mg), and maintained a high level of ER degradation (89%).[10] ARV-471 advanced to phase III clinical trials for breast cancer in 2024.

Following the lead of ARV-110, PROTAC technology has advanced significantly. Approximately 29 PROTAC drugs have entered clinical trials, which marks their successful translation into the clinic.[26] Notably, this rapid expansion includes treatments targeting previously undruggable proteins, such as transcription factors and RNA-binding proteins. Additionally, these trials primarily focus on oncology, targeting cancers with poor prognoses, including metastatic prostate cancer, breast cancer, and solid tumors.

In some of the latest generations of PROTACs, additional elements have been introduced to give another dimension of control over PROTAC activity.[31,32] These classes of controllable PROTACs aim to address off-tissue effects by controlling PROTAC action in a spatiotemporal manner.[32] Some are designed to be activated or deactivated by specific wavelengths of light, allowing for controlled degradation processes in target cells, with potentially reduced side effects and enhanced therapeutic index. These PROTACs include phospho-dependent PROTACs that degrade targets with activated kinase-signaling clues, and light-controllable PROTACs that use light as an external clue to trigger target degradation. Notable light-controllable PROTACs, also commonly referred to as PHOTACs, include photo-caged and photo-switchable PROTACs.[33] Photo-caged PROTACs are designed to be inactive in their initial form and activated by light exposure, which removes the photo-cage group and enables the degradation of the POI. Photo-switchable PROTACs, on the other hand, are designed to reversibly control the

degradation process *via* the incorporation of photoswitchable groups such as azobenzene, which can switch between active and inactive states under different wavelengths of light. In-cell click-formed proteolysis-targeting chimeras (CLIPTACs) share similar ambitions to PHOTACs and have been used to degrade two key oncology targets successfully.[34] The reader is referred to these excellent reviews for a more detailed analysis of milestones in PROTAC development.[1,19,26,32,35]

1.2 The ubiquitin–proteasome pathway

PROTACs work by ingeniously harnessing the ubiquitin–proteasome system (UPS), an important cellular pathway, which naturally degrades over 80% of cellular proteins to regulate protein levels and turnover.[36] The UPS selectively targets misfolded and damaged proteins within the cell for degradation, maintaining proper protein homeostasis.[37] However, if this system falters such that old, damaged, or surplus proteins are not promptly degraded, they can form aggregates resistant to degradation. These aggregates can interfere with cellular functions and are the hallmark of several neurodegenerative diseases such as Alzheimer's, Parkinson's, and Huntington's diseases.[38–41] Moreover, a recent multi-omics study of >9000 human tumors and 33 cancer types found that >19% of all cancer driver genes impact UPS function.[42] The UPS operates by tagging target proteins for degradation through the attachment of ubiquitin (Ub) protein chains. These Ub tags mark the protein for degradation by the proteasome, a protein complex responsible for protein degradation *via* proteolysis.[43] The process of Ub conjugation involves an enzyme cascade, starting with E1 activating enzymes, proceeding to E2 conjugating enzymes, and culminating with E3 ligases.[44] Initially, Ub is activated by an E1 enzyme, a reaction that requires adenosine triphosphate and results in an E1–Ub conjugate. The activated Ub is then transferred to an E2 enzyme through a transthioesterification reaction, forming an E2–Ub complex. The most crucial step is mediated by the E3 ubiquitin ligase, which confers specificity to the ubiquitination process.[45]

E3 ubiquitin ligases are categorized into two main types based on their mechanism of action (MoA) for transferring Ub's to their target proteins: HECT-domain and RING-type E3 ligases. HECT-domain E3 ligases first form a thioester bond with Ub. This means that Ub is temporarily attached to the E3 ligase itself. Subsequently, the E3 ligase transfers the Ub from itself directly onto the substrate protein that is to be tagged for degradation.[46] Unlike HECT-domain ligases, RING E3 ligases do not form a direct bond with Ub. Instead, they facilitate the transfer of Ub directly from an E2 enzyme (which is conjugated with Ub) to the substrate protein.[47] In essence, RING-type E3 ligases act as mediators that bring the E2–Ub conjugate close to the substrate, enabling the direct transfer of Ub. This ubiquitination cycle repeats, leading to the transfer of multiple Ub's and the polyubiquitination of the substrate. Once a protein is polyubiquitinated, it is tagged for degradation. The proteasome recognizes the tagged protein, binds to it, unfolds it, and breaks it down into smaller peptides.[48] The PROTAC is then recycled for additional ubiquitination rounds of additional substrates.[49]

1.3 PROTAC structure design

The UPS's sophisticated mechanism forms the basis for PROTAC structure design. PROTACs are bifunctional molecules, designed to harness the UPS for TPD. Each PROTAC comprises three key components: a ligand that binds to the POI, also frequently referred to as the "warhead"; a ligand that recruits an E3 ubiquitin ligase; and an organic linker that connects these two ligands. This dual engagement enables PROTACs to bring the POI and E3 ligase into proximity, forming a ternary complex that facilitates the transfer of Ub from the E3 ligase to the target protein.[11,50,51] Notably, the linker connecting these moieties is not merely a passive scaffold; it has been shown to play a vital role in determining the overall efficacy and specificity of the PROTAC molecule.[52–56] It ensures efficient ubiquitination by correctly positioning the two ligands. This can be achieved by carefully designing the linker length and composition to maintain the required distance, flexibility (or rigidity), and spatial orientation between the POI and the E3 ligase.[10,57,58] Additionally, linker modification can affect properties like hydrogen bond donors (HBDs) and acceptors (HBAs), lipophilicity, molecular weight, rotatable bonds, and polar surface area, which are all critical factors in absorption, distribution, metabolism, and excretion (ADME).[59] Improving these properties can make PROTACs function better as drugs. A study on BET degraders provides a notable example,[60] where researchers replaced an amide bond in BET degraders MZ1 and ARV-771 with an ester group. This change removed one HBD and increased the lipophilicity for each molecule, leading to increased cell membrane permeability. This was reflected in improvements in parallel artificial membrane permeability assay (PAMPA) and $A \log P$ values, among other measurements. Besides being able to enter cells more easily, each molecule's degradation activity also improved after this otherwise "small" change to the linker.

1.4 Advantages of PROTACs over small-molecule drugs

While small-molecule drugs (SMDs) have demonstrated success in treating various diseases,[61–63] PROTACs can target and degrade proteins regardless of their function by hijacking the cell's natural disposal mechanism – the ubiquitin–proteasome pathway.[64] This approach circumvents the inherent limitations of standard SMDs, which must occupy specific binding sites on target proteins.[3] Notably, ~85% of the human proteome has been deemed undruggable by SMDs;[64–67] these proteins typically lack well-defined binding pockets, offering limited opportunities for ligand interaction or binding.[3,68] For instance, 63% of the known 600 cancer-related proteins are classified as undruggable, including transcription factors, scaffold proteins, and membrane-bound proteins.[69] While SMDs are constrained to a limited pool of proteins that can be effectively targeted, PROTACs don't necessarily require binding to a specific well-defined pocket on a POI to trigger degradation; in theory, they can bind to any reachable region on a POI's surface that facilitates induced proximity between the POI and an E3 ligase, even in the case of low binding affinities with the POI.[1,70] This flexibility in target engagement expands the scope of the proteome

that can be drugged with this modality, and is a mechanism that can be particularly pertinent in cancer treatments, where target proteins often develop resistance to SMDs through mechanisms like genetic mutations, overexpression, or altered signaling pathways. PROTACs offer a potential alternative capable of overcoming these resistance barriers.[71–75]

Furthermore, a key advantage of PROTACs lies in their substoichiometric catalytic activity, which operates on an event-driven basis.[76] This means that PROTACs do not need to fully occupy their target proteins to be effective, in contrast to traditional inhibitors that function in an occupancy-driven manner.[77] In SMDs, the effectiveness of the drug is often dependent on stoichiometrically occupying the target binding site.[78] This means that a significant portion of a target protein must be bound by an inhibitor molecule for the desired therapeutic effect to be observed. This often requires relatively high concentrations of the drug to achieve sufficient occupancy, since the effects are proportional to the extent of binding.[78] PROTACs operate differently: they bind transiently to their targets and, after facilitating ubiquitination, dissociate. This allows them to cycle through multiple rounds of activity, repeatedly initiating the degradation of additional instances of the same POI.[1] In contrast to SMDs that act in a dose-dependent manner, this catalytic feature allows PROTACs to achieve potent effects at possibly lower doses, offering potential advantages in terms of efficacy, safety, negative side effects, and off-target effects.[79]

One final advantageous characteristic of PROTACs worth mentioning is that they are able to selectively target and induce the degradation of specific protein isoforms. These are distinct forms of the same protein arising from a single gene. The ability to selectively target them is significant because it implies that PROTACs can be used to differentiate between closely related forms of a protein and target only the isoform(s) associated with a disease without affecting others that may have essential functions in normal cellular processes.[6,69,76]

1.5 PROTACs as probes for target identification

PROTACs have been used as tools not only for therapeutic applications but also for target deconvolution, known as PROTAC probes.[26] Target deconvolution is the process of identifying the exact biological target that a drug or compound interacts with to produce its therapeutic effect. In drug discovery, researchers often do not know exactly which protein(s) a particular compound is affecting, especially with natural products or new drug candidates. When PROTACs are used as tools for target deconvolution, the goal is to identify or confirm the biological targets that a drug or compound is acting upon. In this case, PROTACs are used as research tools rather than drugs. Yan et al.[26] describe how PROTAC technology has been used in target deconvolution, particularly in the context of natural products like artemisinin and lathyrane diterpenoids. These natural products have therapeutic potential, but their exact targets were not fully understood. Notably, lathyrane diterpenoids were found to have anti-inflammatory properties, but their precise molecular targets were not initially known. To

uncover the biological target, researchers used a PROTAC probe based on the lathyrane diterpenoid ZCY-001. This PROTAC was designed to degrade proteins that the compound might interact with. Using this approach, the protein MAFF was identified as the target responsible for the anti-inflammatory effects. MAFF is a transcription factor involved in regulating stress response pathways, and its degradation helped reduce inflammation. Researchers were able to identify the proteins responsible for the therapeutic effects (e.g., MAFF for lathyrane diterpenoids and PCLAF for artemisinin derivatives) by using PROTAC probes designed to degrade the proteins they interact with.

These attractive characteristics make PROTACs a prime focus of drug design endeavors. To maximize the potential of this innovative class of compounds, researchers are increasingly turning to data-driven approaches for design strategies. Machine learning (ML) has thus demonstrably advanced drug discovery and development by enhancing target identification, small-molecule design, predictive biomarker discovery, and the prediction of clinical trial success.[80,81] ML methods can help researchers analyze large amounts of data to identify potential drug targets, optimize compound properties, and predict how patients will respond to treatments. This makes drug development more efficient and increases the likelihood of success in vitro.[80,81] Given the complexity of designing PROTACs due to the large chemical space they span and their multivalent nature, leveraging ML will likely be crucial in making the development of this new modality more feasible. Despite numerous reviews on PROTACs, there is a notable gap in the literature: an in-depth review that delves into the use of ML for PROTAC design is still lacking. In this comprehensive literature review, we explore the impact of ML on de novo PROTAC design to date. First, we delve into the distinct characteristics of PROTAC linker design, underscoring the features required to create effective bifunctional molecules capable of TPD. We then examine how ML in the context of fragment-based drug discovery (FBDD; Fig. 2a), honed for small-molecule drug discovery, is paving the way for PROTAC linker design. Our review provides a critical assessment of the obstacles inherent in the application of these methods to PROTAC development. This assessment seeks to shed light on the pressing need for specialized algorithms, enhanced data quality, and the adaptation of ML models to address the multifaceted nature of PROTAC engineering. Moreover, we review existing ML works that have been tailored to PROTAC design, highlighting pioneering efforts as well as the limitations associated with these existing approaches. We also offer perspectives on potential avenues for future ML research in this field.

2 Machine learning in PROTAC linker design

2.1 The peculiarities of PROTAC linker design

The role the linker plays in PROTAC function is both unique and complex, offering the broadest scope in terms of where structural modifications can be made when designing a PROTAC (Fig. 2b).[52–56,82] Unlike the linker, the structures of the
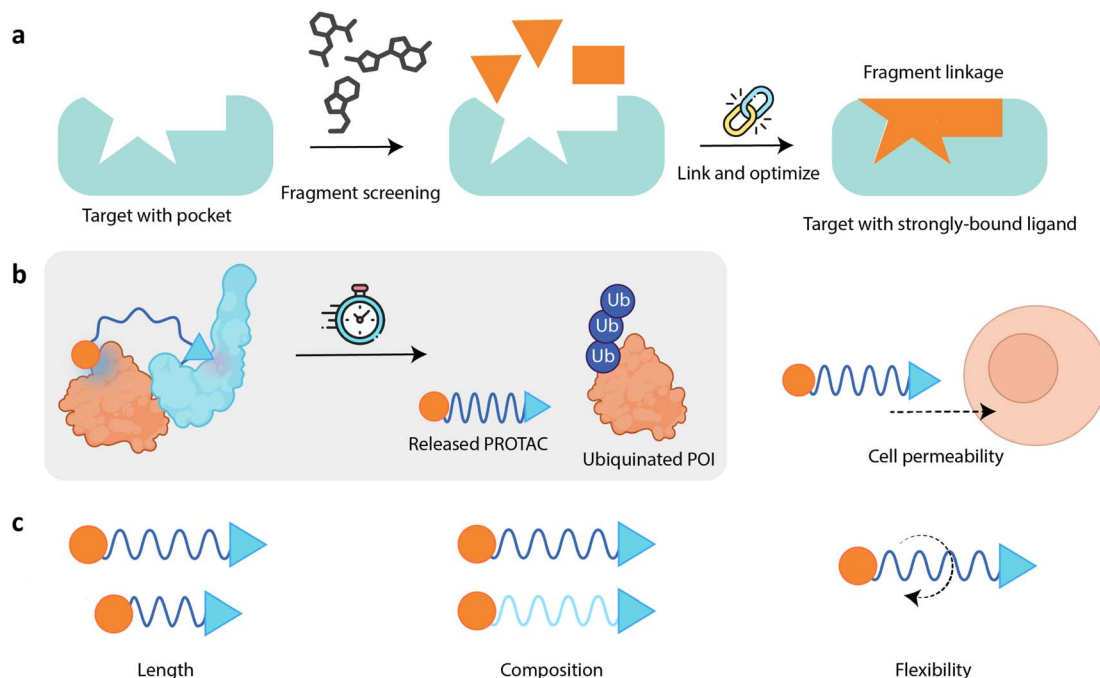
Fig. 2 (a) An overview of fragment-based drug design (FBDD). The initial step involves fragment screening to identify potential fragments that can bind to the pocket of the target protein. These fragments are then linked and optimized to improve their binding properties. The result is a strongly-bound ligand that fits precisely within the target protein's pocket. (b) Left – The linker in a PROTAC isn't just a passive bridge. It's an important component that enhances the interaction dynamics between the POI and the E3. Right – The linker also contributes to the PROTAC's overall PK profile, including cell permeability. Center – Because its MoA relies on transient ternary complex formation, the PROTAC is eventually released, meaning it is catalytic and can go on to be reused for other processes inside the cell. (c) The large and multivalent nature of PROTACs means they require a more complex design approach than FBDD methods developed for small molecules. The linker must be long and/or flexible enough to allow the warhead and E3 ligase ligand to adopt the necessary conformations for effective ternary complex formation, but not too flexible that the PROTAC cannot maintain the correct spatial orientation of the warhead and E3 ligase ligand. The linker may also need to incorporate specific chemical groups to enhance the overall potency of the PROTAC.

ligands that bind the POI and E3 ligase are generally more restricted. These restrictions stem from the need to maintain specific structure–activity relationships (SAR) and effective target binding, limiting the options for modifying the warhead and E3 ligase ligand according to their functional requirements.[82] Consequently, the linker becomes a primary focus for design optimization in PROTACs. Modifications to the linker, such as altering its length, tuning its flexibility or rigidity, and incorporating different chemical groups, can influence the pharmacokinetic (PK) and pharmacodynamic (PD) profiles of PROTACs, as well as their degradation activity and overall efficacy.[82–84] Notably, the geometry (conformation) of the ternary complex is heavily influenced by the nature of the linker.[50,85] Evidently, the linker not only dictates the spatial arrangement necessary for successful TPD in a given system but also the efficiency with which the PROTAC can facilitate the degradation of the POI.[84,86] This is often quantified by metrics like the DC_{50}, the concentration at which half-maximal degradation is observed, and D_{max}, the maximum level of degradation achievable. This implies that PROTAC linker design requires a multifaceted approach that balances several key properties to ensure the successful degradation of a POI while maintaining a desirable drug-like profile. In this section, we present example case studies that highlight the unique aspects of linker design in PROTACs.

2.1.1 Linker length. The conformation of the ternary complex is heavily dependent on the linker length within PROTACs. A long linker can lead to the formation of no functional complex since the ubiquitination of the POI might not occur. A short linker might result in what's called a binary complex where the PROTAC is only effectively linked to either the POI or the E3 ligase, but not both.[87] Similar results were observed by Qin *et al.*,[88] where the potency of PROTACs targeting BET proteins to inhibit cell growth was highly dependent on the linker length. This was demonstrated *via* a series of five PROTACs with progressively longer linkers. Notably, the authors observed an optimal linker length; further extending the linker beyond this optimal length did not enhance the potency. This shows that a very long linker might not provide additional benefits, and could inadvertently introduce steric constraints leading to decreased binding affinity. The most effective PROTAC from this study, QCA570, was tested in xenograft mouse models of leukemia, where it induced complete and durable tumor regression at low picomolar concentrations.

Furthermore, the length of the linker affects the range of spatial configurations accessible to potential ternary complexes during formation, restricting which protein interfaces are accessible for interaction. Smith *et al.*[55] demonstrated how

differences in linker lengths and attachment points enable selective degradation of closely related kinase isoforms using PROTACs. The study developed isoform-selective PROTACs for the p38 mitogen-activated protein kinase (MAPK) family using the same warhead and E3 ligase but varying the linker features (linker attachment points and lengths). Two different linker attachment points (an amide and phenyl series) and varying linker lengths (10, 11, 12, and 13 atoms) were used to create distinct PROTACs that differentially recruit VHL. This selective recruitment controls the degradation of either the p38α or p38δ isoforms. For instance, PROTACs with 12- and 13-atom linkers in the amide series became highly selective for p38α degradation, showing much higher degradation efficacy compared to degraders with shorter linkers, which were also less selective. Conversely, a 10-atom linker in the phenyl series led to selective degradation of p38δ, with very minimal impact on other isoforms. This selective degradation ability is attributed to how variations in linker lengths and attachment points influence the formation of the ternary complex. By fine-tuning the linkers, PROTACs can achieve selective degradation profiles – in this particular study, shorter linkers may bring the E3 ligase into a position that is optimal for ubiquitinating p38δ but not p38α.

● Linker length is an important factor in determining the spatial configuration necessary for effective ternary complex formation. Adequate length ensures optimal potency. Both too-long and too-short linkers can negatively impact the potency of PROTACs.

● Small changes in linker length can shift the degradation selectivity between closely related protein isoforms.

2.1.2 Linker composition. The linker composition has an impact on the physicochemical properties of PROTACs, such as solubility and membrane permeability, among other factors.[59] For instance, the substitution of amide bonds with ester bonds in BET degraders MZ1 and ARV-771 results in improvements to their permeability and cellular activity.[60] The ester-linked versions, OMZ1 and OARV-771, demonstrated a 10- and 1.5-fold increase in PAMPA permeability, respectively. This improved permeability contributes to their enhanced ability to degrade target proteins, with OMZ1 achieving a 1.5- to 2-fold increase in degradation potency, while OARV-771 achieved a 5.5-fold increase. Despite concerns about the stability of esters compared to amides, both OMZ1 and OARV-771 maintained stability in plasma, indicating that their increased permeability does not compromise their overall stability. Additionally, an optimal lipophilicity range ($A \log P$ between 3 and 5) was established for these ester-linked PROTACs, balancing effective membrane crossing with adequate aqueous solubility and minimal efflux. Klein *et al.*[60] conclude that amide-to-ester substitution can benefit the optimization of PROTACs, and potentially other compounds, falling beyond the Rule of 5.

The composition of the linker can also improve the PK properties of PROTACs, such as metabolic stability, and biodistribution.[82] These properties influence how the drug is adsorbed, distributed, and eventually metabolized inside the body. However, the majority of linkers in PROTACs have been based on a limited set of chemical motifs, with PEG and alkyl chains being the most common. Approximately 55% of linkers utilize PEG, while about 30% use alkyl chains of various lengths.[57] These motifs are favored due to their versatility, ease of synthesis, and ability to modulate the solubility and permeability of PROTAC molecules. Around 65% of published PROTAC structures incorporate both alkyl and PEG segments within their linkers. This combination aims to leverage the beneficial properties of both motifs, such as the flexibility and hydrophilicity provided by PEG, and the structural simplicity and modifiability of alkyl chains. A further 15% of linkers involve modifications to the basic glycol units in PEG, such as adding methylene groups.[57] Such modifications are typically done to explore different chain lengths and thus influence the potential structural configurations accessible to PROTACs.

● Amide-to-ester substitution can benefit the optimization of PROTACs, and potentially other compounds, falling beyond the Rule of Five.

● Modifications in PROTAC linker composition, such as altering chemical groups and combining different motifs, directly influence the physicochemical properties of PROTACs.

2.1.3 Linker flexibility. The flexibility of the linker can allow a PROTAC to more easily adapt to different spatial configurations, though too much flexibility may also lead to less predictable interactions.[54,89] A certain degree of rigidity can thus confer stability during ternary complex formation, leading to more consistent degradation activity. A 2018 study on PROTACs for ANK-binding kinase 1 (TBK1) degradation underscores the role linker flexibility plays in PROTAC activity.[89] Flexibility was imparted *via* long alkyl and ether chains to achieve potent compounds due to the inherent flexibility of these chemical units. A systematic exploration of linker length was thus conducted, and the degradation activity of each PROTAC variant was measured, focusing on those with sub-μM potency to identify effective linker lengths and compositions. The study found that PROTACs with linkers shorter than 12 atoms showed no appreciable degradation activity. In contrast, longer linkers, despite their higher polar surface area and challenges in cellular penetration, were generally well-tolerated and effective in degrading TBK1. The very flexible nature of the linker allowed long linkers to orient the ligands in a way that facilitated the association of TBK1 and VHL into suitable ternary complexes.

Another study describes how the ability of PROTACs to induce selective protein degradation is enhanced by the plastic nature of the binding interactions between CRBN and BRD4 bromodomains.[54] Plasticity here means that the proteins can adopt multiple conformations at the binding interface depending on the linker length, composition, and linkage position. It was shown that different linkers can promote different binding conformations between the CRBN and BRD4. This plasticity allows the PROTAC to effectively bring the proteins into proximity in orientations that are conducive to ubiquitination. Using X-ray crystallography and molecular docking, the authors shed light on how different linker configurations lead to distinct low-energy binding conformations between CRBN and BRD4. The varying conformations accessible to PROTACs in this study illustrate how linker-induced flexibility directly impacts biological outcomes.

• The flexibility of a PROTAC's linker can be tuned by adjusting, for instance, the length and chemical composition of the linker.

• Flexibility can allow for conformational adaptability and access to multiple binding orientations.

2.1.4   Bottom line. There are no definitive guidelines that guarantee the design of an effective PROTAC for any given E3 ligase–POI pair.[57] This means that developing a potent degrader generally requires trial and error, with a reliance on empirical metrics to identify effective linker features that establish the optimal SAR. The large combinatorial space makes linker design ideally suited to data-driven approaches, which provide a valuable complement to traditional, labor-intensive experimental methods. Computational models can furthermore simulate how ternary complexes form and behave at a level of detail generally inaccessible in most experiments. Used wisely, computational tools can aid in tricky tasks such as linker optimization without the need to synthesize and test numerous variants experimentally, potentially speeding up their development and reducing the costs associated with the synthesis and empirical testing of unpromising compounds.

One recent study that nicely illustrates these points used a combination of crystallographic data and mathematical modeling to explore the conformational dynamics of protein–protein interactions induced by PROTACs, to understand how these dynamics influence ubiquitination and eventual protein degradation.[90] Interestingly, the authors found that the stability of the ternary complex did not necessarily correlate with increased protein degradation efficiency, suggesting that excessive stability might inhibit degradation efficiency. Notably, the spatial arrangement and kinetic properties of the ternary complex were crucial in this context: effective PROTACs brought lysine residues on the POI close to the active site of the E2 enzyme, facilitated by the E3 ligase within the complex. Lysine residues are the most common sites for ubiquitination in proteins. The authors also confirmed that the kinetics of the ternary complex, especially its dissociation rate, also play a role in determining the degradation efficiency. Salt bridges and the hydrophobicity of the interactions within the ternary complex were found to contribute positively both to the cooperativity and to the half-life of the interaction. These findings suggest prioritizing compounds that can induce the necessary conformational dynamics without overly stabilizing the ternary complex, highlighting how valuable insights can be gained using computational tools.

2.2   PROTAC linker design goes beyond fragment linking

While PROTAC design shares similarities with traditional small-molecule drug design, it is fundamentally distinct, notably in linker optimization (Fig. 2c). For instance, the approach to optimizing PROTAC linkers differs significantly from the concept of "fragment linking" used in fragment-based drug design (FBDD). In essence, FBDD is a strategy used in drug discovery where small, low-complexity molecules, *i.e.*, fragments, are screened for binding to a specific pocket on the target protein. First, a library of small chemical fragments is created. These fragments are typically smaller than traditional

drug-like molecules, with a molecular weight of less than 300 Da. The fragment library is then screened against the target protein to identify fragments that bind to the protein. This can be done using various techniques such as nuclear magnetic resonance (NMR) spectroscopy, X-ray crystallography, surface plasmon resonance (SPR), and thermal shift assays. Fragments that show binding affinity to the target protein are identified as "hits". These hits often bind with low affinity but serve as a starting point for further optimization. Subsequently, the identified fragments are optimized to improve their binding affinity and drug-like properties. This can involve growing the fragment by adding more atoms, merging fragments that bind to adjacent sites, or linking fragments that bind to different parts of the target protein. Finally, the optimized fragments are developed into lead compounds, which have improved pharmacological properties and can be further tested in biological assays and *in vivo* studies.[91,92]

Fragments can be an ideal starting point for drug design, with fragment growing and linking strategies allowing for the optimization of their potency and physicochemical properties. Fragment linking in particular gives the possibility for significant potency gains by ensuring that the linked molecule maintains the interactions of the original fragments, a phenomenon known as super-additivity.[93] However, achieving this is in practice very challenging, as a bad linker can instead lead to the disruption of fragment binding poses.

Despite its success in drug discovery, FBDD may fall short when applied to PROTAC linker design. PROTACs are substantially larger and more complex than the small fragments typically dealt with in FBDD. The linker in a PROTAC must connect two distinct binding moieties, facilitating the formation of a stable ternary complex, and does not simply focus on improving the binding affinity. To reiterate, the linker in a PROTAC must be flexible enough to allow the formation of a ternary complex but rigid enough to maintain the correct spatial arrangement of the ligands. This balance is difficult to tackle using traditional FBDD approaches, which focus on optimizing single-binding interactions rather than complex multi-protein assemblies. The unique challenges posed by the size, complexity, and spatial requirements of PROTACs necessitate more advanced methodologies. While direct application of typical fragment-linking strategies used in FBDD is not generally feasible in PROTAC design, a modular approach can certainly be beneficial. As we show in the next section, researchers are already taking inspiration and lessons learned from FBDD and applying them to PROTAC design.

2.3   FBDD and ML pave the way for PROTAC development

The advent of ML in drug discovery has transformed the way researchers approach the design and optimization of therapeutics. ML algorithms streamline labor-intensive design-make-test-analyze (DMTA) cycles by automating the identification of promising compounds and enabling researchers to avoid synthesizing and testing ineffective compounds, thereby reducing both time and costs.[80,81] Before the widespread use of ML in drug discovery seen today, computational tools for linker

design typically involved searching a database, limiting proposed linkers to those already documented in the database.[93] While such approaches have been successful, ML-based solutions can give models an advantage by allowing them to better explore the possible chemical space. Generative models, in particular, are used to create new molecules from scratch, designing possible structures with enhanced or desirable property profiles based on patterns learned from data.[94,95] However, generative models trained for FBDD are optimized for designing linkers between small fragments. These models might not account for the larger size and higher complexity of PROTAC structures, potentially leading to inaccurate linker design. Presented below is a detailed overview of generative models used in *de novo* linker design within the framework of FBDD.

2.3.1 2D generative models in linker design. SyntaLinker[96] is a method that employs syntactic pattern recognition with molecular string representations *via* deep conditional transformer neural networks.[96] It uses molecular representations based on the simplified molecular-input line-entry system (SMILES), one of the most popular methods for representing 2D molecular graphs as strings. SyntaLinker allows for the automatic linking of molecular fragments by learning from patterns in drug-like compounds in ChEMBL.[97] This enables the generation of new molecular structures starting from pairs of fragments in a conditional way; possible constraints include the shortest-linker bond distance (SLBD), HBDs, number of rotatable bonds, and number of rings. The constraints are incorporated as control codes. For example, "[L_4]" denotes a linker with a bond distance of four, which acts as a guiding prefix in the sequence. ChEMBL data was filtered using Lipinski's "Rule of Five," PAINS substructures, and synthetic accessibility (SAscore) to ensure lead-like and synthesizable molecules were used to construct the training data. The dataset was constructed using matched-molecular pairs (MMPs) to build molecules into fragment molecule triplets (fragment 1, linker, fragment 2). SyntaLinker consists of multiple encoder–decoder stacks where each encoder layer has a multi-head self-attention sub-layer and a feedforward network sub-layer. The model was trained using the prepared dataset, with SLBD and the aforementioned constraints to learn the implicit rules of fragment linking. The generated molecules were evaluated using several metrics, including validity (97.2%), uniqueness (88.1%), recovery (84.7%), and novelty (91.8%).

Link-INVENT[93] is an extension to the existing *de novo* molecular design platform REINVENT; it uses policy-based reinforcement learning (RL) for multi-parameter optimization, and can be applied to both fragment linking and scaffold hopping given a desired property profile. *Via* RL, the Link-INVENT agent learns to generate linkers connecting molecular fragments while satisfying diverse objectives, facilitating the practical application of the model for real-world drug discovery projects. In the original study, Link-INVENT used the drug-like compound SMILES extracted from ChEMBL for training. Lenient criteria were applied to ensure the dataset's effectiveness for PROTAC applications (*e.g.*, larger warheads). ChEMBL compounds were sliced using reaction SMIRKS to create triplets

(linker, warheads, full molecule). Unrealistic data points were removed, and datasets were augmented *via* SMILES randomization for improved generalizability. Link-INVENT is trained based on the conditional probabilities of observing a linker given both molecular subunits, similar to SyntaLinker. The agent is initialized with the same parameters as the prior and is updated *via* RL to generate linkers that increasingly satisfy the desired multi-parameter optimization (MPO) objectives. The scoring function combines various components (physicochemical properties, structural features, predictive models, and binding energy approximations) to evaluate the desirability of generated linkers. Link-INVENT was tested in various experiments, demonstrating its capability to generate linkers that meet specific criteria. Notably, Link-INVENT has also been demonstrated to be effective in PROTAC linker design, successfully optimizing the properties of generated linkers, including effective length, the presence of rings, and flexibility.

Due partly to the surprising effectiveness of 2D representations like SMILES, the majority of molecular generative models used for *de novo* molecular design and FBDD have made limited use of 3D structural information, including SyntaLinker and Link-INVENT. Nevertheless, the PROTAC MoA suggests that incorporating 3D information may come to play an important role in designing PROTAC structures, which lead to favorable ternary complexes. In the next subsection, we cover ML models that seek to incorporate structural information into their molecular design workflows.

2.3.2 3D generative models in linker design. DeLinker[98] is a graph-based deep generative model that incorporates 3D structural information for designing molecules using a multimodal encoder–decoder. Training data was derived from a 250k molecule subset of ZINC, better known as ZINC-250k.[99] The dataset was further processed to create fragment–molecule pairs using standard transformations from MMP analysis. 3D conformers were generated using RDKit, and the lowest-energy conformation was used as the reference structure. Molecules are represented as graphs, where atoms are nodes and bonds are edges. The model builds new molecules iteratively, bond by bond, from a pool of atoms that can be initialized with partial structures. The model uses a gated graph neural network (GGNN) for encoding and a single-layer neural network for edge prediction and labeling. This setup allows the model to use local, global, and 3D structural information for molecule generation, encoding multi-modal molecular information in a low-dimensional latent space. The model is trained under the VAE framework to reconstruct known linkers from fragment pairs and the linked molecule. The latent vector is derived from the embedding of the linked molecule, and the model is regularized to follow a standard normal distribution. The training objective includes a reconstruction loss and a Kullback–Leibler (KL) regularization term, which together ensure that generated molecules are valid and structurally sound. DeLinker can generate novel molecular structures, including those with longer linkers of at least five atoms. It was applied to the design of PROTACs targeting SMARCA2 and SMARCA4 subunits, where it generated linkers maintaining the geometry observed in a ternary complex with VHL. Imrie *et al.*[100] further improves

DeLinker in their more recent DEVELOP framework by introducing a convolutional neural network (CNN), which operates on the 3D structure of the starting fragments. DEVELOP not only improves the proportion of generated molecules with high 3D similarity to the reference molecule, but also recovers 10× more of the original molecules compared to DeLinker.

3DLinker[101] is a conditional generative model for designing molecular linkers using 3D spatial information and is capable of generating linker graphs along with their 3D structures and anchor atoms. This is achieved through an E(3)-equivariant graph VAE, addressing challenges such as the conditional generation of linkers based on two input ligands and the requirement for 3D structural awareness to avoid atom clashes. It predicts both the graph (2D) structure of the linker and its 3D coordinates while ensuring the model's outputs are equivariant with respect to E(3) group symmetries (*i.e.*, rotation, translation, and reflection). The training data was derived from the ZINC database,[102] from which 3D conformers were generated for each molecule using RDKit and the lowest-energy conformation chosen as the reference structure. The final curated dataset contains ∼366k (fragment, linker, coordinate) triplets and was roughly divided into 99.8%/0.1%/0.1% training/validation/testing splits. Using this generous training split, the model outperforms other baselines, including DeLinker and other 2D graph generative models (coupled with ConfVAE[103] for 3D structure generation), in recovering molecular graphs and accurately predicting the 3D coordinates of atoms. Nevertheless, it is unclear if the reported metrics are for the training, validation, or test set. While 3DLinker demonstrates improved performance in generating 3D molecular structures with accurate geometry, precise connection of molecular fragments, and higher recovery rates, the authors observed that this comes with the trade-off of lower uniqueness and novelty in sampled molecules compared to the benchmarked approaches.

DiffLinker[104] is an E(3)-equivariant 3D-conditional diffusion model for the design of molecular linkers. This approach uniquely generates molecular linkers for a set of input fragments represented as 3D atomic point clouds, overcoming the limitations of previous methods by not being restricted to linking pairs of fragments. DiffLinker automatically determines the number of atoms in the linker and its attachment points to the input fragments. As the previous approaches, DiffLinker was trained and evaluated on a dataset derived from ZINC-250k, but the authors also took things a step further by benchmarking on two additional datasets: one derived from CASF-2016, and another derived from GEOM.[105] The molecules derived from GEOM can be decomposed into three or more fragments with one or two linkers connecting them, creating a more challenging benchmark that better approximates real-world usage. DiffLinker demonstrates an ability to generate diverse and synthetically accessible molecules with minimal clashes, especially when conditioned on target protein pockets. It represents a significant advancement in FBDD, providing a powerful tool for the generation of chemically relevant molecules in a flexible and efficient manner. Nevertheless, the authors did not apply their fragment-linking approach to PROTAC design.

Building upon the success of SyntaLinker, DRlinker[106] is a similar approach that incorporates RL, and, indirectly, 3D information, for the generation of linkers with specific 2D and 3D attributes. It was trained and evaluated for FBDD on datasets derived not only from ChEMBL, but also from CASF-2016.[107] On tasks like optimizing bioactivity, it achieves a 91.0% and 93.9% success rate in generating compounds with desired linker length and $\log P$, respectively. Despite being based on 2D SMILES representation, DRlinker can also perform scaffold-hopping in a way that generates molecules with high 3D similarity but low 2D similarity to lead inhibitors. Two years later, the same team followed up with another model for FBDD, which aims to better incorporate 3D information. GRELinker[108] combines a gated-graph neural network (GGNN[109]) with RL and curriculum learning (CL) to design linkers with desirable property profiles. Its architecture is very similar to that of GraphINVENT.[110] It outperforms DRlinker in tasks such as controlling $\log P$, optimizing synthesizability and bioactivity, and generating molecules with high 3D similarity but low 2D similarity to lead compounds. It has also been evaluated in scenarios representative of real-world use-cases, where the aim is to optimize for molecular affinity using docking scores. The authors found that the use of CL improved its efficiency in generating complex linkers.

Despite the successes of the aforementioned works in FBDD, and, in particular, of DeLinker and Link-INVENT in PROTAC linker design, the methods reviewed above all face a key limitation – they were all trained and optimized on small-molecule binders rather than on an actual PROTAC dataset. Although careful filtering was done to make the datasets more generalizable beyond small-molecule binders, due to the fact that warheads can be much larger than the typical fragments used in FBDD, we argue that the training of these models may not fully capture the unique features and complexities of larger, multivalent molecules like PROTACs, nor their unique chemistry. As previously discussed, PROTACs not only have larger sizes but also exhibit different biophysical and chemical properties compared to the small molecules typically found in drug discovery databases (Fig. 3). This training limitation can affect the applicability of these methods for designing effective PROTAC linkers, as the chemical space and design strategies for PROTACs diverge significantly from those of small molecules. This underscores the necessity for specialized tools for PROTAC linker design that can accommodate their unique size, complexity, and 3D structural requirements. The next section reviews ML models specifically tailored for PROTAC design.

2.4 Previous work in PROTAC linker design

In this section, we focus on prior ML studies that have been designed to advance the optimization of PROTAC linkers. By highlighting how computational models have been intentionally developed to refine PROTAC linker design, we hope to illustrate how the application of these technologies has evolved from traditional drug design to TPD. As ML techniques continue to evolve, they are expected to play an increasingly central role in PROTAC development. Notable ML-driven works
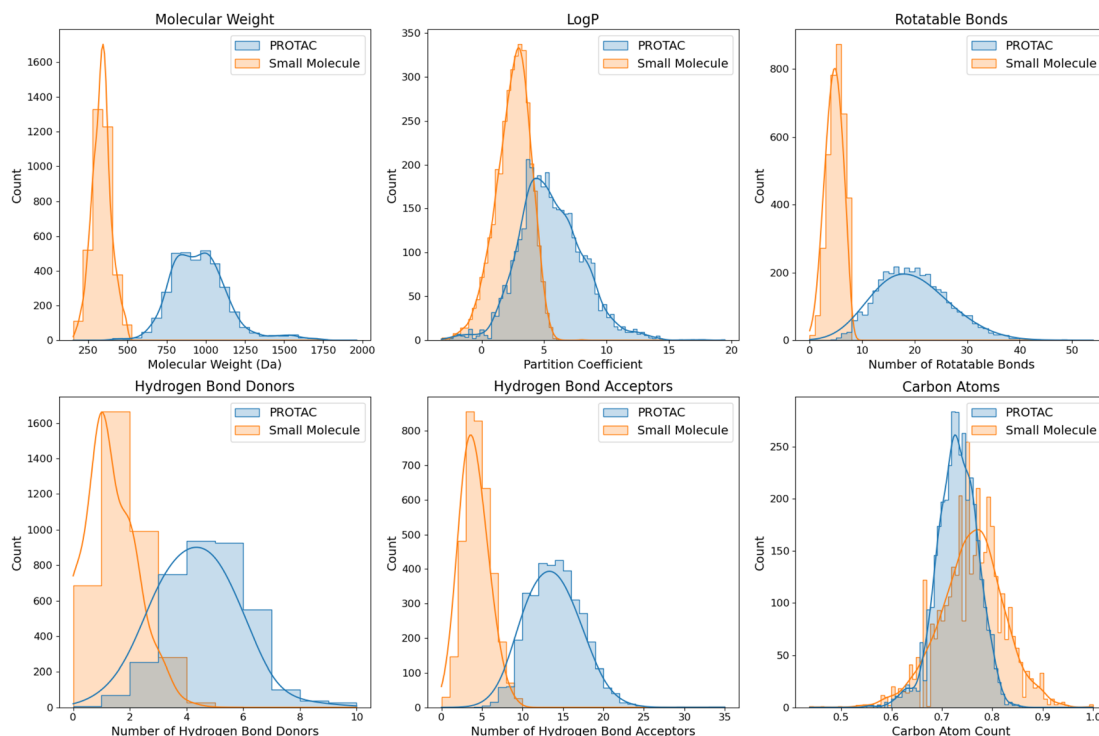
Fig. 3 The distributions of various molecular descriptors in PROTACs *versus* small molecules. PROTACs were downloaded from PROTAC-DB and PROTACpedia, while small molecules were randomly sampled from ZINC-250k,[102] a popular database used in drug discovery containing commercially-available compounds for virtual screening (*e.g.*, drug-like compounds). This comparative analysis of their chemical and physical properties highlights the differences between both classes of molecules. The descriptors include molecular weight, partition coefficient (Log *P*), number of rotatable bonds, number of hydrogen bond donors (HBDs) and acceptors (HBAs), and normalized atom counts for carbon.

such as AIMLinker[111] and PROTAC-RL[112] shed light on the complex nature of PROTAC linker design, though they are not the only recent works to tackle this problem.

PROTAC-RL is a deep generative model combining an augmented transformer architecture with memory-assisted RL capable of generating PROTACs with favorable PK properties, including solubility, stability, and bioavailability.[112] Notably, the authors experimentally validated their model by testing the synthetic feasibility of six of their designs. To address the challenge of limited training data, the model was pre-trained using a large dataset of PROTAC-like structures, termed quasi-PROTACs, followed by fine-tuning on actual PROTAC data. Given a pair of E3 ligand and warhead SMILES, the model generates optimized linkers, which aim to optimize the PK attributes of the returned PROTACs. PROTAC-RL achieved a recovery rate of 43.0%, much higher than the recovery rates of the baseline models, DeLinker and SyntaLinker, even after these were retrained using the PROTAC training datasets. After retraining, Delinker and SyntaLinker achieved recovery rates of 4.8% and 10.4%, respectively. This stark contrast in recovery rates between PROTAC-RL and the benchmarked models after retraining further strengthens the argument that models designed and trained for small molecular fragments cannot adequately capture the unique aspects of PROTACs, as the design strategies and principles for these two classes of molecules are fundamentally different. Because the RL component

allows for the conditional generation of PROTACs with specific properties, such as a desired protein target, the authors applied PROTAC-RL to the design of BRD4-targeting PROTACs. To this end, they generated 5k compounds, which they then filtered through a combination of ML classifiers and molecular simulations to identify candidates with favorable PK properties and synthetic accessibility. Of the six candidate PROTACs, which were synthesized and experimentally tested, three showed inhibitory activity against BRD4 in cell-based assays. One lead candidate demonstrated high anti-proliferative potency and a favorable PK profile in mice.

AIMLinker is a GGNN[109] model for autoregressive PROTAC linker generation at the atomic/bond level.[111] Like GRELinker, it seeks to improve upon previously-developed graph-based deep generative models like DeLinker, CGVAE,[113] and Graph-INVENT[110] *via* the incorporation of 3D information. AIMLinker was trained on a dataset combining molecules from ZINC and PROTAC-DB.[114] The training focused on predicting viable 2D linker structures from fragment–molecule pairs. Generated molecules were then validated *via* molecular docking and simulations to verify binding to the target proteins *via* binding affinity and conformational predictions. AIMLinker was used to successfully generate a diverse library of novel PROTACs. The model demonstrated superiority over other fragment-linking methods (DeLinker and DiffLinker) in generating molecules with favorable PK properties and high binding affinities, with

a few designed PROTACs even outperforming the reference compound dBET6 in binding affinity and structural alignment (Fig. 1c). Despite a promising performance in PROTAC linker design, AIMLinker does have two current limitations, namely the focus on a single PROTAC target (BRD4) and the reliance on docking predictions, which are known to be inaccurate.[115]

Finally, ShapeLinker[116] is a model based on Link-INVENT, but with an important shape alignment contribution to the scoring function, and less significant but still important contributions from the ratio of rotatable bonds and the linker length ratio. The authors train on PROTAC-DB[114] data, as well as on ten well-known ternary complexes from the Protein Data Bank (PDB): 5T35, 7ZNT, 6BN7, 6BOY, 6HAY, 6HAX, 7S4E, 7JTP, 7Q2J, and 7JTO. All of these complexes have binding PROTACs that were optimized in individual structure-based drug studies and cover a diverse range of PROTAC (and linker) "shapes". These are included in the training of the shape alignment model. Nevertheless, it is not clear whether these additions indeed improve the performance of ShapeLinker over that of the base Link-INVENT. The results suggest that perhaps larger changes to the model architecture are required for step-changes in performance.

3 Machine learning in *de novo* PROTAC design: going beyond linker optimization

3.1 Comprehensive PROTAC design strategies

The design of a PROTAC involves more than the design of the linker; it also includes the optimization of the ligands that bind to the target protein and the E3 ligase. The challenge lies in designing a molecule where these components work in harmony to achieve ubiquitination and subsequent degradation of the target protein. The design process begins with the selection of a ligand, if known, that can selectively bind to the POI. This step is crucial because the efficacy of a PROTAC largely depends on the ability of this warhead to recognize and attach to the intended POI.[18] When numerous known binders exist for a protein, the binding affinity, physicochemical properties, and synthetic feasibility of these binders are crucial factors to consider. Approved drugs, drug candidates in clinical trials, or highly active inhibitors may be preferred starting points due to their optimized PK and PD properties, for example.[117]

Equally important is the choice of an E3 ligase and a corresponding ligand. The selection of the E3 ligase often involves considering a range of factors, such as the ubiquitination efficiency, the specificity of the ligase, and its expression levels within the relevant cells or tissues. For instance, if a PROTAC is being developed for cancer therapy, the chosen E3 ligase should be highly expressed in cancer cells and less so in healthy cells to minimize off-target effects.[118] Furthermore, different E3 ligases can induce varying degrees of degradation even with the same POI ligands and linkers. The selection of an effective E3 ligase and ligand is thus a critical aspect of PROTAC design, and structural knowledge of the E3 ligase–POI interaction can significantly aid this process. Interestingly, neither the binding

affinity of the warhead nor of the E3 ligase ligand seems to directly influence the degradation efficiency of the PROTAC.[117]

Once the individual ligands have been selected, the next challenge is to design a molecule where these components work in concert. This harmony is essential for the formation of an effective ternary complex between the PROTAC, the target protein, and the E3 ligase. The spatial and temporal dynamics of this complex formation are critical. It's not just about bringing these entities into proximity; it's also about ensuring that they interact in a manner that facilitates the transfer of ubiquitin from the E3 ligase to the target protein. For instance, the spatial arrangement in a potential ternary complex needs to allow the POI's ubiquitination site to be accessible to the E3 ligase once the complex is formed. This may involve tweaking the linker length, rigidity, or chemical composition to achieve the optimal orientation.

Nevertheless, some ML-driven methods for PROTAC design seek to tackle the problem in a more holistic manner. Though less common than modular approaches, which focus heavily on PROTAC linker optimization, comprehensive PROTAC design strategies can be advantageous for a few reasons. They allow, in principle, for the simultaneous optimization of multiple parameters, such as flexibility, cell permeability, and degradation efficiency—factors determined not only by the linker composition, but also by that of the warhead and E3 ligand. A holistic approach may also better account for the complex interactions between the different PROTAC components, leading to the design of more effective and specific PROTACs; nevertheless, this remains to be rigorously demonstrated. In the next section, we examine the only study which, to our knowledge, has tackled the problem of engineering PROTACs in a holistic fashion, challenging traditional FBDD principles.

3.2 Previous work in comprehensive PROTAC design

A case study in ML-driven *de novo* PROTAC design is the application of GraphINVENT, a graph-based deep generative model, to the generation of novel PROTAC graphs predicted to represent highly active degraders.[119] The authors used policy-gradient reinforcement learning (RL) and a surrogate model for protein degradation activity to guide the model toward a chemical space of more active potential degraders. The open-source PROTAC-DB[114] database was used to train the model, which included 638 "complete" entries detailing degradation activity in various systems. For an entry to be considered complete, it needed to include a PROTAC SMILES, an E3 ligase, a POI, a defined cell type, and, crucially, a DC_{50} value. The authors reported that, despite their large size, GraphINVENT did not struggle to propose novel PROTAC-like structures starting from empty graphs; they go on to show that, following RL fine-tuning, the model could generate diverse molecules with not only higher predicted activity than the prior, but also substructures found in known degraders. An analysis was conducted to demonstrate that the model could be used to generate novel compounds with high predicted activity for IRAK3 degradation, but none of the proposed structures were experimentally validated.

3.3 Modeling degradation activity in PROTACs

A critical component of any generative model for PROTACs is a reliable surrogate model for degradation activity. To this end, a few data-driven approaches have been developed to tackle the prediction of degradation activity in PROTACs. Nori *et al.*[119] first used eXtreme Gradient Boosting (XGBoost) and Morgan fingerprints to classify potential degraders into "active" or "inactive" compounds. Following this work, Li *et al.*[120] introduced DeepPROTACs, a deep neural network architecture integrating graph convolutional networks (GCNs) and bidirectional long short-term memory (LSTM) layers for predicting degradation activity in PROTACs. DeepPROTACs was trained using data from PROTAC-DB and other public sources, including 2832 labeled datasets split into 988 "good" degraders and 1844 "bad" degraders based on their DC_{50} and D_{max} values. When tested on a dataset of 16 PROTACs targeting ER and VHL, the model achieved a prediction accuracy of 68.75%. For other PROTAC targets like EZH2, STAT3, eIF4E, and FLT3, the accuracy rates ranged from 65% to 80%. More recently, Ribes *et al.*[121] introduced a neural network ensemble model for the classification of PROTACs into "active" and "inactive" compounds. Here, the authors combined data from PROTAC-DB and PROTACpedia,[122] where only entries that had both a DC_{50} and D_{max} value reported were used. PROTACs were represented using Morgan fingerprints, while target proteins, E3 ligases, and cell types are each embedded into feature vectors using linear layers, and normalized. The model demonstrated superior performance to previous models for degradation activity classification, reaching a top test accuracy of 82.6% when using a stratified data split, and a test accuracy of 61% on unseen POIs. The authors conclude that the model will generalize well to novel PROTAC structures so long as both the POI and E3 ligase have been seen before in training.

3.4 Innovations and emerging trends

The field of PROTAC design is rapidly evolving along with new ML approaches to molecular design. One common thread in many of the aforementioned methods is the use of RL to learn optimal policies for PROTAC design through trial and error. Another emerging trend is the integration of 3D information into generative approaches. This allows for a more holistic view of the interactions between proteins and can lead to more effective PROTAC designs. Additionally, there is a growing trend towards the use of transfer learning, where a model developed for one task is reused as the starting point for a model on a second task. This is particularly useful in PROTAC design where the limited amount of public data poses a challenge.

Although diffusion models have not yet been applied to PROTAC design (only FBDD, as in DiffLinker[104]), we believe they present a promising direction in PROTAC engineering, both for linker-only and holistic design strategies. Firstly, diffusion models excel at generating high-quality molecular structures by gradually transforming simple distributions into complex data distributions.[123] Secondly, diffusion models can naturally integrate 3D information, which allows for the design of PROTACs that account for the spatial arrangement and interactions between the POI, the PROTAC, and the E3 ligase. Diffusion models are also known for their robustness in handling noisy data,[124] and they can be integrated with existing generative and predictive frameworks in an online setting.[125] These capabilities of diffusion models make them natural choices to explore further for generating diverse and novel PROTAC structures.

For a detailed summary of all models surveyed in this work, please see Table 1.

3.5 Datasets

There are currently two main sources of openly-accessible, structured PROTAC data: PROTAC-DB and PROTACpedia.

PROTAC-DB is a public database designed to support the research and development of PROTACs.[114] It offers an online repository of structural and experimental data related to these molecules. Data in the database is manually extracted from the literature or calculated using specific programs. In the second release, the number of PROTACs was expanded to 3270 and featured ∼360 warheads, ∼1500 linkers, and ∼80 E3 ligands. As of June 2024, PROTAC-DB contains 5388 entries. It also includes ternary complex structures for PROTACs. PROTAC-DB covers key aspects of PROTAC activity, including degradation capacity, quantified by metrics like DC_{50} and D_{max}; binding affinities between PROTACs (or PROTAC ligands) and target proteins and E3 ligases; cellular activities such as IC_{50}, EC_{50}, GI_{50}, and GR_{50}; and PAMPA and Caco-2 permeability data. Nevertheless, entries are not necessarily complete and there is a lot of missing data in PROTAC-DB, often because the original source does not report all aforementioned metrics.[121]

PROTACpedia is a curated database focused on PROTACs, containing detailed entries on 1190 PROTAC molecules as of the latest update (October 2022).[122] It contains high-quality data that has been carefully curated by experts, including information on ∼202 warheads, ∼65 E3 ligands, and ∼806 linkers. This platform facilitates the sharing and dissemination of critical PROTAC-related data to help expand PROTACpedia. Its collaborative nature encourages contributions to ensure that the database remains an up-to-date resource for researchers exploring PROTACs.

There is a significant overlap in the activity distributions of structures deposited in PROTACpedia with those in PROTAC-DB. As of June 2024, there are 807 PROTACs present in both databases, identified *via* string comparison following canonicalization of PROTAC SMILES from both databases. In other words, roughly 68% of PROTACs in PROTACpedia and 25% of PROTACs in PROTAC-DB are present in both databases. We did not explore what fraction of the duplicate PROTAC structures correspond to duplicate entries between the two databases, as it is possible that a PROTAC may be present in both databases but still contain information about different sets of experiments.

As PROTACs represent a relatively new therapeutic modality, there is a relative scarcity in the number of publicly available crystal structures, especially for ternary complexes. Structures that are available in the PDB have most frequently been determined using cryogenic electron microscopy (cryo-EM), a technology that has revolutionized the field of protein structural

Table 1 Summary of previous work on ML-guided PROTAC design. SMILES is a molecular string representation constructed from 2D graphs. RL: reinforcement learning. LSTM: long short-term memory, a class of recurrent neural network. JT-VAE: the junction-tree variational autoencoder.[126] CNN: convolutional neural network. GGNN: gated-graph neural network.[109] CL: curriculum learning. GCNs: graph convolutional networks. MLP: multi-layer perceptron. The annotation "(w/3D coords)" indicates that 3D information was used indirectly during structure generation (e.g., scoring), whereas "+3D coords" indicates 3D information was directly used in structure generation

Model	Year	Data	Type	Focus
SyntaLinker[96]	2020	SMILES	Transformers	Fragment linking
PROTAC-RL[112]	2022	SMILES	Transformers + RL	Fragment linking & PROTAC linker design
Link-INVENT[93]	2023	SMILES	LSTM + RL	Fragment linking & PROTAC linker design
ShapeLinker[116]	2023	SMILES (w/3D coords)	Link-INVENT	PROTAC linker design
DRlinker[106]	2022	SMILES (w/3D coords)	Transformers + RL	Fragment linking
DeLinker[98]	2020	2D graphs (w/3D coords)	JT-VAE	Fragment linking
DEVELOP[100]	2021	2D graphs + 3D coords	JT-VAE + CNN	Fragment linking
3DLinker[101]	2022	2D graphs + 3D coords	E(3) eq. graph VAE	Fragment linking
Nori et al.[119]	2022	2D graphs	GraphINVENT[110] (GGNN + RL)	Full ("holistic") PROTAC design
AIMLinker[111]	2023	2D graphs (w/3D coords)	GGNN	PROTAC linker design
GRELinker[108]	2024	2D graphs (w/3D coords)	GGNN + RL + CL	Fragment linking
DiffLinker[104]	2024	2D graphs + 3D coords	2D GNN + E(3) eq. 3D diffusion	Linker size prediction & fragment linking
Nori et al.[119]	2022	Morgan fingerprints	XGBoost	Degradation activity prediction
DeepPROTACs[120]	2022	SMILES + 3D graphs	GCNs + LSTMs	Degradation activity prediction
Ribes et al.[121]	2024	Morgan fingerprints	MLP	Degradation activity prediction

biology. Nevertheless, because the PROTAC MoA is not particularly well-understood, generalizations are being made across a range of PROTACs based on limited mechanistic data. Researchers would be wise to exercise caution when generalizing too far beyond the scope of their models or experiments.

Part of the challenge in ML-driven PROTAC engineering stems from the limited amount of structured data available. While public databases such as these have been greatly influential thus far in driving the development of ML tools for PROTAC design, without more comprehensive datasets, data-driven models will only be able to access a fraction of the vast chemical space accessible with PROTACs. Data scarcity becomes even more of a concern when considering factors like bioactivity, PK properties, and 3D structure in PROTACs. Low-data and low-resource learning can provide valuable strategies in the current scarce data landscape,[127] but, ultimately, more high-quality, structured data will need to be systematically generated and deposited following FAIR data-sharing principles for researchers to truly harness the powers of ML in PROTAC design. We hope that, just as ML has become an invaluable tool for identifying hits and optimizing leads in small-molecule drug discovery pipelines, it will also transform the current paradigm of PROTAC engineering, making us wonder how we ever managed without it.

4 Discussion

ML models trained on small-molecule datasets often struggle with generalization to novel chemical spaces not represented in the training data.[128,129] This can limit their predictive accuracy for entirely new compounds. PROTACs, with their bifunctional nature and larger size, represent a significant departure from small molecules. PROTACs generally have larger molecular weights and require higher degrees of conformational flexibility to achieve their function. Notably, the

physicochemical properties and PK profiles of PROTACs are markedly different from those of traditional small molecules. In Fig. 3, we highlight some of the key physicochemical differences between PROTACs and SMDs. These include differences in molecular weight (MW), partition coefficient (Log P), number of rotatable bonds (flexibility and conformational dynamics), number of HBDs and HBAs, and number of carbon atoms. By analyzing these descriptors, we can gain precise insights into the structural and physicochemical differences between PROTACs and small molecules.

● Molecular weight: the small-molecule MW distribution peaks around 250–500 Da. This is the typical range expected for drug-like small molecules as it is considered optimal for oral bioavailability according to Lipinski's Rule of Five. The PROTAC distribution peaks around 750–1000 Da, highlighting how much larger they are than traditional small molecules. The small-molecule distribution is relatively narrow and sharply peaked, indicating a more uniform range of MWs, while the PROTAC distribution is broader, reflecting greater variability in the size of these complex molecules. The clear separation between the two distributions highlights a key difference between PROTACs and traditional small molecules: their size.

● Partition coefficient: the Log P distribution for small molecules peaks around 2–3. This is consistent with drug-likeness criteria, where a Log P value between 1 and 3 is typically considered favorable for oral bioavailability. The Log P distribution for PROTACs is broader and peaks around 5. Higher Log P values indicate that PROTACs are generally more hydrophobic than small molecules, which can affect their solubility and cellular permeability. The higher Log P values for PROTACs may pose challenges for their solubility in aqueous environments like the extracellular environment and the cytosol, and may require formulation strategies to enhance their solubility and bioavailability.

● Rotatable bonds: the distribution in the number of rotatable bonds peaks around 1–5 rotatable bonds for small molecules. Fewer rotatable bonds are associated with greater rigidity. For PROTACs, the distribution instead peaks around 15–20 rotatable bonds. The higher number of rotatable bonds can be largely attributed to the flexible linker regions in PROTACs.

● Hydrogen bonds: the HBD distribution of small molecules peaks around 0–2. This is in line with the drug-likeness criteria that suggest a limited number of hydrogen bond donors to ensure good membrane permeability. On the other hand, the HBD distribution of PROTACs peaks around 3–5. Similar trends are observed for the HBA distributions.

● Carbon composition: both PROTACs and small molecules have a high normalized carbon count, peaking between 0.7–0.8, with the peak being slightly lower for PROTACs. Small molecules display a slightly broader distribution in normalized carbon atom count. No significant differences in normalized nitrogen, oxygen, or fluorine atom composition were observed, although small molecules do display marginally broader distributions for all these atom types.

This comparative analysis highlights the unique challenges and opportunities facing ML models for PROTAC design. Furthermore, it should be evident that models trained on small fragments will not capture the distinct features of PROTACs, as small molecule fragments and PROTACs exist in largely non-overlapping areas of chemical space. Extending ML models developed for small molecules to PROTACs requires modifications; this could entail small changes, like re-training on larger molecules and/or more diverse datasets that include PROTACs, or changes to fundamental design principles. It is well-known in deep learning, including generative modeling, that the predictive accuracy of a model depends heavily on the availability of high-quality datasets.[130] However, in drug design, datasets often have various forms of inconsistencies and missing data.[131] These challenges are even more pronounced when focusing on PROTAC design; here, experimental data on the efficacy and specificity of PROTAC molecules is even scarcer. The multifaceted nature of PROTACs necessitates detailed and high-quality datasets to uncover the subtle patterns underlying their biological activity. This is especially crucial for turning predictive models into tools for robust PROTAC optimization and design.

Models that fail to incorporate structural, or even dynamical, information regarding PROTACs and their target proteins might not effectively capture the feasibility of ternary complex formation. We know that the specificity and activity of PROTACs towards a specific POI are influenced by precise 3D interactions at the molecular level, more so than by the binding affinity. Without detailed 3D structural data, ML models may not be accurate enough to generalize to new PROTAC structures or even new POIs, a concern reflected in the changing landscape of ML models for PROTAC design: while early work focused primarily on 2D representations or simplified 3D information, all work we surveyed from the past two years involved the incorporation of more complex 3D data. We don't believe this change is due solely to advances in computing hardware and software. Rather, models capable of handling 3D data offer

a superior capability to capture the interplay of molecular shapes and complex spatial arrangements especially relevant in PROTAC function.

Another big challenge with PROTAC design is getting them into the cell.[59] As they depend on the proteasome for degrading their target proteins, PROTACs can only be used to target proteins found in the cytosol or with cytosolic domains (for membrane proteins), thus excluding as targets any proteins found outside the cell. According to the Human Protein Atlas, ~25% of all protein-coding human genes have been shown to encode proteins that localize to the cytosol and its substructures,[132] though this estimate does not include proteins that transiently reside in the cytosol. How to improve cellular permeability in PROTACs is thus an active area of research, as it imposes hard constraints on their efficacy. Notwithstanding, exactly which mechanism PROTACs use for entering the cell is not fully understood and may very well vary depending on the specific molecule and cell type, adding another layer of complexity to the task of cell permeability prediction.

To overcome the many challenges facing *de novo* PROTAC design, future ML methods must place a greater emphasis on accurately modeling the 3D structures of PROTACs and the corresponding ternary complexes they form. This could involve the development of molecular dynamics or physics-based approaches that leverage ML to simulate important molecular interactions and conformational changes at a coarse-grained or even atomistic level, and it could also involve experimental advances that allow us to better isolate and characterize these complexes, possibly with the assistance of active learning or other ML-driven strategies. Scientists leveraging ML have undeniably driven many recent breakthroughs in protein structure prediction[133] and conditional protein structure generation.[134] Perhaps it's time to apply similar guiding principles to PROTAC engineering (*e.g.*, systematic data collection and accessibility, better algorithms harnessing biological knowledge), and see what breakthroughs we can achieve in this domain.

5 Conclusion

PROTACs differ fundamentally from small-molecule drugs (SMDs) in their mechanism of action. SMDs such as inhibitors typically function by blocking a protein's active site and thus its activity, whereas PROTACs instead carry out a complicated dance inside the cell, which, if performed correctly, will lead to the degradation of the target protein. This means that rather than simply inhibiting a protein's function, a PROTAC removes it from the cell. Notably, it is not consumed in the process, which means it can go on to cause the degradation of many other copies of the target protein before it is metabolized and/or excreted. This can be more effective in cases where simple inhibition of activity is insufficient for a therapeutic effect, as is the case in many diseases lacking effective first-line treatments, including many cancers. PROTACs offer an alternative pathway to drugging otherwise undruggable proteins in the cytosol.

In this comprehensive review, we have highlighted the significant impact of ML on PROTAC design. The complexities involved in PROTACs make traditional ML in the context of

FBDD less effective. These complexities include the unique mechanism of action of PROTACs, the delicate spatial configuration required for effective protein degradation, and the need for favorable PK profiles for drug-like compounds, which are not adequately captured by models designed for small molecular fragments. Advanced ML techniques, such as generative models tailored to PROTAC peculiarities offer promising solutions for optimizing PROTAC design.

In the hope of spurring more research in what we view as a hugely impactful but formidable research direction, we have prepared this comprehensive review on ML for PROTAC design. We hope that this review and the insights described in it serve as a comprehensive guide to researchers looking to apply their deep ML knowledge to the design of an exciting "new" therapeutic modality, or conversely, to enable biologists to venture into the rewarding world of deep generative models. The synergy between ML and PROTAC design holds immense potential, and we encourage further research in this pioneering domain.

Data availability

This data analysis was carried out using publicly available data from PROTAC-DB at **https://cadd.zju.edu.cn/protacdb/downloadswith**, PROTACPedia at **https://protacpedia.weizmann.ac.il/ptcb/download** and ZINC-250 dataset at **https://www.kaggle.com/datasets/basu369victor/zinc250k**. The data analysis script of this paper is available in the interactive Google Colab notebook "PROTACs *vs.* small molecules.ipynb" at **https://colab.research.google.com/drive/1MIYJcEytzz_glOjyKcyXCVHGgDGChNX0?usp=sharing**.

Author contributions

YG conducted the initial literature review, with guidance from RM. YG prepared the initial draft of the manuscript and figures. RM and YG engaged in extensive discussions and iterations of writing and feedback throughout the process. Both authors contributed significantly to the development and refinement of the manuscript. Both authors have read and approved the final version of the manuscript.

Conflicts of interest

There are no conflicts to declare.

Acknowledgements

Notes and references

1 A. C. Lai and C. M. Crews, *Nat. Rev. Drug Discovery*, 2017, **16**, 101–114.

2 M. Schapira, M. F. Calabrese, A. N. Bullock and C. M. Crews, *Nat. Rev. Drug Discovery*, 2019, **18**, 949–963.

3 M. Békés, D. Langley and C. Crews, *Nat. Rev. Drug Discovery*, 2022, **21**, 181–200.

4 L. Zhao, J. Zhao, K. Zhong, A. Tong and D. Jia, *Signal Transduction Targeted Ther.*, 2022, **7**, 113.

5 J. Kim, H. Kim and S. B. Park, *J. Am. Chem. Soc.*, 2014, **136**, 14629–14638.

6 P. Martín-Acosta and X. Xiao, *Eur. J. Med. Chem.*, 2021, **210**, 112993.

7 M. J. Clague, C. Heride and S. Urbé, *Trends Cell Biol.*, 2015, **25**, 417–426.

8 S. Gu, D. Cui, X. Chen, X. Xiong and Y. Zhao, *BioEssays*, 2018, **40**, 1700247.

9 M. Xi, Y. Chen, H. Yang, H. Xu, C. Du, C. Wu, Y. Xu, L. Deng, X. Luo, L. Yu, *et al.*, *Eur. J. Med. Chem.*, 2019, **174**, 159–180.

10 Z. Liu, M. Hu, Y. Yang, C. Du, H. Zhou, C. Liu, Y. Chen, L. Fan, H. Ma, Y. Gong, *et al.*, *Mol. Biomed.*, 2022, **3**, 46.

11 K. G. Coleman and C. M. Crews, *Annu. Rev. Cancer Biol.*, 2018, **2**, 41–58.

12 K. M. Sakamoto, K. B. Kim, A. Kumagai, F. Mercurio, C. M. Crews and R. J. Deshaies, *Proc. Natl. Acad. Sci. U. S. A.*, 2001, **98**, 8554–8559.

13 D. C. Cannatella and R. O. De Sá, *Syst. Biol.*, 1993, **42**, 476–507.

14 K. M. Sakamoto, K. B. Kim, R. Verma, A. Ransick, B. Stein, C. M. Crews and R. J. Deshaies, *Mol. Cell. Proteomics*, 2003, **2**, 1350–1358.

15 J.-M. Renoir, V. Marsaud and G. Lazennec, *Biochem. Pharmacol.*, 2013, **85**, 449–465.

16 O. Astapova, C. Seger and S. Hammes, *J. Endocr. Soc.*, 2021, **5**, bvab035.

17 R. P. Bhole, P. R. Kute, R. V. Chikhale, C. Bonde, A. Pant and S. S. Gurav, *Bioorg. Chem.*, 2023, **139**, 106720.

18 X. Li and Y. Song, *J. Hematol. Oncol.*, 2020, **13**, 1–14.

19 J. Liu, J. Ma, Y. Liu, J. Xia, Y. Li, Z. P. Wang and W. Wei, *Semin. Cancer Biol.*, 2020, 171–179.

20 A. R. Schneekloth, M. Pucheault, H. S. Tae and C. M. Crews, *Bioorg. Med. Chem. Lett.*, 2008, **18**, 5904–5908.

21 H. Li, J. Dong, M. Cai, Z. Xu, X.-D. Cheng and J.-J. Qin, *J. Hematol. Oncol.*, 2021, **14**, 1–23.

22 Y. Ma, Y. Gu, Q. Zhang, Y. Han, S. Yu, Z. Lu and J. Chen, *Mol. Cancer Ther.*, 2013, **12**, 286–294.

23 L. Bai, H. Zhou, R. Xu, Y. Zhao, K. Chinnaswamy, D. McEachern, J. Chen, C.-Y. Yang, Z. Liu, M. Wang, *et al.*, *Cancer Cell*, 2019, **36**, 498–511.

24 Y. Wang, Y. Zhou, S. Cao, Y. Sun, Z. Dong, C. Li, H. Wang, Y. Yao, H. Yu, X. Song, *et al.*, *Bioorg. Chem.*, 2021, **111**, 104833.

25 P. S. Dragovich, T. H. Pillow, R. A. Blake, J. D. Sadowsky, E. Adaligil, P. Adhikari, J. Chen, N. Corr, J. dela Cruz-Chuh, G. Del Rosario, *et al.*, *J. Med. Chem.*, 2021, **64**, 2576–2607.

26 S. Yan, G. Zhang, W. Luo, M. Xu, R. Peng, Z. Du, Y. Liu, Z. Bai, X. Xiao and S. Qin, *Eur. J. Med. Chem.*, 2024, **276**, 116725.

27 T. Neklesa, L. B. Snyder, R. R. Willard, N. Vitale, J. Pizzano, D. A. Gordon, M. Bookbinder, J. Macaluso, H. Dong, C. Ferraro, et al., J. Clin. Oncol., 2019, 37, 10–1200.

28 D. P. Petrylak, X. Gao, N. J. Vogelzang, M. H. Garfield, I. Taylor, M. Dougan Moore, R. A. Peck and H. A. Burris III, First-in-human phase I study of ARV-110, an androgen receptor (AR) PROTAC degrader in patients (pts) with metastatic castrate-resistant prostate cancer (mCRPC) following enzalutamide (ENZ) and/or abiraterone (ABI), J. Clin. Oncol., 2020, 38, 3500.

29 L. B. Snyder, J. J. Flanagan, Y. Qian, S. M. Gough, M. Andreoli, M. Bookbinder, G. Cadelina, J. Bradley, E. Rousseau, J. Chandler, et al., Cancer Res., 2021, 81, 44.

30 G. M. Burslem and C. M. Crews, Cell, 2020, 181, 102–114.

31 J. Liu, Y. Peng and W. Wei, Front. Cell Dev. Biol., 2021, 9, 678077.

32 M. Xiao, J. Zhao, Q. Wang, J. Liu and L. Ma, Biomolecules, 2022, 12, 1257.

33 M. Reynders, B. S. Matsuura, M. Bérouti, D. Simoneschi, A. Marzio, M. Pagano and D. Trauner, Sci. Adv., 2020, 6, eaay5064.

34 H. Lebraud, D. J. Wright, C. N. Johnson and T. D. Heightman, ACS Cent. Sci., 2016, 2, 927–934.

35 M. Pettersson and C. M. Crews, Drug Discovery Today: Technol., 2019, 31, 15–27.

36 Y.-J. Chen, H. Wu and X.-Z. Shen, Cancer Lett., 2016, 379, 245–252.

37 N. R. Jana, Neurochem. Int., 2012, 60, 443–447.

38 N. P. Dantuma and L. C. Bott, Front. Mol. Neurosci., 2014, 7, 70.

39 M. F. Schmidt, Z. Y. Gan, D. Komander and G. Dewson, Cell Death Differ., 2021, 28, 570–590.

40 B. Liu, J. Ruan, M. Chen, Z. Li, G. Manjengwa, D. Schlüter, W. Song and X. Wang, Mol. Psychiatry, 2022, 27, 259–268.

41 Y. Liang, G. Zhong, M. Ren, T. Sun, Y. Li, M. Ye, C. Ma, Y. Guo and C. Liu, NeuroMol. Med., 2023, 1–18.

42 C. Tokheim, X. Wang, R. T. Timms, B. Zhang, E. L. Mena, B. Wang, C. Chen, J. Ge, J. Chu, W. Zhang, et al., Mol. Cell, 2021, 81, 1292–1308.

43 J. A. Bard, E. A. Goodall, E. R. Greene, E. Jonsson, K. C. Dong and A. Martin, Annu. Rev. Biochem., 2018, 87, 697–724.

44 S. Lorenz, A. J. Cantor, M. Rape and J. Kuriyan, BMC Biol., 2013, 11, 1–12.

45 Q. Yang, J. Zhao, D. Chen and Y. Wang, Mol. Biomed., 2021, 2, 1–17.

46 J. Weber, S. Polo and E. Maspero, Front. Physiol., 2019, 10, 443919.

47 E. Branigan, J. Carlos Penedo and R. T. Hay, Nat. Commun., 2020, 11, 2846.

48 G. A. Collins and A. L. Goldberg, Cell, 2017, 169, 792–806.

49 N. I. Sincere, K. Anand, S. Ashique, J. Yang and C. You, Molecules, 2023, 28, 4014.

50 M. S. Gadd, A. Testa, X. Lucas, K.-H. Chan, W. Chen, D. J. Lamont, M. Zengerle and A. Ciulli, Nat. Chem. Biol., 2017, 13, 514–521.

51 R. P. Wurz, H. Rui, K. Dellamaggiore, S. Ghimire-Rijal, K. Choi, K. Smither, A. Amegadzie, N. Chen, X. Li, A. Banerjee, et al., Nat. Commun., 2023, 14, 4177.

52 A. Zorba, C. Nguyen, Y. Xu, J. Starr, K. Borzilleri, J. Smith, H. Zhu, K. A. Farley, W. Ding, J. Schiemer, et al., Proc. Natl. Acad. Sci. U. S. A., 2018, 115, E7285–E7292.

53 K.-H. Chan, M. Zengerle, A. Testa and A. Ciulli, J. Med. Chem., 2018, 61, 504–513.

54 R. P. Nowak, S. L. DeAngelo, D. Buckley, Z. He, K. A. Donovan, J. An, N. Safaee, M. P. Jedrychowski, C. M. Ponthier, M. Ishoey, et al., Nat. Chem. Biol., 2018, 14, 706–714.

55 B. E. Smith, S. L. Wang, S. Jaime-Figueroa, A. Harbin, J. Wang, B. D. Hamman and C. M. Crews, Nat. Commun., 2019, 10, 131.

56 R. G. Guenette, S. W. Yang, J. Min, B. Pei and P. R. Potts, Chem. Soc. Rev., 2022, 51, 5740–5756.

57 R. Troup, C. Fallan and M. Baud, Explor. Targeted Anti-Tumor Ther., 2020, 1, 273.

58 A. Zagidullin, V. Milyukov, A. Rizvanov and E. Bulatov, Explor. Targeted Anti-Tumor Ther., 2020, 1, 381.

59 H. Yokoo, M. Naito and Y. Demizu, Expert Opin. Drug Discovery, 2023, 18, 357–361.

60 V. G. Klein, A. G. Bond, C. Craigon, R. S. Lokey and A. Ciulli, J. Med. Chem., 2021, 64, 18082–18101.

61 Z. Xie, X. Yang, Y. Duan, J. Han and C. Liao, J. Med. Chem., 2021, 64, 1283–1345.

62 J. Dong, G. Huang, Q. Cui, Q. Meng, S. Li and J. Cui, Eur. J. Med. Chem., 2021, 209, 112895.

63 J. Dodson and P. A. Lio, Curr. Allergy Asthma Rep., 2022, 22, 183–193.

64 T. K. Neklesa, J. D. Winkler and C. M. Crews, Pharmacol. Ther., 2017, 174, 138–144.

65 A. L. Hopkins and C. R. Groom, Nat. Rev. Drug Discovery, 2002, 1, 727–730.

66 J. P. Overington, B. Al-Lazikani and A. L. Hopkins, Nat. Rev. Drug Discovery, 2006, 5, 993–996.

67 T. I. Oprea, C. G. Bologa, S. Brunak, A. Campbell, G. N. Gan, A. Gaulton, S. M. Gomez, R. Guha, A. Hersey, J. Holmes, et al., Nat. Rev. Drug Discovery, 2018, 17, 317–332.

68 C. V. Dang, E. P. Reddy, K. M. Shokat and L. Soucek, Nat. Rev. Cancer, 2017, 17, 502–508.

69 A. T. Vicente and J. A. Salvador, Int. J. Mol. Sci., 2022, 23, 11068.

70 X. Han, C. Wang, C. Qin, W. Xiang, E. Fernandez-Salas, C.-Y. Yang, M. Wang, L. Zhao, T. Xu, K. Chinnaswamy, et al., J. Med. Chem., 2019, 62, 941–964.

71 Y. Sun, X. Zhao, N. Ding, H. Gao, Y. Wu, Y. Yang, M. Zhao, J. Hwang, Y. Song, W. Liu, et al., Cell Research, 2018, 28, 779–781.

72 A. D. Buhimschi, H. A. Armstrong, M. Toure, S. Jaime-Figueroa, T. L. Chen, A. M. Lehman, J. A. Woyach, A. J. Johnson, J. C. Byrd and C. M. Crews, Biochemistry, 2018, 57, 3564–3575.

73 Y. Sun, N. Ding, Y. Song, Z. Yang, W. Liu, J. Zhu and Y. Rao, Leukemia, 2019, 33, 2105–2110.

74 M. Brand, B. Jiang, S. Bauer, K. A. Donovan, Y. Liang, E. S. Wang, R. P. Nowak, J. C. Yuan, T. Zhang, N. Kwiatkowski, *et al.*, *Cell Chem. Biol.*, 2019, **26**, 300–306.

75 B. Jiang, E. S. Wang, K. A. Donovan, Y. Liang, E. S. Fischer, T. Zhang and N. S. Gray, *Angew. Chem., Int. Ed.*, 2019, **58**, 6321–6326.

76 D. P. Bondeson, A. Mares, I. E. Smith, E. Ko, S. Campos, A. H. Miah, K. E. Mulholland, N. Routly, D. L. Buckley, J. L. Gustafson, *et al.*, *Nat. Chem. Biol.*, 2015, **11**, 611–617.

77 M. Konstantinidou, J. Li, B. Zhang, Z. Wang, S. Shaabani, F. Brake, K. Essa and A. Domling, *Expert Opin. Drug Discovery*, 2019, **14**, 1–14.

78 D. Nath and S. Shadan, *Nature*, 2009, **458**, 421.

79 C. M. Olson, B. Jiang, M. A. Erb, Y. Liang, Z. M. Doctor, Z. Zhang, T. Zhang, N. Kwiatkowski, M. Boukhali, J. L. Green, *et al.*, *Nat. Chem. Biol.*, 2018, **14**, 163–170.

80 J. Vamathevan, D. Clark, P. Czodrowski, I. Dunham, E. Ferran, G. Lee, B. Li, A. Madabhushi, P. Shah, M. Spitzer, *et al.*, *Nat. Rev. Drug Discovery*, 2019, **18**, 463–477.

81 S. Dara, S. Dhamercherla, S. S. Jadav, C. M. Babu and M. J. Ahsan, *Artif. Intell. Rev.*, 2022, **55**, 1947–1999.

82 C. Cecchini, S. Tardy and L. Scapozza, *Chimia*, 2022, **76**, 341.

83 T. A. Bemis, J. J. La Clair and M. D. Burkart, *J. Med. Chem.*, 2021, **64**, 8042–8052.

84 D. P. Bondeson, B. E. Smith, G. M. Burslem, A. D. Buhimschi, J. Hines, S. Jaime-Figueroa, J. Wang, B. D. Hamman, A. Ishchenko and C. M. Crews, *Cell Chem. Biol.*, 2018, **25**, 78–87.

85 M. J. Roy, S. Winkler, S. J. Hughes, C. Whitworth, M. Galant, W. Farnaby, K. Rumpel and A. Ciulli, *ACS Chem. Biol.*, 2019, **14**, 361–368.

86 K. Cyrus, M. Wehenkel, E.-Y. Choi, H.-J. Han, H. Lee, H. Swanson and K.-B. Kim, *Mol. BioSyst.*, 2011, **7**, 359–364.

87 X. Wang, S. Feng, J. Fan, X. Li, Q. Wen and N. Luo, *Biochem. Pharmacol.*, 2016, **116**, 200–209.

88 C. Qin, Y. Hu, B. Zhou, E. Fernandez-Salas, C.-Y. Yang, L. Liu, D. McEachern, S. Przybranowski, M. Wang, J. Stuckey, *et al.*, *J. Med. Chem.*, 2018, **61**, 6685–6704.

89 A. P. Crew, K. Raina, H. Dong, Y. Qian, J. Wang, D. Vigil, Y. V. Serebrenik, B. D. Hamman, A. Morgan, C. Ferraro, *et al.*, *J. Med. Chem.*, 2018, **61**, 583–598.

90 H. Zhao, *ChemMedChem*, 2024, e202400171.

91 M. Bon, A. Bilsland, J. Bower and K. McAulay, *Mol. Oncol.*, 2022, **16**, 3761–3777.

92 Q. Li, *Front. Mol. Biosci.*, 2020, **7**, 180.

93 J. Guo, F. Knuth, C. Margreitter, J. P. Janet, K. Papadopoulos, O. Engkvist and A. Patronov, *Digital Discovery*, 2023, **2**, 392–408.

94 Y. Cheng, Y. Gong, Y. Liu, B. Song and Q. Zou, *Briefings Bioinf.*, 2021, **22**, bbab344.

95 C. Pang, J. Qiao, X. Zeng, Q. Zou and L. Wei, *J. Chem. Inf. Model.*, 2023, **64**(7), 2174–2194.

96 Y. Yang, S. Zheng, S. Su, C. Zhao, J. Xu and H. Chen, *Chem. Sci.*, 2020, **11**, 8312–8322.

97 A. Gaulton, L. J. Bellis, A. P. Bento, J. Chambers, M. Davies, A. Hersey, Y. Light, S. McGlinchey, D. Michalovich, B. Al-Lazikani, *et al.*, *Nucleic Acids Res.*, 2012, **40**, D1100–D1107.

98 F. Imrie, A. R. Bradley, M. van der Schaar and C. M. Deane, *J. Chem. Inf. Model.*, 2020, **60**, 1983–1995.

99 R. Gómez-Bombarelli, J. N. Wei, D. Duvenaud, J. M. Hernández-Lobato, B. Sánchez-Lengeling, D. Sheberla, J. Aguilera-Iparraguirre, T. D. Hirzel, R. P. Adams and A. Aspuru-Guzik, *ACS Cent. Sci.*, 2018, **4**, 268–276.

100 F. Imrie, T. E. Hadfield, A. R. Bradley and C. M. Deane, *Chem. Sci.*, 2021, **12**, 14577–14589.

101 Y. Huang, X. Peng, J. Ma and M. Zhang, *arXiv*, 2022, preprint, arXiv:2205.07309, DOI: 10.48550/arXiv.2205.07309.

102 J. J. Irwin, K. G. Tang, J. Young, C. Dandarchuluun, B. R. Wong, M. Khurelbaatar, Y. S. Moroz, J. Mayfield and R. A. Sayle, *J. Chem. Inf. Model.*, 2020, **60**, 6065–6073.

103 M. Xu, W. Wang, S. Luo, C. Shi, Y. Bengio, R. Gomez-Bombarelli and J. Tang, *International Conference on Machine Learning*, 2021, pp. 11537–11547.

104 I. Igashov, H. Stärk, C. Vignac, A. Schneuing, V. G. Satorras, P. Frossard, M. Welling, M. Bronstein and B. Correia, *Nat. Mach. Intell.*, 2024, 1–11.

105 S. Axelrod and R. Gomez-Bombarelli, *Sci. Data*, 2022, **9**, 185.

106 Y. Tan, L. Dai, W. Huang, Y. Guo, S. Zheng, J. Lei, H. Chen and Y. Yang, *J. Chem. Inf. Model.*, 2022, **62**, 5907–5917.

107 M. Su, Q. Yang, Y. Du, G. Feng, Z. Liu, Y. Li and R. Wang, *J. Chem. Inf. Model.*, 2018, **59**, 895–913.

108 H. Zhang, J. Huang, J. Xie, W. Huang, Y. Yang, M. Xu, J. Lei and H. Chen, *J. Chem. Inf. Model.*, 2024.

109 Y. Li, D. Tarlow, M. Brockschmidt and R. Zemel, *arXiv*, 2015, preprint, arXiv:1511.05493, DOI: 10.48550/arXiv.1511.05493.

110 R. Mercado, T. Rastemo, E. Lindelöf, G. Klambauer, O. Engkvist, H. Chen and E. J. Bjerrum, *Mach. Learn.: Sci. Technol.*, 2021, **2**, 025023.

111 C.-T. Kao, C.-T. Lin, C.-L. Chou and C.-C. Lin, *J. Chem. Inf. Model.*, 2023, **63**, 2918–2927.

112 S. Zheng, Y. Tan, Z. Wang, C. Li, Z. Zhang, X. Sang, H. Chen and Y. Yang, *Nat. Mach. Intell.*, 2022, **4**, 739–748.

113 Q. Liu, M. Allamanis, M. Brockschmidt and A. Gaunt, *Adv. Neural Inf. Process. Syst.*, 2018, **31**.

114 G. Weng, X. Cai, D. Cao, H. Du, C. Shen, Y. Deng, Q. He, B. Yang, D. Li and T. Hou, *Nucleic Acids Res.*, 2023, **51**, D1367–D1372.

115 T. Pantsar and A. Poso, *Molecules*, 2018, **23**, 1899.

116 R. M. Neeser, M. Akdel, D. Kovtun and L. Naef, *arXiv*, 2023, preprint, arXiv:2306.08166, DOI: 10.48550/arXiv.2306.08166.

117 S. He, G. Dong, J. Cheng, Y. Wu and C. Sheng, *Med. Res. Rev.*, 2022, **42**, 1280–1342.

118 X. Li, W. Pu, Q. Zheng, M. Ai, S. Chen and Y. Peng, *Mol. Cancer*, 2022, **21**, 99.

119 D. Nori, C. W. Coley and R. Mercado, *arXiv*, 2022, preprint, arXiv:2211.02660, DOI: 10.48550/arXiv.2211.02660.

120 F. Li, Q. Hu, X. Zhang, R. Sun, Z. Liu, S. Wu, S. Tian, X. Ma, Z. Dai, X. Yang, *et al.*, *Nat. Commun.*, 2022, **13**, 7133.

121 S. Ribes, E. Nittinger, C. Tyrchan and R. Mercado, *arXiv*, 2024, preprint, arXiv:2406.02637, DOI: **10.48550/arXiv.2406.02637**.

122 N. London and J. Prilusky, PROTACpedia, **https://protacpedia.weizmann.ac.il/**, accessed: 2024-05-21.

123 L. Yang, Z. Zhang, Y. Song, S. Hong, R. Xu, Y. Zhao, W. Zhang, B. Cui and M.-H. Yang, *ACM Comput. Surv*, 2023, **56**, 1–39.

124 Z. Guo, J. Liu, Y. Wang, M. Chen, D. Wang, D. Xu and J. Cheng, *Nat. Rev. Bioeng.*, 2024, **2**, 136–154.

125 M. Uehara, Y. Zhao, K. Black, E. Hajiramezanali, G. Scalia, N. L. Diamant, A. M. Tseng, S. Levine and T. Biancalani, *arXiv*, 2024, preprint, arXiv:2402.16359, DOI: **10.48550/arXiv.2402.16359**.

126 W. Jin, K. Yang, R. Barzilay and T. Jaakkola, *arXiv*, 2018, preprint, arXiv:1812.01070, DOI: **10.48550/arXiv.1812.01070**.

127 D. van Tilborg, H. Brinkmann, E. Criscuolo, L. Rossen, R. Özçelik and F. Grisoni, *Curr. Opin. Struct. Biol.*, 2024, **86**, 102818.

128 M. Glavatskikh, J. Leguy, G. Hunault, T. Cauchy and B. Da Mota, *J. Cheminf.*, 2019, **11**, 1–15.

129 F. Kretschmer, J. Seipp, M. Ludwig, G. W. Klau and S. Boecker, *bioRxiv*, 2023, preprint, DOI: **10.1101/2023.03.27.534311**.

130 *Nat. Biotechnol.*, 2023, **41**, 433.

131 W. R. Myers, *Drug Inf. J.*, 2000, **34**, 525–533.

132 The Human Protein Atlas, **https://www.proteinatlas.org/humanproteome/subcellular/cytosol**, accessed: 2024-06-18.

133 J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, *et al.*, *Nature*, 2024, 1–3.

134 J. L. Watson, D. Juergens, N. R. Bennett, B. L. Trippe, J. Yim, H. E. Eisenach, W. Ahern, A. J. Borst, R. J. Ragotte, L. F. Milles, *et al.*, *Nature*, 2023, **620**, 1089–1100.