Computer Methods and Programs in Biomedicine Update 10 (2026) 100257

Contents lists available at ScienceDirect

Computer Methods and Programs in Biomedicine Update

journal homepage:
www.sciencedirect.com/journal/computer-methods-and-programs-in-biomedicine-update

PharMistral: Drug generation for novel protein targets by Mistral Large Language Model

Asma Bendjeddou [a], Kamyar Zeinalipour [a], Dalal Bardou [b], Marco Maggini [a], Franco Scarselli [a], Monica Bianchini [a,*]

[a] University of Siena, Dipartimento di Ingegneria dell'Inforrmazione e Scienze Matematiche, Via Roma 56, 53100 Siena, Italy
[b] LMIA, Department of Computer Science, Abbes Laghrour University of Khenchela, 40000, Algeria

ARTICLE INFO

Keywords:
Drug discovery
Mistral Large Language Model
Drug generation
Conditional molecule generation
Drug candidates

ABSTRACT

Designing molecules for targets lacking abundant data or 3D structures remains a bottleneck, as most models rely on known binders or pocket geometries. To address this, we present *PharMistral*, a sequence-conditioned ligand generator based on Mistral-7B. Unlike DeepTarget (task-specific) or ChemGPT (SMILES-only), PharMistral adapts a single 7B-parameter model via a two-stage process: (1) joint pre-training on more than 300,000 unpaired human proteins and more than 1 million drug-like SMILES to learn a shared biochemical language; and (2) end-to-end fine-tuning on ~300,000 high-affinity human protein–ligand pairs curated from ChEMBL-26. At inference, PharMistral autoregressively generates ligands from raw protein sequences without target-specific retraining or structural input. On unseen proteins, the model achieved 99.5% chemical validity. Of valid molecules, 57% were unique, and 36% of those were novel compared to the training set. After drug-likeness and toxicity filtering, 3383 unique molecules were retained (55% of the unique valid set and 31% of all valid molecules).

1. Introduction

Large language models (LLMs) and deep generative models offer vast potential for drug discovery [1–3]. However, designing ligands for emerging targets remains bottlenecked by the need for extensive protein–ligand pairs or high-quality 3D structures, data that are often unavailable despite advances like AlphaFold2 [4]. While methods like DeepTarget [5] bypass structural constraints via sequence conditioning, they lack the scale of modern LLMs. Conversely, models like ChemGPT [6] or BioMistral [7] excel in specific modalities (molecules or text) but lack joint protein–ligand understanding.

To address this limit, we introduce *PharMistral*, the first 7B-parameter open-weight LLM jointly adapted to unpaired protein sequences and small-molecule SMILES, then fine-tuned end-to-end for sequence-only ligand generation. We investigate whether this approach can generate valid, novel, drug-like ligands for unseen targets without structural data, focusing on four Research Questions (RQs):

RQ1 Chemical quality & diversity: What fraction of generated ligands are chemically valid, synthetically accessible, and novel with respect to the training set?

RQ2 Target relevance: How do docking scores and interactions with DeepTarget and baseline methods?

RQ3 Target generalization: How does the model perform on strictly excluded targets and under asymmetric training splits?

RQ4 Drug-likeness: Do generated candidates satisfy ADMET, QED, and synthetic accessibility (SA) criteria?

PharMistral employs a two-stage training pipeline: (1) joint continued pre-training on >300 K human protein sequences and >1 M drug-like SMILES to learn a shared representation; and (2) fine-tuning on ~300 K high-affinity protein–ligand pairs from ChEMBL-26 [8]. Our main contributions include:

(1) *Joint biochemical adaptation:* A single transformer jointly processes protein sequences and molecular representations using the original Mistral tokenizer.
(2) *Sequence-only ligand generation for unseen proteins:* PharMistral generates valid, novel, and diverse ligands for fully unseen targets using only protein sequences.
(3) *Multi-objective evaluation:* A comprehensive benchmark against DeepTarget [5], covering chemical validity, docking performance, ADMET properties, QED [9], and synthetic accessibility [10].
(4) *Strong performance on unseen targets:* PharMistral achieves 99.5% chemical validity on held-out targets. Among valid molecules, 57%

* Correspondence to: Via Roma 56, I-53100, Siena, Italy.
E-mail address: monica.bianchini@unisi.it (M. Bianchini).

https://doi.org/10.1016/j.cmpbup.2026.100257
Received 9 February 2026; Received in revised form 30 March 2026; Accepted 7 June 2026
Available online 12 June 2026

are unique and 36% are novel, with 55% of unique molecules passing physicochemical filtering. For 9 out of 17 unseen proteins, generated ligands achieve docking scores up to 1.9 kcal/mol better than the best ChEMBL reference ligands, as evaluated using AutoDock Vina [11]

(5) *Open resources:* Public release of checkpoints and scripts to facilitate reproducibility.

The paper follows this structure: Related Work (Section 2), Methodology (Section 3), Evaluation (Section 4), Results (Section 5), and Conclusion (Section 6).

2. Related work

Recent advances in deep generative models have substantially impacted *de novo* drug design by enabling molecular generation conditioned on biochemical constraints such as protein sequences, three-dimensional structures, or desired molecular properties. Existing approaches broadly fall into four paradigms: large pre-trained language models, structure-conditioned generation, sequence-conditioned generation, and SMILES-based molecular language modeling. We briefly review different approaches and position our work within this landscape.

2.1. Large pre-trained language models

Large language models (LLMs) such as GPT-3 [12] and LLaMA [13] exhibit strong zero-shot generalization, motivating their adaptation to biological domains. Protein-focused models, including ESM-2 [14] and ProtGPT2 [15], demonstrate that structure and function can be inferred directly from amino acid sequences.

Joint modeling of proteins and small molecules further improves interaction prediction. Methods such as DeepAffinity [16] learn unified compound-protein representations from sequences and achieve accurate binding affinity estimation. More recently, lightweight adaptation of Mistral-7B in BioMistral [7] showed competitive performance on protein-centric tasks despite limited data, highlighting the promise of large decoder-style LLMs for biomedical applications.

2.2. Structure-conditioned generation

Structure-based methods aim to generate ligands *in situ* by explicitly modeling protein binding pockets. Approaches such as RELATION [17] and DeepLigBuilder [18] assemble atoms or fragments within three-dimensional binding sites using graph-based representations. While effective in achieving favorable docking scores, these methods depend heavily on high-quality structural data and are computationally expensive. Related geometric learning approaches, such as EquiBind [19], focus primarily on pose prediction rather than de novo ligand generation.

2.3. Sequence-conditioned generation

To bypass reliance on three-dimensional structural data, several works condition molecular generation directly on protein sequences. The task was first framed as a sequence-to-sequence translation problem from amino acid sequences to SMILES in [20]. DeepTarget [5] later introduced a contrastive learning framework aligning protein and ligand embeddings in a shared latent space, enabling sequence-only ligand generation without structural input. However, its limited scale and lack of extensive pretraining constrain chemical diversity and generalization to low-homology or fully unseen targets.

Other models, such as DrugGPT [21], adopt a two-stage autoregressive framework combining SMILES pretraining with protein–ligand fine-tuning, but still require target-specific supervision, limiting applicability to truly novel proteins.

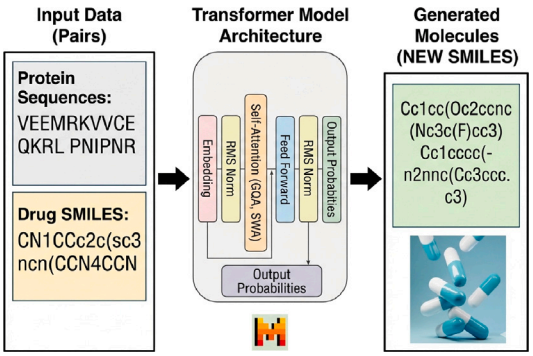


Fig. 1. Overview of the *PharMistral* architecture for generating novel drug-like SMILES from protein–drug input pairs.

2.4. SMILES-based language modeling

Early molecular generators modeled SMILES strings using recurrent neural networks [22,23], introducing key ideas such as SMILES randomization and variational decoding. Transformer-based architectures later improved scalability and generation quality. ChemBERTa [24] demonstrated the benefits of large-scale self-supervised pretraining for molecular property prediction, while MolGPT [25] achieved high validity and novelty through autoregressive SMILES generation.

More recently, large-scale chemical language models such as ChemGPT [6] extended this paradigm to the billion-parameter regime, achieving near-perfect validity using SELFIES. These results highlight the effectiveness of large decoder-only models for molecular generation, but such approaches lack explicit conditioning on protein targets.

Despite these advances, designing ligands for proteins without known structures or experimentally verified binders remains challenging, particularly for unseen targets [5,15,26]. Existing methods often exhibit limited generalization or rely on target-specific fine-tuning.

To address this gap, we propose **PharMistral**, a unified transformer-based generator that jointly models protein sequences and molecular representations. The model is first pre-trained on large-scale unpaired protein sequences and SMILES strings to learn a shared biochemical representation, and subsequently fine-tuned on curated protein–ligand pairs. Unlike prior approaches, PharMistral is evaluated in a target–agnostic, protein-held-out setting, enabling the generation of valid, diverse, and target-specific ligands for entirely unseen proteins. We demonstrate state-of-the-art chemical validity and diversity, and consistent improvements over DeepTarget across validity, binding, and ADMET-related metrics [5,27].

3. PharMistral processing pipeline

In this section, we describe the end-to-end PharMistral processing pipeline. We first outline how three complementary corpora are constructed from public databases: human protein sequences, uncoupled drug-like small molecules, and high-affinity protein–ligand pairs (Section 3.1). We then explain how proteins and SMILES strings are tokenized with the Mistral-7B SentencePiece vocabulary (Section 3.2), followed by Stage I joint unpaired pre-training on mixed protein and small-molecule sequences (Section 3.3) and Stage II sequence-conditioned fine-tuning on paired protein–ligand data (Section 3.4). Finally, we detail how the fine-tuned model is used at inference time to generate candidate ligands for previously unseen protein targets (Section 3.5). An overview of the PharMistral architecture and processing flow is shown in Fig. 1.

3.1. Data acquisition

We prepared three complementary corpora.

- **Proteins.** 318,421 non-redundant, full-length Homo sapiens sequences were obtained from the UniProt release 2024_03 and filtered for retaining only those composed of the 20 standard amino acids to ensure compatibility with subsequent modeling [28].
- **Uncoupled small molecules.** 1,083,562 drug-like canonical SMILES were sampled from the ZINC15 "drug–like" and "clean" subsets after deduplication [29].
- **Protein–ligand couples.** 314,920 activity records (IC_{50}, K_d or $EC_{50} \leq 1$ μM) were extracted from ChEMBL 26 (human), giving a one-to-one mapping between each protein sequence and its most affine ligand [8].

To ensure uniformity, all SMILES strings were preprocessed by removing stereochemical markers (such as `@`, `@@`, `/` and `\`), eliminating chirality and cis/trans information that could introduce unwanted variability. After this preprocessing, the dataset retained 304,139 protein–ligand couples, which were separated into training (90%) and test (10%) sets. The split was defined based on exact protein sequence identity: proteins in the test set were treated as unseen if their amino acid sequences did not exactly match any sequence in the training set. No additional sequence-similarity clustering or homology-aware partitioning methods (e.g., CD-HIT) were employed during the construction of the original split.

3.2. Tokenization

Protein sequences are written with a 20-letter amino-acid alphabet, whereas drug-like SMILES strings employ ∼ 120 ASCII symbols, including multi-character constructs such as "`Cl`" or "`@@`" for chirality. In all our experiments we use the *original* SentencePiece tokenizer shipped with Mistral-7B [30], without any modification or retraining. Although this tokenizer was optimized for natural-language text, it already covers the ASCII range and can therefore be applied directly to both amino-acid sequences and SMILES strings.

In preliminary experiments we trained several domain-specific tokenizers on the union of protein sequences and drug SMILES strings, but these variants did not improve perplexity or downstream generation quality and, in some cases, slightly degraded performance. We therefore keep the default Mistral tokenizer and let the model adapt to the biochemical domain through weight fine-tuning.

Before tokenization, all strings are white-space stripped and stored as UTF-8. Protein sequences and SMILES are then converted to token sequences using the shared Mistral vocabulary. We prepend the literal tags "`Protein:`" and "`Drug:`" and append the built-in end-of-sequence token (`<eos>`) to mark the type and termination of each record.

Using a shared tokenizer for both modalities ensures that protein and molecular tokens live in the same discrete space and can be processed by an unchanged Mistral-7B architecture [31].

3.3. Stage I — Joint unpaired pre-training

Proteins and small molecules are presented to the model *individually* (there is no notion of pairing at this stage). Each record is initiated by a type control tag (`Protein:`, `Drug:`) and terminated with the end-of-string (`<eos>`) tag, with the following patterns:

`Protein:` AA_string `<eos>`

`Drug:` SM_string `<eos>`

where AA_string is an amino-acid string (20 canonical residues) and SM_string is a SMILES. The two data streams are concatenated and randomly permuted in a 1:3 (protein:SMILES) ratio, yielding a corpus of 345,539 sentences (318,421 proteins and 1,083,562 SMILES, deduplicated).

Records are then truncated or padded to 1500 vocabulary tokens (see Section 3.2), with four sequences forming a micro-batch. Therefore, at every optimization step, 96,000 tokens are processed (64 sequences across four GPUs). Training was run on four RTX A6000 (48 GB) GPUs with PyTorch 2.1, native Distributed Data Parallel, `bfloat16` precision and Flash-Attention v2 kernels [32].

Sequences are tokenized with the original Mistral-7B tokenizer (Section 3.2) and passed through the corresponding embedding matrix, which is initialized from the publicly available Mistral-7B checkpoint. The entire embedding matrix and all transformer layers are updated during this pre-training phase. Under this configuration, the tokenized lengths of individual protein or small-molecule records show a right-skewed distribution with a median of 72 tokens, and almost all sequences remain well below our 1500-token limit (Fig. 2a).

For a tokenized sequence $\mathbf{x} = (x_1, \ldots, x_T)$, where T is the (padded) sequence length and each $x_t \in V$ is a discrete token from the Mistral vocabulary V, the model defines a conditional probability distribution $p_\theta(x_t \mid \mathbf{x}_{<t})$ over the next token given all previous tokens $\mathbf{x}_{<t} = (x_1, \ldots, x_{t-1})$. Here, θ denotes all trainable parameters of the transformer (embeddings, attention layers and feedforward blocks). Training follows the standard *autoregressive language modeling* paradigm, in which the model is asked to predict each token in the sequence from its left context. Concretely, we minimize the average negative log-likelihood (cross-entropy) of the true sequence under the model:

$$\mathcal{L}_{\mathrm{CE}} = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta\big(x_t \mid \mathbf{x}_{<t}\big), \tag{1}$$

where log denotes the natural logarithm. Positions corresponding to padding tokens are masked out so that they do not contribute to the loss. Minimizing $\mathcal{L}_{\mathrm{CE}}$ is equivalent to maximizing the likelihood of the training sequences, and its exponential, $\exp(\mathcal{L}_{\mathrm{CE}})$, corresponds to the usual perplexity metric used in language modeling. One epoch ($\approx 5.3\times10^3$ optimizer steps) covers the full corpus once and is sufficient to make the perplexity converge on a held-out validation subset. The resulting checkpoint, which captures a joint biochemical language over proteins and small molecules, is then used to initialize Stage II.

The optimization of $\mathcal{L}_{\mathrm{CE}}$ is carried out with the AdamW optimizer [33], with first- and second-moment coefficients $\beta_{1,2} = [0.90, 0.95]$ and numerical stabilizer $\epsilon = 10^{-8}$. AdamW is a variant of Adam that decouples weight decay from adaptive gradient updates, providing more stable training for large transformers. We use the following additional hyperparameter values: peak learning rate 5×10^{-5}; 30 warm-up steps, during which the learning rate is linearly increased from zero to its peak value; a cosine decay schedule that smoothly anneals the learning rate back to zero over the remainder of training. A weight decay of 0.01 acts as L_2 regularization on the parameters, dropout of 0.10 is applied within the transformer layers to reduce overfitting, and gradient clipping with threshold $\|g\|_2 \leq 1$ prevents exploding gradients by rescaling the update when the gradient norm $\|g\|_2$ exceeds 1.

To further improve the modeling of long-range dependencies, we apply fill-in-the-middle (FIM) augmentation [34] to 50% of the training sequences: for a given sequence, a contiguous internal span is removed and the model is trained to reconstruct the missing middle segment from the surrounding prefix and suffix. This encourages the network to integrate information from distant positions in the sequence, which is particularly important for capturing long-range biochemical motifs. All hyperparameters were chosen by starting from the recommended configuration for Mistral-7B and performing small-scale preliminary experiments, monitoring perplexity on a held-out validation subset.

Finally, it is worth noting that Stage I is not intended to learn protein–ligand interactions directly. Rather, its role is to provide large-scale distributional pretraining, allowing the model to acquire a better

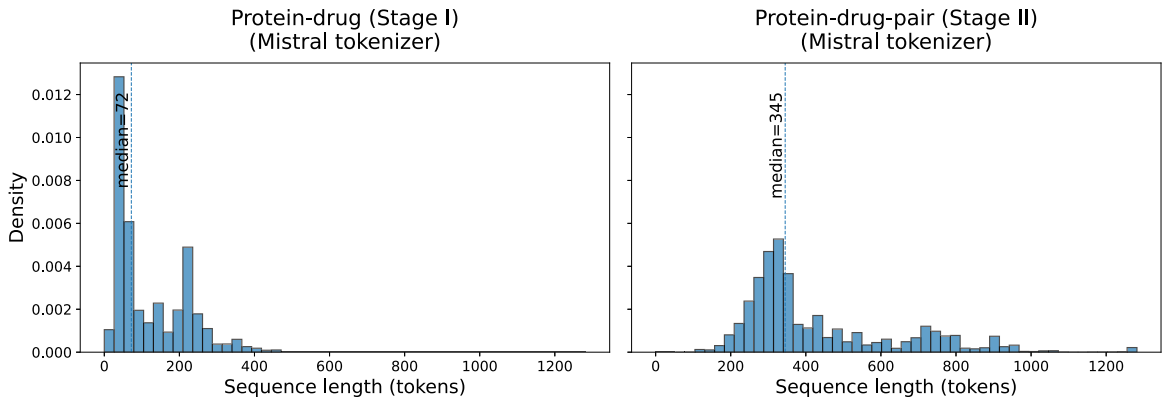


Fig. 2. Token-length distributions under the Mistral tokenizer for (a) individual protein and small-molecule records used in Stage I and (b) protein–ligand pairs used in Stage II. Vertical dashed lines mark the median sequence lengths (72 and 345 tokens, respectively).

representation of the grammar of protein sequences and ligand chemistry. In practice, this initialization improves training stability and supports generalization in the subsequent conditional fine-tuning stage.

3.4. Stage II — Sequence-conditioned fine-tuning

In Stage II, we turn the general biochemical language model obtained in Stage I (Section 3.3) into a conditional generator that maps protein sequences to ligand structures. To this end, we fine-tune the Stage I checkpoint on paired protein–ligand data so that the model learns the conditional distribution of a ligand drug SMILES string given the corresponding protein sequence.

For this phase, we retain the 304,139 high-affinity ($\leq$ 1 μM) protein–ligand pairs from ChEMBL-26 described in Section 3.1. Each pair consists of a protein sequence and a canonical SMILES string. We represent every pair as a *single* token sequence containing both modalities:

`Protein:` AA_string `Drug:` SM_string `<eos>`

where `Protein:` and `Drug:` are literal tags, AA_string is the amino-acid sequence (20 canonical residues), SM_string is the SMILES representation of the ligand, and `<eos>` is the end-of-sequence token from the Mistral tokenizer (Section 3.2).

For Stage II, concatenated protein–ligand sequences exhibit a broader token-length distribution with a median of 345 tokens, which still comfortably fits within the 1500-token budget and leaves ample headroom for longer targets (Fig. 2b).

Let $\mathbf{u} = (u_1, \ldots, u_M)$ denote the tokenized protein prefix, including the `Protein:` and `Drug:` tags, and let $\mathbf{v} = (v_1, \ldots, v_N)$ denote the tokenized ligand SMILES. The full input sequence is then $\mathbf{x} = (\mathbf{u}, \mathbf{v}) = (x_1, \ldots, x_T)$ with $T = M + N$. We reuse the autoregressive loss defined in Eq. (1), but now restrict it to the ligand tokens only, so that the model is trained to predict the SMILES conditioned on the entire protein prefix:

$$\mathcal{L}_{\text{cond}} = -\frac{1}{N} \sum_{t=1}^{N} \log p_\theta\big(v_t \mid \mathbf{u}, \mathbf{v}_{<t}\big), \tag{2}$$

where $\mathbf{v}_{<t} = (v_1, \ldots, v_{t-1})$ and θ collects the model parameters. In practice, this is implemented by computing the same cross-entropy loss as in Stage I (Eq. (1)) and masking out all positions that belong to the protein prefix $\mathbf{u}$. Optimizing $\mathcal{L}_{\text{cond}}$ encourages the network to capture the statistical relationship between the protein sequence and its ligand, thereby approximating the conditional distribution $p_\theta(\mathbf{v} \mid \mathbf{u})$.

Although this objective does not explicitly model binding physics or optimize affinity during training, it can still capture useful statistical regularities from real protein–ligand pairs. In this sense, the model is trained to approximate the conditional distribution of ligands given a protein sequence, which may provide an implicit proxy for protein–ligand compatibility when such relationships are reflected in the data. Accordingly, PharMistral should not be interpreted as a direct model of binding mechanisms, but rather as a method for learning useful target-conditioned signals from large-scale paired data.

The optimization setup closely follows Stage I (Section 3.3): we again use the AdamW optimizer with the same $\beta_{1,2}$ and ϵ values and the same cosine learning rate schedule with linear warm-up. For Stage II, we use the hyperparameter values listed below.

- Optimizer steps: 5000 (one full epoch over the paired corpus).
- Peak learning rate 6×10^{-6}; 30 warm-up steps; cosine decay.
- No FIM augmentation (the full protein–ligand context is always visible).
- Token dropout 10% on the protein prefix: during training, a random 10% of protein tokens are replaced by a special "dropped" token, which makes the model more robust to small changes or noise in the input sequence.
- Effective batch size: 64 sequences per update, with the same packing strategy and maximum length as in Stage I.

These settings were selected by starting from the Stage I configuration and running small-scale experiments, monitoring validation loss on a held-out subset of protein–ligand pairs.

3.5. Inference

At inference time, the fine-tuned model is used purely as a sequence-conditioned generator. Given a new protein sequence `AA_string`, we construct the textual prompt

`Protein:` AA_string `Drug:`

and feed its tokenized version to the model. Then, the transformer autoregressively samples a sequence of tokens for the `Drug:` segment until the end-of-sequence token `<eos>` is generated, which yields a candidate ligand in SMILES format.

This design was deliberately chosen for the present study to evaluate whether a large sequence-conditioned language model can generate ligands for unseen protein targets using only protein sequence information, without relying on structural inputs or additional guidance modules during generation. For this reason, we adopt a standard autoregressive SMILES generation framework, which is simple, scalable, and computationally efficient. In addition, this generation strategy remains relatively exploratory, which is useful in early hit-discovery settings where broad exploration of chemical space is important before later optimization stages. At the same time, the model does not explicitly optimize binding, ADMET, or other properties during decoding.

Table 1
Molecular filtering criteria.

Param	Value	Param	Value
MW	200–500 Da	LogP	[−1, 6]
HBA	0–10	HBD	[0, 5]
QED	0–1	SA	[0, 5]
TPSA	20–130 Å^2	Arom. Rings	≥0
Rot. Bonds	–	Heavy Atoms	–

Key – MW: Molecular Weight; HBA/HBD: H-Bond Acceptor/Donor; SA: Synthetic Accessibility (1=easy, 5=hard); TPSA: Topol. Polar Surface Area; QED: Quant. Est. of Drug-likeness.

These aspects are instead evaluated after generation through docking and downstream predictors.

Token sampling follows a standard stochastic decoding scheme designed to balance diversity and chemical plausibility. Let $p_\theta(\cdot \mid \mathbf{u}, \mathbf{v}_{<t})$ denote the probability distribution over the next token at decoding step t, conditioned on the protein prefix $\mathbf{u}$ and the partial SMILES $\mathbf{v}_{<t}$ (Section 3.4).

Before sampling, the logits are divided by a temperature parameter (equal to 0.8), which sharpens the distribution and encourages selection of higher-probability tokens. We then apply nucleus (top-p) sampling with $p_{\text{top}} = 0.9$ — the vocabulary is restricted to the smallest set of tokens whose cumulative probability mass is at least 0.9 —, this restricted distribution is renormalized, and the next token is drawn from it. Finally, a repetition penalty of 1.1 is applied, which slightly downweights tokens that have already appeared many times in the generated prefix and discourages degenerate loops or trivial repetitions.

Unless otherwise stated, for each protein, we generate 19 independent SMILES candidates by repeating this sampling procedure with different random seeds. Each candidate $\mathbf{v}$ has an associated log-likelihood $\log p_\theta(\mathbf{v} \mid \mathbf{u})$, obtained by summing the token-wise log-probabilities under the model. By default, we select the candidate with the highest internal log-likelihood for downstream analyses, while the full set of samples is used to assess diversity and novelty.

After Stage II, the network operates as a sequence-conditioned generator that can translate previously unseen protein sequences into candidate ligands without any target-specific retraining, enabling the target-agnostic generation experiments on held-out proteins reported in Sections Section 4–5.

4. Evaluation methods

To rigorously assess the performance of our generative framework, we employed a comprehensive evaluation pipeline combining cheminformatics, docking simulation, and pharmacokinetic profiling. Our evaluation aimed to measure not only the chemical validity and novelty of the generated molecules, but also their structural relevance, drug-likeness, synthetic feasibility, and binding potential.

4.1. Evaluation metrics

We categorized evaluation metrics into three main groups reflecting our analysis stages: (i) Validity, Uniqueness, and Novelty, (ii) Structural Deviation and Similarity Analysis, and (iii) Statistical Comparison and Distribution Analysis.

4.1.1. Validity, uniqueness, and novelty

To assess the generative capacity of our model, we evaluated the three core metrics described below.

- **Validity** measures the proportion of syntactically and chemically valid molecules among all generated SMILES strings. A molecule is considered valid if its SMILES representation corresponds to a chemically plausible structure. It is defined as:

$$\text{Validity} = \frac{n(P)}{n(G)}$$

where:

  - $n(P)$: number of valid compounds;
  - $n(G)$: total number of generated compounds.

- **Uniqueness** quantifies the diversity within the valid set by calculating the fraction of distinct molecules. It is expressed as:

$$\text{Uniqueness} = \frac{n(U)}{n(P)}$$

where:

  - $n(U)$: number of unique compounds;
  - $n(P)$: number of valid compounds.

- **Novelty** measures the percentage of valid and unique molecules that do not appear in the paired training set used in Stage II for the supervised protein–ligand task. This metric is intended to assess whether the model generates compounds beyond those seen in the target-conditioned training data. It is defined as:

$$\text{Novelty} = \frac{n(Z)}{n(P)}$$

where

  - $n(Z)$ is the number of novel compounds;
  - $n(P)$ number of valid and unique compounds.

This definition does not imply absolute novelty with respect to the full public chemical space, since molecules may still exist in external databases such as ZINC or PubChem.

These metrics collectively assess the syntactic correctness of the generated SMILES, their internal diversity, and their novelty with respect to the supervised training split.

4.1.2. Structural deviation and similarity analysis

We have evaluated the structural relevance of the generated molecules using the following analyses.

- **Tanimoto Similarity.** We assessed structural similarity using Tanimoto similarity [35] computed from Morgan fingerprints in RDKit (radius 2, 2048 bits). We performed two separate analyses: generated molecules versus the training set, and generated molecules versus the test set. In both cases, for every generated molecule, we retained the maximum Tanimoto similarity to the corresponding reference set. Similarity values range from 0 (completely different) to 1 (identical), and we used a threshold of 0.8 to identify structurally conserved candidates.
- **Root Mean Square Deviation (RMSD).** RMSD was computed using RDKit alignment tools. It measures the average distance between matched atoms in generated and reference molecules after 3D alignment. Lower values indicate that the generated conformation is closer to the reference structure.
- **Bemis–Murcko Scaffold Analysis.** To evaluate scaffold-level novelty, we extracted standard Bemis–Murcko scaffolds from canonical SMILES using RDKit. We then performed two separate scaffold-overlap analyses: generated versus training set molecules, and generated versus test set molecules. For each comparison, we counted the unique scaffolds in the generated set and in the reference set, the number of shared scaffolds, and the number of generated scaffolds not present in the reference set. This analysis was used to evaluate whether the generated molecules reproduced known ligand families or explored scaffold-level novelty.

Table 2
Optimal ranges of ADMET properties.

Property	Criteria/Range
FDAMDD, T1/2, F20%	**Ex:** 0–0.3; **Med:** 0.3–0.7; **Poor:** 0.7–1.0
Caco-2	Acceptable: > -5.15 *(log cm/s)*
NPscore	> 0 *(Natural Product-like)*
BBB Penetration	**High:** BBB+(able to cross); **Low:** BBB-(unable)

Key – Ex: Excellent; Med: Medium; F20%: Bioavailability; T1/2: Half-life; FDAMDD: Toxicity.

4.1.3. Statistical comparison and distribution analysis

To better understand how our generated molecules relate to the training and test sets, we applied several analytical methods, as described below.

- **SMILES Overlap.** To assess exact-molecule redundancy or novelty, we computed the overlap between the generated molecules and the training set and, separately, between the generated molecules and the test set, considering only valid canonical SMILES representations.
- **Wasserstein Distance.** We used this statistical metric to compare the distributions of key molecular properties, such as molecular weight (MW), LogP, and topological polar surface area (TPSA), to quantify how these properties differ between the generated molecules and the molecules belonging to the test set.
- **Kernel Density Estimation (KDE).** We applied KDE to visualize and compare the distribution of predicted binding affinities across generated molecules and the test set. To highlight high-affinity compounds, we emphasized those with binding scores ≤ -8.18 kcal/mol.

4.2. Molecular property analysis and energy optimization

4.2.1. Physicochemical property filtering

To ensure drug-likeness, we applied a set of strict physicochemical filters using the Python library `RDKit` [36]. The properties that were included are the molecular weight, hydrogen bond donors and acceptors, lipophilicity (LogP), synthetic accessibility (SA) [10], drug-likeness (QED) [9], and topological polar surface area (TPSA). The complete list of parameters and corresponding thresholds is summarized in Table 1.

In addition to standard drug-likeness rules (e.g., Lipinski, Ghose, and Veber), we incorporated further structural constraints such as aromaticity, rotatable bond count, and heavy atom count to ensure molecular relevance and chemical feasibility.

Moreover, to eliminate potentially toxic or unstable molecules, we applied additional substructure-based filters:

- **PAINS filter** to remove assay-interfering compounds;
- **BRENK filter** to exclude problematic groups such as nitro and thiol functionalities;
- **NIH filter** to eliminate reactive moieties like aldehydes and alkyl halides.

4.2.2. Molecular optimization and feature encoding

Filtered SMILES strings were converted into three-dimensional structures and geometrically optimized using the UFF and MMFF94 force fields to minimize conformational energy. Following the optimization phase, we computed extended-connectivity fingerprints (ECFPs) using Morgan fingerprints with a radius of 2 and a length of 2048 bits to encode the chemical features of each molecule.

4.2.3. Conformational energy calculation

To evaluate the thermodynamic stability of the generated molecules, we calculated the conformational energy of each optimized 3D structure using RDKit. The energies were expressed in Hartree (Ha), a standard unit in quantum chemistry that allows for consistent comparison between compounds. These energy values were then analyzed to identify low-energy conformations, indicative of stable and synthesizable molecules.

4.3. Molecular docking evaluation

We used molecular docking to evaluate the binding interactions between our designed compounds and the protein targets from the test set. Docking simulations were performed using AutoDock Vina [11], which estimates binding affinity based on electrostatic and van der Waals interactions.

Out of 19 proteins in the test set, 17 had experimentally determined 3D structures available in the Protein Data Bank (resolved by X-ray crystallography, NMR, or cryo-EM) and were selected for docking evaluation. For binding affinity assessment, we considered docking scores ≤ -8.18 kcal/mol as indicative of moderate to strong bioactivity. We then calculated the percentage of generated molecules meeting this threshold and compared their performance against known reference ligands.

We used **Success rate (Succ)** to evaluate the percentage of generated molecules that pass predefined thresholds on desired properties. We define a qualified molecule to have QED ≥ 0.25, SAscore ≥ 0.59 and Vina score ≤ -8.18 kcal/mol. QED and SAscore thresholds are defined as the 10th percentile of approved drugs in DrugCentral [37]. The Vina score threshold corresponds to a binding affinity less than 1 μM, which is a widely used value to guarantee moderate bioactivity in medicinal chemistry [38].

4.4. Toxicity and ADMET evaluation

A compound's ability to bind well to a protein is important, but it is not enough on its own. To be considered a viable drug, a molecule also needs to behave safely and predictably in the body. That is where ADMET properties come in: they cover Absorption, Distribution, Metabolism, Excretion, and Toxicity, all of which are critical for assessing a compound's overall suitability.

We used ADMETlab 2.0 [39], a machine learning-based tool, to predict a range of important characteristics for the molecules generated by our model. The specific properties we evaluated are listed in the following.

- **FDAMDD (FDA Maximum Recommended Daily Dose):** Predicts oral toxicity by ensuring doses remain within safe limits.
- **$T_{1/2}$ (Half-Life):** Indicates the time required for drug concentration to reduce by half; less than 3 h is considered short.
- **F20% (Oral Bioavailability):** Represents the probability that less than 20% of the administered dose reaches systemic circulation.
- **Caco-2 Permeability:** An estimate of intestinal absorption.
- **BBB Penetration:** Predicts a compound's ability to cross the blood–brain barrier.
- **NPscore:** A measure of similarity to natural products.

Table 2 reports practical ADMET thresholds used for in silico screening, reflecting established best practices for oral drug discovery and enabling early assessment of drug-likeness and development risk.

We applied these criteria to evaluate the ADMET profiles of the most promising compounds generated by our model, focusing on four representative protein targets: 3ZCW, 4A69, 1HY7, and 3D59.[1] This

Table 3
Affinity analysis: PharMistral (PM) vs. DeepTarget (DT).

Protein	Mean		SD		Top-N		Random-N		Best	
	PM	*DT*	*PM*	*DT*	*PM*	*DT*	*PM*	*DT*	*PM*	*DT*
7KK4	−8.54	−7.99	1.17	0.82	−10.22	−9.21	−8.52	−8.01	−12.72	−11.93
6CM4	−6.86	−8.40	0.81	0.96	−7.44	−9.73	−6.90	−8.40	−9.13	−11.46
7DUA	−9.92	−8.83	1.02	0.87	−10.05	−10.72	−9.94	−8.85	−12.46	−12.22
8V3O	−6.99	−6.58	0.72	0.60	−7.15	−8.10	−7.02	−6.57	−8.81	−9.11

step allowed us to assess not only how effective these molecules might be, but also their potential safety, before moving on to more advanced filtering or validation stages.

4.5. Evaluating PharMistral in comparison to DeepTarget

To evaluate the performance of our model, PharMistral, we performed a direct comparison with a recent method, DeepTarget. To ensure the fairness of this evaluation, we searched the training datasets of PharMistral and DeepTarget to identify proteins present in the training set of one model but absent in the other. Our analysis of the two datasets yielded the following results:

- We identified **only** two proteins, 6CM4 and 7KK4,[2] that were present in DeepTarget's training set but not included in PharMistral's. We then used these two proteins to test PharMistral's performance on data unknown to it.
- Conversely, we found **only** two proteins, 7DUA and 8V3O,[3] present in PharMistral's test set but not included in DeepTarget's training set. We used these two proteins to test DeepTarget's performance on data unknown to it.

Each model was tasked with generating 1000 candidate molecules per protein. All generated molecules were subjected to the same post-processing pipeline, including physicochemical filtering and molecular docking, so that both models were evaluated under identical and transparent conditions.

In particular, we compared how well PharMistral and DeepTarget performed in terms of docking affinity (in kcal/mol) across the four protein targets 6CM4, 7DUA, 7KK4 and 8V3O. For each protein, we calculated the average and standard deviation of the docking scores, highlighting the best-performing molecules, based on randomly sampled subsets to verify the robustness of the results. After filtering, the number of valid molecules available for the analysis varied across proteins: 500 for 6CM4 and 7KK4, 80 for 7DUA, and 40 for 8V3O. Detailed results are reported in Table 3.

Finally, to evaluate the quality of the molecules generated by PharMistral and DeepTarget, we compared them to real, approved drugs using three key criteria: binding affinity (how strongly a molecule interacts with its target protein), drug-likeness measured by QED (Quantitative Estimate of Drug-likeness), and synthetic accessibility (SA), which reflects how easily a compound can be synthesized in a laboratory setting. Tables 4–7 present the detailed metrics across all target proteins.

Table 4
Binding affinity, QED, and SA scores for **6CM4**.

Model	Binding affinity (kcal/mol)	QED ↑	SA ↓
DeepTarget	−11.464	**0.699**	0.770
Pharmi-stral	−9.125	0.680	**0.511**
Real Drug	**−12.200**	0.657	0.647

Table 5
Binding affinity, QED, and SA scores for **7KK4**.

Model	Binding affinity (kcal/mol)	QED ↑	SA ↓
DeepTarget	−11.929	0.602	0.790
PharMistral	−12.720	0.667	**0.596**
Real Drug	**−13.900**	**0.683**	0.758

Table 6
Binding affinity, QED, and SA scores for **7DUA**.

Model	Binding affinity (kcal/mol)	QED ↑	SA ↓
DeepTarget	−12.218	**0.759**	**0.457**
PharMistral	−12.463	0.564	0.714
Real Drug	**−13.213**	0.413	0.704

Table 7
Binding affinity, QED, and SA scores for **8V3O**.

Model	Binding affinity (kcal/mol)	QED ↑	SA ↓
DeepTarget	**−9.107**	**0.731**	0.589
PharMistral	−8.8005	0.511	0.764
Real Drug	−9.035	0.429	**0.568**

5. Results

In this section, we present the key findings from our experimental evaluation. The results cover several aspects of model performance, including the quality of the generated molecules, their structural similarity to known compounds, physicochemical properties, docking performance and ADMET predictions. We also provide a direct comparison with a baseline generative model to highlight the strengths and limitations of our approach.

5.1. Performance of the generative model

Out of a total of 10,782 SMILES strings generated by our model, 10,733 were found to be chemically valid, resulting in an impressive validity rate of 99.55%. Among these, 6140 molecules (57.21%) were unique, while the remaining 4593 (42.79%) were duplicates. Finally, 2189 of the unique molecules (35.65%) were novel, meaning they did not appear in the training set. These results indicate that the model is capable of generating chemically valid structures, also showing substantial diversity and novelty relative to the training set.

To better evaluate structural novelty, we computed the maximum Tanimoto similarity between each generated molecule and the

[1] The reported four-character code corresponds to the protein accession number in the PDB biobank. Human kinesin 3ZCW is a target for drug development in cancer chemotherapy with compounds in phase II clinical trials. Protein 4A69 regulates gene expression by removing acetyl groups from lysine residues in histone tails, resulting in chromatin condensation, while 1HY7 is a carboxylic acid-based inhibitor in complex with Matrix metalloproteinases 3, a biological marker for inflammatory osteoarthritis. Finally, 3D59 is a human plasma platelet-activating factor, which acts as an anti-inflammatory.

[2] 6CM4 is dopamine receptor that participates in processes like reward, movement, and mood and is a target for drugs treating neurological and psychiatric conditions such as schizophrenia and Parkinson's disease. Poly-ADP-ribosyltransferase (7KK4) plays a critical role in DNA repair and cell death.

[3] 7DUA is a human kinase fragment related to the phosphorylation of proteins, specifically tyrosine residues, which is crucial for its biological role in the body cellular processes and signaling pathways. Protein 8V3O is involved in the cellular antiviral defense response.

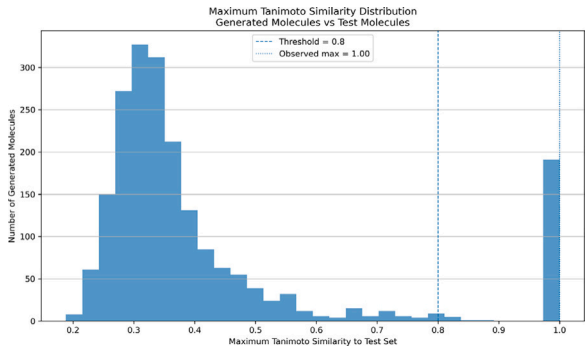


Fig. 3. Distribution of maximum Tanimoto similarity values between generated molecules and the test set. Only 9.89% of generated molecules reached a maximum similarity ≥ 0.8, while most remained below this threshold.

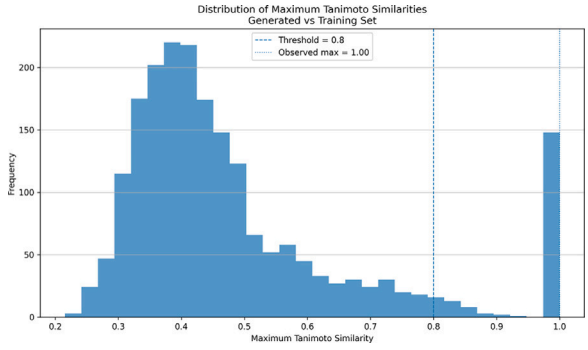


Fig. 4. Distribution of maximum Tanimoto similarity values between generated molecules and the training set. Similarity is slightly higher than for the test set, but only 9.10% of generated molecules reached a maximum similarity ≥ 0.8.

molecules in the test set and, separately, between each generated molecule and the molecules in the training set, using Morgan fingerprints. As shown in Fig. 3, the generated molecules had an average maximum similarity of 0.4176 to the test set (median 0.3393), and only 202 out of 2043 molecules (9.89%) reached a similarity of at least 0.8. When compared with the training set (Fig. 4), the average maximum similarity was 0.4881 (median 0.4262), with 186 out of 2043 molecules (9.10%) exceeding the same threshold. In both cases, about 90% of the generated molecules remained below 0.8, indicating that most generated compounds are not close fingerprint-level analogues of either the training or the test ligands.

We then examined scaffold-level overlap using standard Bemis–Murcko scaffolds. The generated set contained 2605 unique scaffolds, compared with 2205 in the test set and 62994 in the much larger training set. Compared with the test set (Fig. 5), only 352 scaffolds were shared, while 2253 generated scaffolds (86.49%) were not present in the test set. Compared with the training set (Fig. 6), 1980 scaffolds were shared, whereas 625 generated scaffolds (23.99%) were not found in the training data.

These findings demonstrate that, although the model frequently reproduces scaffolds from training-set families, it also has the ability to generate genuinely novel scaffolds absent from the training data.

We also computed RMSD after an RDKit-based 3D alignment for structurally matched molecule pairs. The low RMSD values indicate close geometric agreement in these matched cases, providing a complementary 3D structural perspective to the fingerprint and scaffold analyses.

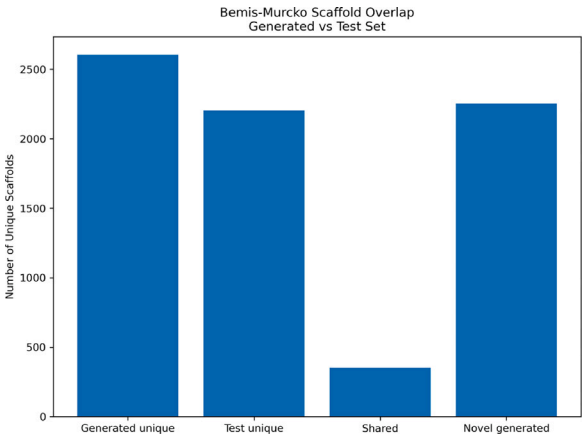


Fig. 5. Bemis–Murcko scaffold overlap between generated molecules and the test set. Only 352 scaffolds were shared, while most generated scaffolds were absent from the test set.

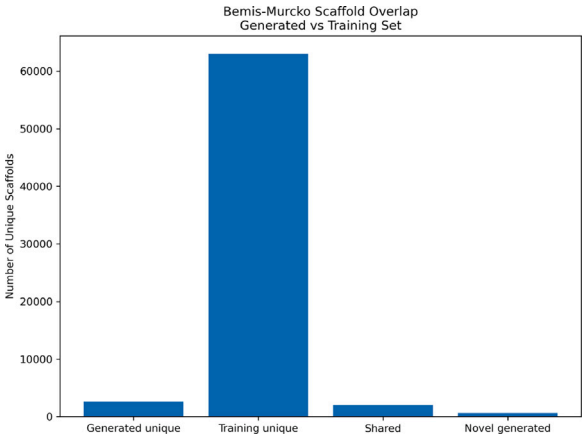


Fig. 6. Bemis–Murcko scaffold overlap between generated molecules and the training set. A large fraction of generated scaffolds was shared with the training data, although 625 scaffolds were not present in it.

5.2. Approved drugs rediscovered by the model

5.2.1. Generation of approved drugs not present in the training set

A detailed comparison between the generated molecules and the reference datasets showed that **752 unique compounds** were shared between the generated and test sets. Among these, **only four molecules (0.53%)** were not part of the training data but matched real, approved drugs listed in ChEMBL: **CHEMBL111000**, **CHEMBL444341**, **CHEMBL470831**, and **CHEMBL1210208** (Fig. 7). Although this represents a small fraction of the generated set, it is an important finding. It shows that the model was able to *rediscover real, pharmacologically validated drugs* without ever seeing them during training. This suggests that the model generative process goes beyond simple pattern memorization and can capture *chemical and biological features* typical of real drugs.

5.2.2. Protein–drug relationships

A closer look revealed that these four approved drugs are linked to specific protein structures already included in the test set, suggesting that the model correctly understood *functional relationships between proteins and ligands*. In particular, **CHEMBL111000** and **CHEMBL444341** were associated with **PDB ID 7E32**, **CHEMBL470831** with **7WC8**, and **CHEMBL1210208** with **1HY7**. This consistent matching between generated molecules and their corresponding proteins demonstrates

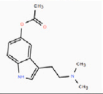
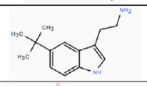
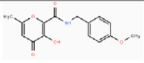
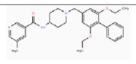
Parent_Chemblid	SMILES	Structure	Protein_PDBid
CHEMBL111000	COC1CCC(CNC(=O)C2OC(C)CC(=O)C2O)CC1		7E32
CHEMBL444341	CC(C)(C)C1CCC2[NH]CC(CCN)C2C1		7E32
CHEMBL470831	CC(=O)OC1CCC2[NH]CC(CCN(C)C)C2C1		7WC8
CHEMBL1210208	CCOC1CC(CN2CCC(NC(=O)C3CNCC(C)C3)CC2)CC(OCC)C1-C1CCCCC1		1HY7

Fig. 7. Approved drugs rediscovered by the model and their corresponding protein targets.

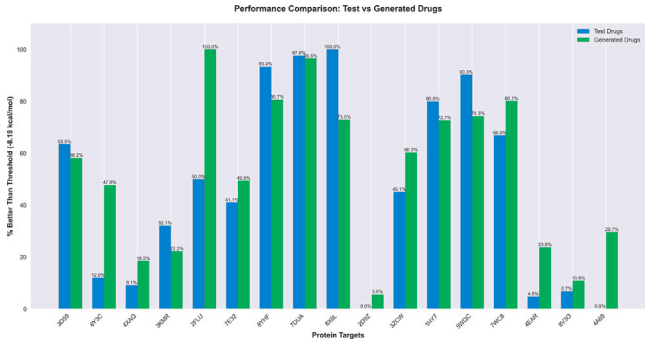


Fig. 8. Percentage of compounds with a binding affinity ≤ −8.18 kcal/mol for each protein target. Green bars represent generated molecules; blue bars represent test/reference drugs.

that the model can *connect chemical structure to biological context*, a key requirement for rational drug design. Overall, these results confirm that the model not only generates valid chemical structures but also learns meaningful protein–drug relationships, enabling the rediscovery of existing therapeutics through a generative approach.

5.3. Molecular property and energy profile

After physicochemical and toxicity filtering, 3383 unique molecules (55.1% of valid SMILES) were retained, exhibiting improved drug-likeness and reduced structural risk. Subsequent 3D optimization and fingerprint-based analysis identified 2043 stable and structurally diverse compounds (33% of valid SMILES). Conformational energy analysis showed that most molecules lie near local energy minima, indicating thermodynamically stable and energetically realistic structures suitable for downstream screening and synthesis.

5.4. Molecular docking performance

Docking simulations on 17 protein targets showed that a large portion of our generated molecules achieved strong binding scores, with many being less than the binding affinity threshold (≤ −8.18 kcal/mol). In most cases, the percentage of high-affinity binders among generated molecules was greater than the molecules in the test set, demonstrating the model's ability to produce promising drug candidates. The above described results are shown in Tables 8 and 9, and in Fig. 8.

To evaluate how well our generated molecules resemble real drugs, we compared their key physicochemical properties using the Wasserstein distance. This analysis, conducted across all generated compounds, focused on ten key descriptors, such as molecular weight (MW), LogP, TPSA, QED, and others. As an illustrative example, we present some results for the 3D59 protein target.

Table 8
Molecules with binding affinity ≤ −8.18 kcal/mol.

PDB	Test ≤ *T*		Gen ≤ *T*		PDB	Test ≤ *T*		Gen ≤ *T*	
	(N)	*(%)*	*(N)*	*(%)*		*(N)*	*(%)*	*(N)*	*(%)*
3D59	47	63.5	53	58.2	2D92	0	0.0	1	5.6
6Y3C	85	12.0	129	47.8	3ZCW	115	45.1	76	60.3
4XAO	24	9.1	20	18.5	1HY7	12	80.0	40	72.7
3KMR	9	32.1	6	22.2	5WQC	65	90.3	55	74.3
2FLU	1	50.0	9	100	7WC8	555	66.9	366	80.1
7E32	85	41.1	91	49.5	4EAR	3	4.8	10	23.8
8YHF	310	93.4	163	80.7	8V3O	11	6.8	5	10.9
7DUA	166	97.7	83	96.5	4A69	0	0.0	57	29.7
8X8L	9	100	27	73.0					

Table 9
Performance against test set maximum affinity.

PDB	Best	Gen > Test		PDB	Best	Gen > Test	
	(min)	*(N)*	*(%)*		*(min)*	*(N)*	*(%)*
3D59	−9.48	6	6.59	2D92	−7.63	2	11.11
6Y3C	−9.67	25	9.26	3ZCW	−10.16	5	3.97
4XAQ	−9.56	1	0.93	1HY7	−10.24	8	14.55
3KMR	−13.78	0	0.00	5WQC	−11.46	0	0.00
2FLU	−11.98	0	0.00	7WC8	−12.55	1	0.22
7E32	−11.46	0	0.00	4EAR	−9.41	0	0.00
8YHF	−12.08	0	0.00	8V3O	−9.04	0	0.00
7DUA	−13.21	0	0.00	4A69	−8.05	67	34.90
8X8L	−9.97	4	10.81				

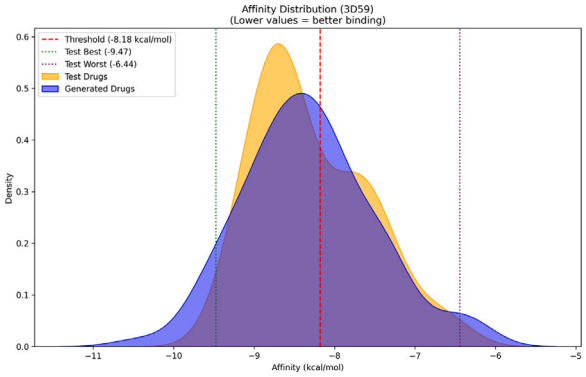


Fig. 9. Binding affinity distribution for protein 3D59. The plot compares generated and test set compounds using KDE.

Table 10
Wasserstein distance of chemical properties for protein 3D59.

Property	Wasserstein distance
MW	7.942
HBA	0.426
HBD	0.720
LogP	0.488
QED	0.053
SA	0.038
TPSA	5.538
Aromatic Rings	0.460
Rotatable Bonds	0.787
Heavy Atoms	0.525

As shown in Table 10, the generated molecules closely match the reference compounds in most properties, with particularly low distances for QED and SA. Slight deviations were observed in MW and TPSA, suggesting potential areas for refinement.

These results are also supported by Wasserstein distance plots (see Fig. 10), which show that the generated molecules and known ligands

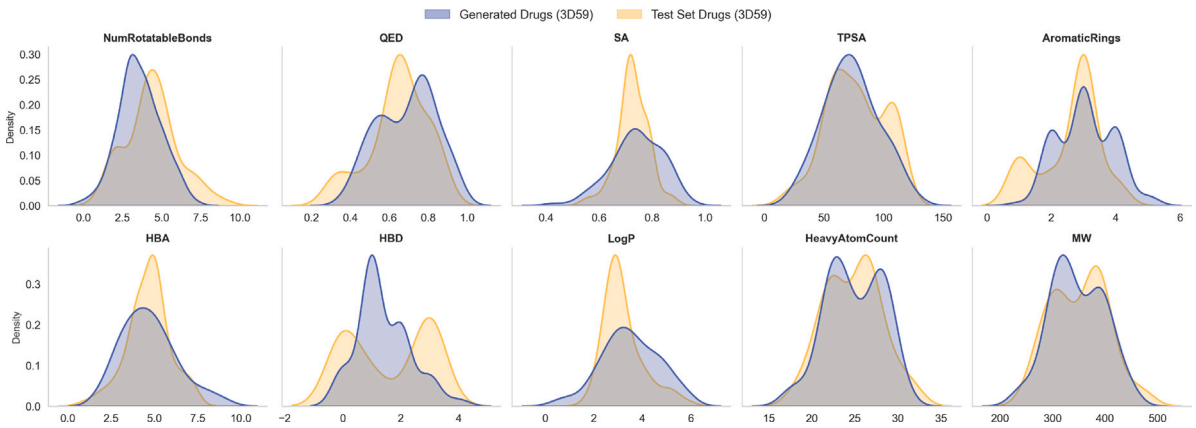


Fig. 10. Wasserstein distance plots comparing key physicochemical property distributions between generated and test set compounds for the 3D59 protein target. Properties include QED, SA, TPSA, HBA, HBD, NumRotatableBonds, AromaticRings, and HeavyAtomCount.

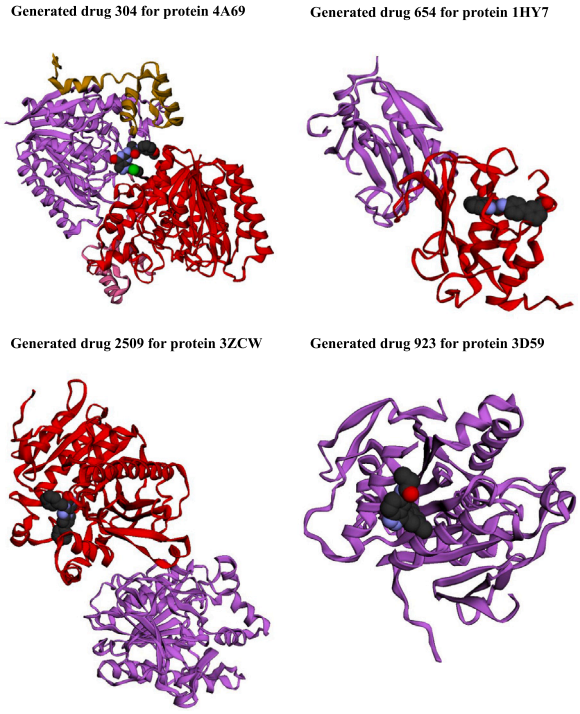


Fig. 11. Docked conformations of promising AI-generated compounds bound to their respective target proteins: HDAC3 (304), KIF11 (2509), MMP3 (654), and PAF-AH (923). Binding poses indicate strong protein–ligand interactions.

have very similar property distributions. This suggests that our generated compounds are chemically realistic and comparable to real drug molecules.

Fig. 9 illustrates the binding affinity distribution for protein 3D59. Using kernel density estimation (KDE), we compared the predicted binding affinities of generated and test set compounds. While both distributions largely overlap, the generated molecules tend to show slightly weaker affinities on average, though several are still less than −8.18 kcal/mol.

These results confirm that the generated molecules are chemically realistic and often biologically promising, reinforcing the model's ability to produce viable drug-like candidates.

Beyond overall performance, several AI-generated compounds showed stronger binding affinities than their ChEMBL reference ligands. Indeed, improved molecules were found for 1HY7 (7 molecules), 2D9Z (1),[4] 3ZCW (3), 3D59 (5), 4A69 (49), 4XAQ (1),[5] 6Y3C (22),[6] 7WC8 (1),[7] and 8X8L (4).[8]

To further validate their potential, all top-performing molecules were subjected to ADMET analysis, which confirmed their suitability as promising drug candidates. A complete summary of these findings is provided in Table 11.

5.5. ADMET and toxicity results

The ADMET analysis revealed a mixed picture across different protein targets. For 4A69, the generated molecule showed excellent safety, with a low FDAMDD score of 0.19, compared to 0.682 for the reference drug, placing the latter in a medium/high risk category. In contrast, compounds generated for 3ZCW and 3D59 showed higher predicted toxicity levels, suggesting these candidates may require further refinement. Predictions for oral bioavailability (F20%) varied. While the reference drugs generally performed well, the generated compounds showed more diverse outcomes. The molecule for 3ZCW had an acceptable score of 0.07, whereas the 3D59 candidate scored poorly (0.943), likely due to its high molecular weight or polarity. In terms of half-life ($T_{1/2}$), reference drugs consistently showed shorter durations (all under 0.3 h), while generated compounds fell into the medium range (0.4–0.6 h), indicating there is room for pharmacokinetic optimization. When assessing absorption through Caco-2 permeability, results were favorable for the generated compounds targeting 3ZCW and 3D59, suggesting good intestinal absorption. However, molecules for 4A69 and 1HY7 performed poorly on this metric, potentially limiting their effectiveness when taken orally.

Interestingly, the molecules generated for 3ZCW were the only ones predicted to cross the blood–brain barrier (BBB+), suggesting potential suitability for central nervous system (CNS) applications. All other generated compounds, as well as their reference ligands, were classified as BBB–.

[4] 2D9Z is a signaling protein involved in converting diacylglycerol signals into prolonged physiological effects. It also plays a role in resistance to oxidative stress.

[5] 4XAQ modulates synaptic transmission by inhibiting presynaptic calcium channels, which reduces glutamate release and dampens excitatory signaling in the brain.

[6] 6Y3C plays a vital role in physiological processes like gastrointestinal cytoprotection, hemostasis, and renal hemodynamics.

[7] 7WC8 is involved in neurotransmission, influencing neural activity, mood, cognition, and perception.

[8] 8X8L provides a three-dimensional model that helps elucidate the receptor's activation mechanism for designing new drugs for neuroendocrine disorders.

Table 11
Generated molecules with improved binding affinity compared to the test set.

Protein	Test best affinity (kcal/mol)	Gen best affinity (kcal/mol)	Num Gen > Test
1HY7	−10.24	−11.45	7
2D9Z	−7.63	−8.28	1
3ZCW	−10.16	−10.79	3
3D59	−9.48	−10.43	5
4A69	−8.05	−9.97	49
4XAQ	−9.56	−9.76	1
6Y3C	−9.67	−10.80	22
8X8L	−9.97	−10.41	4
7WC8	−12.552	−12.694	1

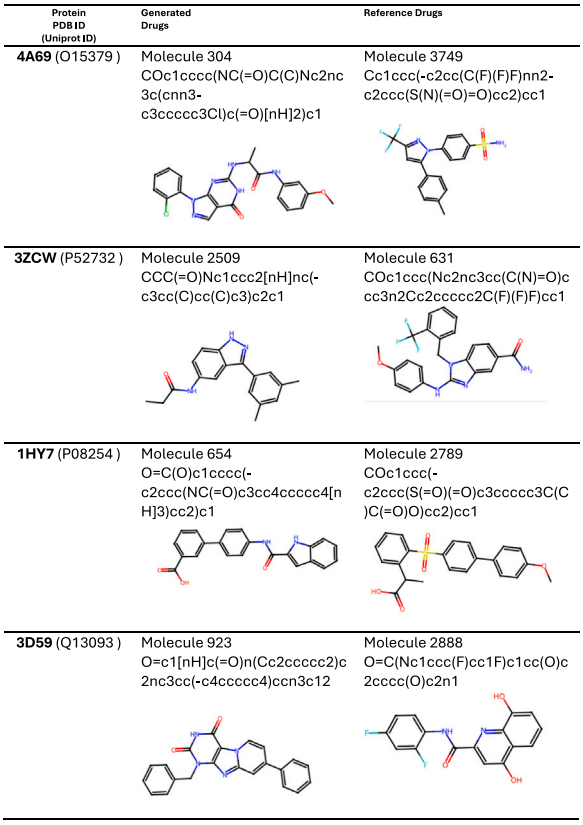


Fig. 12. Visual comparison of generated vs. reference compounds. Each row corresponds to a protein target, showing the generated molecule, its SMILES string, and its 2D structure, alongside the respective reference drug.

Finally, as for drug-likeness (measured by NPscore), reference drugs generally performed better. However, the generated molecule for 3D59 outperformed its reference, suggesting that the model is capable of producing compounds with strong natural product-like features, a valuable trait in drug development.

A full comparison of reference drugs and generated molecules based on ADMET properties is reported in Table 12.

5.6. Case studies: Promising AI-generated compounds compared to reference drugs

As part of our analysis, we identified several AI-generated molecules that performed better than known reference drugs in both binding affinity and drug-likeness. These compounds were designed to target human proteins that play important roles in different therapeutic areas, such as cancer, inflammation, cardiovascular diseases, and neurodegenerative disorders.

In this study, we focused on four representative protein targets: **HDAC3** (Histone deacetylase 3, PDB ID: 4A69, UniProt: O15379), **KIF11** (Kinesin-like protein 11, PDB ID: 3ZCW, UniProt: P52732), **MMP3** (Matrix metalloproteinase 3, PDB ID: 1HY7, UniProt: P08254), and **PAF-AH** (Platelet-activating factor acetylhydrolase, PDB ID: 3D59, UniProt: Q13093). These targets were selected as case studies to demonstrate the potential of our generative approach across diverse biological functions and disease contexts.

Fig. 11 presents four standout generated compounds that demonstrated better docking scores and more favorable physicochemical properties than their corresponding reference drugs. In fact, these improvements span several key drug-like features, including lower molecular weight, higher QED scores (indicating stronger drug-likeness), better lipophilicity, and improved synthetic accessibility. All of these factors contribute to the molecular viability as a drug candidate. A detailed quantitative comparison of these attributes is provided in Table 13, whereas, for a more complete picture, additional details such as 2D structures, SMILES representations, and side-by-side property comparisons are shown in Fig. 12.

In the following, we will briefly outline the characteristics of the best generated molecules (see Fig. 12) for proteins HDAC3, KIF11, MMP3, and PAF-AH, evidencing their possible advantages with respect to known drugs for the same protein target.

- **Molecule 304** is a **potent HDAC3 Inhibitor** for **Epigenetic Modulation**. This compound shows a 16 % improvement in binding affinity compared to Celecoxib and also offers better solubility. While Celecoxib is an approved drug, its clinical use is limited by known cardiovascular side effects. In contrast, Molecule 304 emerges as a promising alternative for targeting HDAC3 in epigenetic therapies. However, to fully realize its potential in a clinical setting, further optimization of its pharmacokinetic properties will be necessary.
- **Molecule 2509** is a **Selective KIF11 Inhibitor** for **Cancer Therapy**. This compound outperforms the reference molecule 631 by showing stronger binding affinity, a high drug-likeness score (QED = 0.76), and a simpler molecular structure that is easier to synthesize. Together, these characteristics make it a compelling candidate for targeted cancer therapies that focus on KIF11 inhibition, a pathway known to play a critical role in tumor cell division.
- **Molecule 654** is an **Optimized MMP3 Inhibitor** for **Inflammatory Diseases**. This molecule exceeds its reference in several key aspects, including binding strength, lower molecular weight, and improved lipophilicity. These properties suggest strong potential for use in treating inflammatory conditions. However, additional optimization will be important to improve its target selectivity and reduce the risk of off-target effects, ensuring both efficacy and safety.
- **Molecule 923** is a **High-Affinity PAF-AH Inhibitor** for **Cardiovascular and Immune Modulation**. This molecule shows enhanced binding affinity, driven by hydrogen bonding and π–stacking interactions, and outperforms the reference drug in both binding strength and drug-likeness. While there is still room for improvement in terms of solubility and formulation, its overall profile makes it a promising candidate for treating cardiovascular and immune-related disorders.

Table 12
ADMET Comparison of generated vs. reference drugs across protein targets.

Property	3ZCW (Gen 2509 vs. Ref 631)	4A69 (Gen 304 vs. Ref 3749)	1HY7 (Gen 654 vs. Ref 2789)	3D59 (Gen 923 vs. Ref 2888)
FDAMDD (Toxicity)	Gen: High risk (0.727) Ref: Medium (0.658)	Gen: Excellent (0.19) Ref: Medium (0.682)	Both: Medium risk (∼ 0.54)	Both: High risk (> 0.82)
F20% (Bioavailability)	Gen: Medium (0.07) Ref: Excellent (0.014)	Both: Excellent (0.002)	Both: Excellent (< 0.01)	Gen: Poor (0.943) Ref: Excellent (0.002)
T1/2 (Half-life)	Gen: Medium (0.427) Ref: Excellent (0.04)	Gen: Medium (0.46) Ref: Excellent (0.029)	Gen: Medium (0.585) Ref: Excellent (0.143)	Both: Medium (∼ 0.45)
Caco-2 (Permeability)	Gen: Optimal (−4.995) Ref: Poor (−5.166)	Gen: Poor (−5.468) Ref: Optimal (−4.767)	Gen: Poor (−5.165) Ref: Optimal (−4.947)	Both: Optimal (∼ −4.91)
BBB Penetration	Gen: BBB+ (0.553) Ref: BBB- (−0.248)	Gen: BBB- (0.028) Ref: BBB+ (0.586)	Both: BBB- (∼ 0.21)	Both: BBB- (< 0.08)
NPscore	Ref: More NP-like (−1.6) vs. (−1.661)	Ref: More NP-like (−1.566) vs. (−1.982)	Ref: More NP-like (−0.643) vs. (−0.986)	Gen: More NP-like (−1.192) vs. (−1.316)

Table 13
Comparison of generated (Gen.) and reference (Ref.) ligands.

Target	Ligands *(G/R)*	MW *(Da)*	HBA/HBD *(G/R)*	LogP	QED	SA	TPSA *(Å²)*	Arom. *Rings*	Rot. *Bonds*
HDAC3	304/3749	438.9/**381.4**	7/3/**4**/**1**	**3.21**/3.51	0.43/**0.75**	0.65/**0.85**	113.9/**78.0**	4/**3**	6/**3**
KIF11	2509/631	**293.4**/440.4	**2**/**2**/5/2	**4.20**/4.95	**0.76**/0.44	**0.81**/0.75	**57.8**/82.2	**3**/4	**3**/6
MMP3	654/2789	**356.4**/396.5	2/3/**4**/**1**	4.79/**4.38**	0.49/**0.67**	**0.80**/0.69	–	–	–
PAF-AH	923/2888	368.4/**316.3**	5/1/**4**/**3**	**3.05**/3.18	0.53/**0.68**	0.65/**0.74**	**72.2**/82.5	5/**3**	3/**2**

Note: Values shown as Generated/Reference. Bold indicates optimal drug-likeness.

5.7. Comparative analysis with DeepTarget

When generating 10,000 molecules per protein, PharMistral consistently produced a higher percentage of valid molecules compared to DeepTarget:

- **6CM4**: PharMistral **99.90%** vs DeepTarget 74.39%;
- **7DUA**: PharMistral **99.55%** vs DeepTarget 73.10%;
- **7KK4**: PharMistral **99.14%** vs DeepTarget 72.76%;
- **8V3O**: PharMistral **99.55%** vs DeepTarget 74.01%.

However, to better understand how the two models perform, we compared them across three essential metrics: binding affinity, drug-likeness (measured by QED), and synthetic accessibility (SA). The results for each of the four target proteins are summarized below.

- **6CM4**: DeepTarget showed slightly better affinity, but PharMistral maintained a competitive QED **(0.680)**. Its SA score (0.511) suggested higher synthetic complexity, which may still be acceptable depending on the application.
- **7DUA**: PharMistral outperformed DeepTarget in binding affinity (**(−12.46)** vs. (−12.21) kcal/mol) and had a strong SA score of **0.714**, though its QED was slightly lower. Overall, PharMistral generated more synthetically feasible candidates.
- **7KK4**: PharMistral achieved the best binding affinity among generative models (**(−12.72)** kcal/mol), close to the reference drug (−13.90), and recorded the highest QED **(0.667)**, indicating excellent drug-likeness.
- **8V3O**: DeepTarget outperformed PharMistral, achieving stronger binding affinity (−9.107 vs. −8.8005 kcal/mol), higher QED, and better SA, thus generating more drug-like and accessible compounds.

Binding affinity comparisons (Table 3) confirm that PharMistral matched or outperformed DeepTarget on most targets, particularly for 7KK4 and 7DUA. QED and SA scores (Tables 4–7) also show that PharMistral frequently achieved a favorable balance between potency, drug-likeness, and synthetic feasibility.

Overall, PharMistral delivered strong and consistent results, particularly in generating chemically valid molecules and achieving high binding affinities across several important targets. Its ability to strike a balance between potency and chemical plausibility highlights its potential as a powerful tool for generative drug discovery.

6. Conclusion and future work

In this work, we have introduced *PharMistral*, a sequence-conditioned ligand generator based on Mistral-7B, providing a reproducible pipeline for drug generation. Our contributions are twofold: (i) a two-stage training strategy combining unpaired pre-training with end-to-end fine-tuning on protein–ligand pairs, and (ii) a comprehensive evaluation framework benchmarking cheminformatics, docking and ADMET profiles against DeepTarget. All checkpoints and scripts are publicly available.

Addressing our central question, the results demonstrate that PharMistral can generate valid, novel, and drug-like ligands for held-out targets without relying on structural data. It achieves near-perfect validity and substantial novelty relative to the training set, yielding candidates that often outperform known ligands in affinity and synthetic accessibility.

Our findings regarding specific research questions are summarized below.

RQ1 — Chemical quality and diversity. On held-out targets, PharMistral generated 10 782 SMILES with 99.5% validity. Of the valid set, 57% were unique and 36% novel molecules, suggesting that the model is not simply reproducing known compounds from the training set. After physicochemical filtering, 2043 unique molecules retained stable low-energy 3D conformations. Tanimoto similarity and Bemis–Murcko scaffold analyses further indicate that most generated molecules are not highly similar to either training or test ligands, although some share scaffold families known to be bioactive. In this study, however, novelty is defined relative to the training set used for the supervised protein–ligand task and should not be interpreted as absolute novelty with respect to the broader public chemical space.

RQ2 — Target relevance. Docking experiments on 17 unseen structures revealed that generated molecules frequently exceed the

−8.18 kcal/mol threshold. For 9 proteins, candidates surpassed the best ChEMBL ligands by up to 1.9 kcal/mol. Structural analysis of top candidates (e.g., for HDAC3, KIF11, MMP3, PAF-AH) highlighted well-positioned hydrogen bonds and π–π interactions, suggesting that the generated molecules can recover chemically plausible binding patterns. However, these results must be interpreted cautiously, since docking scores are only an imperfect proxy for binding, and stronger experimental or computational validation is needed to support any firm conclusions.

RQ3 — Target generalization. PharMistral demonstrated robust generalization on asymmetric training splits. It maintained high validity and competitive affinity on DeepTarget-exclusive proteins (6CM4, 7KK4) and outperformed DeepTarget on PharMistral-exclusive sets (7DUA, 8V3O). The rediscovery of four approved drugs, not seen during training, further supports the idea that the model captures transferable target-related signals beyond direct memorization of training samples.

RQ4 — Drug-likeness and ADMET. High-affinity candidates largely satisfied standard drug-likeness constraints, with QED and SA scores aligning with approved drugs (confirmed via Wasserstein distance). ADMETlab 2.0 profiling indicated favorable bioavailability and toxicity profiles for many candidates, though some require pharmacokinetic optimization.

Overall, PharMistral stands as a scalable, structure-free generator. Unlike SMILES-only models, it explicitly conditions on targets; unlike smaller task-specific models (DeepTarget), it leverages a 7B-parameter LLM for superior validity; and unlike structure-based methods, it operates without pocket information. In the present study, this design choice was intended to evaluate how far a sequence-conditioned autoregressive language model can go using only protein sequence information, without additional structural constraints or guidance modules during decoding.

Limitations and future work

This study has several important limitations. First, in our setting, ‘‘unseen’’ refers only to protein sequences that are not explicitly present in the training set. It does not imply that test proteins are entirely unrelated to training proteins at the homology level, as no homology-aware splitting strategy was employed. Second, the notion of novelty adopted here is defined relative to the training set. A generated molecule may be absent from the training data yet still be present in larger public databases such as ZINC or PubChem. Accordingly, our analysis should be interpreted as reflecting the model’s generalization capability rather than absolute chemical novelty. Third, PharMistral adopts a conventional autoregressive SMILES generation framework and does not incorporate explicit optimization of binding affinity, ADMET, or related properties during generation. These properties are instead evaluated post hoc via docking and downstream predictive models. Moreover, the training objective does not explicitly encode binding physics, but rather learns statistical regularities from protein–ligand pairs, which may function as an implicit proxy for compatibility while permitting a wider exploration of chemical space. Finally, although docking provides useful evidence for target relevance, it remains an approximate metric, and stronger wet-lab controls would further strengthen the evaluation.

Future work will focus on improving both the model and the experimental framework. In particular, we plan to explore guided generation strategies, such as reinforcement learning, classifier-guided decoding, or contrastive objectives, while carefully considering the risk of biases introduced by noisy auxiliary predictors. We also plan to adopt more rigorous data-splitting protocols, including homology-aware training/test splits, and to strengthen the evaluation through broader novelty comparisons against public databases, stronger docking baselines, and higher-fidelity validation methods.

CRediT authorship contribution statement

Asma Bendjeddou: Writing – original draft, Validation, Investigation, Data curation. **Kamyar Zeinalipour:** Writing – original draft, Software, Formal analysis, Data curation. **Dalal Bardou:** Software, Data curation. **Marco Maggini:** Writing – review & editing, Supervision, Methodology, Formal analysis, Conceptualization. **Franco Scarselli:** Writing – review & editing, Supervision, Funding acquisition, Formal analysis, Conceptualization. **Monica Bianchini:** Writing – review & editing, Supervision, Funding acquisition, Formal analysis, Conceptualization.

Ethics statement

This research presents an accurate account of the work performed, all data presented are accurate and methodologies detailed enough to permit others to replicate the work.

This manuscript represents entirely original works and or if work and/or words of others have been used, that this has been appropriately cited or quoted and permission has been obtained where necessary.

This material has not been published in whole or in part elsewhere.

The manuscript is not currently being considered for publication in another journal.

All authors have been personally and actively involved in substantive work leading to the manuscript and will hold themselves jointly and individually responsible for its content.

Declaration of Generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used ChatGPT in order to improve the quality of English text. After using these tools/services, the authors reviewed and edited the content as needed and took full responsibility for the content of the publication.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Monica Bianchini reports financial support was provided by University of Siena. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors thank **Alessia Lucia Prete** and **Barbara Toniella Corradini** for their valuable assistance and support and gratefully acknowledge the financial support of **Philogen S.p.A.**. This study was co-founded by the European Union -Next Generation EU, in the context of The National Recovery and Resilience Plan - Investment 1.5 Ecosystems of Innovation, Project Tuscany Health Ecosystem (THE), Spoke 3 - Advanced technologies, methods and materials for human health and well-being. ECS00000017, CUP: B63C22000680007. This work was also partially supported by Project DEEP-GRAPH, funded by the Italian Ministry of University and Research (MUR) PRIN 2022 (CUP: B53C24006570006).

References

[1] T.B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, 2020, arXiv preprint arXiv:2005.14165.

[2] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, F. Babaei, S. Bashlykov, Y. Bhosale, et al., LLaMA 2: Open foundation and fine-tuned chat models, 2023, arXiv preprint arXiv:2307.09288.

[3] D. Polykovskiy, A. Zhebrak, B. Sanchez-Lengeling, A. Aspuru-Guzik, M. Veselov, S. Nikolenko, et al., Molecular Sets (MOSES): A benchmarking platform for molecular generation models, Front. Pharmacol. 11 (2020) 565644.

[4] D.V. Laurents, AlphaFold 2 and NMR spectroscopy: Partners to understand protein structure, dynamics and function, Front. Mol. Biosci. 9 (2022) 906437.

[5] Y. Chen, Z. Wang, L. Wang, J. Wang, P. Li, D. Cao, X. Zeng, X. Ye, T. Sakurai, Deep generative model for drug design from protein target sequence, J. Cheminformatics 15 (1) (2023) 38, [Online]. Available: https://doi.org/10.1186/s13321-023-00702-2.

[6] N.C. Frey, A.L.C. D., I.R. Gould, J.R. Lane, A.D. White, Neural scaling of deep chemical models, Nat. Mach. Intell. 5 (11) (2023) 1107–1117, [Online]. Available: https://doi.org/10.1038/s42256-023-00740-3.

[7] Y. Labrak, A. Bazoge, E. Morin, P.-A. Gourraud, M. Rouvier, R. Dufour, BioMistral: A collection of open-source pretrained large language models for medical domains, 2024.

[8] A. Gaulton, A. Hersey, M. Nowotka, A.P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L.J. Bellis, Á. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M.P. Magariños, J.P. Overington, G. Papadatos, I. Smit, A.R. Leach, The ChEMBL database in 2017, Nucleic Acids Res. 45 (D1) (2017) D945–D954.

[9] R.G. Bickerton, G.V. Paolini, J. Besnard, S. Muresan, A.L. Hopkins, Quantifying the chemical beauty of drugs, Nat. Chem. 4 (2) (2012) 90–98.

[10] P. Ertl, A. Schuffenhauer, Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions, J. Cheminformatics 1 (1) (2009) 1–11.

[11] O. Trott, A.J. Olson, AutoDock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading, J. Comput. Chem. 31 (2) (2010) 455–461.

[12] T. Brown, et al., Language models are few-shot learners, Adv. Neural Inf. Process. Syst. (2020).

[13] H. Touvron, et al., LLaMA: Open and efficient foundation language models, 2023, arXiv preprint arXiv:2302.13971.

[14] A. Rives, et al., Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences, PNAS (2021).

[15] N. Ferruz, et al., ProtGPT2 is a deep unsupervised language model for protein design, Nat. Commun. (2022).

[16] M. Karimi, D. Wu, Z. Wang, Y. Shen, DeepAffinity: Interpretable deep learning of compound–protein affinity through unified recurrent and convolutional neural networks, Bioinformatics 35 (18) (2019) 3329–3338, [Online]. Available: https://doi.org/10.1093/bioinformatics/btz111.

[17] M. Wang, C.-Y. Hsieh, J. Wang, D. Wang, G. Weng, C. Shen, X. Yao, Z. Bing, H. Li, D. Cao, T. Hou, RELATION: A deep generative model for structure-based de novo drug design, J. Med. Chem. 65 (13) (2022) 9478–9492, [Online]. Available: https://doi.org/10.1021/acs.jmedchem.2c00732.

[18] Y. Li, J. Pei, L. Lai, Structure-based de novo drug design using 3D deep generative models, Chem. Sci. 12 (41) (2021) 13664–13675, [Online]. Available: https://doi.org/10.1039/D1SC04444C.

[19] H. Stärk, A. Jabbari, L. Pattanaik, N. Bode, R. Barzilay, T. Jaakkola, EquiBind: Geometric deep learning for drug binding structure prediction, in: Proceedings of the 39th International Conference on Machine Learning, ICML, PMLR, 2022, pp. 20503–20521, [Online]. Available: https://proceedings.mlr.press/v162/stark22a.html.

[20] F. Imrie, A.R. Bradley, M. van der Schaar, C.M. Deane, Deep generative models for 3D linker design, J. Chem. Inf. Model. 60 (4) (2020) 1983–1995, [Online]. Available: https://doi.org/10.1021/acs.jcim.9b01120.

[21] Y. Li, C. Gao, X. Song, X. Wang, Y. Xu, S. Han, DrugGPT: a GPT-based strategy for designing potential ligands targeting specific proteins, bioRxiv (2023) http://dx.doi.org/10.1101/2023.06.29.543848.

[22] E.J. Bjerrum, R. Threlfall, SMILES enumeration as data augmentation for neural network modeling of molecules, 2017, arXiv preprint arXiv:1703.07076.

[23] R. Gómez-Bombarelli, et al., Automatic chemical design using a data-driven continuous representation of molecules, ACS Central Sci. 4 (2) (2018) 268–276.

[24] S. Chithrananda, G. Grand, B. Ramsundar, ChemBERTa: Large-scale self-supervised pretraining for molecular property prediction, 2020, arXiv preprint arXiv:2010.09885, arXiv:2010.09885.

[25] V. Bagal, et al., MolGPT: Molecular generation using a transformer-decoder model, J. Chem. Inf. Model. 62 (9) (2021) 2064–2076.

[26] X. Yang, et al., DrugGPT: Generative pretrained transformer for target-aware molecule generation, 2023, arXiv preprint arXiv:2301.01071.

[27] K. Zeinalipour, N. Jamshidi, M. Bianchini, M. Maggini, M. Gori, Design proteins using large language models: Enhancements and comparative analyses, 2024, arXiv preprint arXiv:2408.06396.

[28] U. Consortium, UniProt: the universal protein knowledgebase in 2023, Nucleic Acids Res. 51 (D1) (2023) D523–D531.

[29] T. Sterling, J.J. Irwin, ZINC 15 – ligand discovery for everyone, J. Chem. Inf. Model. 55 (11) (2015) 2324–2337.

[30] T. Kudo, J. Richardson, SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing, in: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Brussels, Belgium, 2018, pp. 66–71.

[31] A.Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D.S. Chaplot, D.d.l. Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, et al., Mistral 7B, 2023, arXiv preprint arXiv:2310.06825.

[32] T. Dao, FlashAttention-2: Faster Attention with Better Parallelism and Work Partitioning, 2023, arXiv preprint arXiv:2307.08691.

[33] I. Loshchilov, F. Hutter, Decoupled Weight Decay Regularization, in: 7th International Conference on Learning Representations, ICLR, 2019, Conference Paper.

[34] M. Bavarian, H. Jun, N. Tezak, J. Schulman, C. McLeavey, J. Tworek, M. Chen, Efficient Training of Language Models to Fill in the Middle, 2022, arXiv preprint arXiv:2207.14255.

[35] T. Taffee, Elementary mathematical theory of classification and prediction, 1958, Unpublished manuscript.

[36] G. Landrum, RDKit: Open-source cheminformatics, 2006, Online, https://www.rdkit.org.

[37] O. Ursu, J. Holmes, C.G. Bologa, J.J. Yang, S.L. Mathias, V. Stathias, D.-T. Nguyen, S. Schürer, T. Oprea, DrugCentral 2018: an update, Nucleic Acids Res. 47 (D1) (2019) D963–D970.

[38] Y. Yang, S. Ouyang, X. Hu, M. Zheng, H. Zhou, L. Li, Structure-based drug design via 3D molecular generative pre-training and sampling, 2024, [Online]. Available: https://arxiv.org/abs/2402.14315.

[39] G. Xiong, Z. Wu, J. Yi, L. Fu, Z. Yang, C.-H. Hsieh, M. Yin, X. Zeng, C. Wu, A. Lu, X. Chen, T. Hou, D. Cao, ADMETlab 2.0: an integrated online platform for accurate and comprehensive predictions of ADMET properties, Nucleic Acids Res. 49 (W1) (2021) W5–W14.