Received: 24 February 2026 | Revised: 21 July 2026 | Accepted: 3 September 2026 | Published online: 20 September 2026

RESEARCH ARTICLE

Medinformatics
2026, Vol. 00(00) 1–14
DOI: 10.47852/bonviewMEDIN62029449

BON VIEW PUBLISHING

Generative AI for Drug Discovery: GPT-2 and LSTM-Based Models for Designing EGFR Inhibitors

Omar Dasser[1,*], Othmane Filali benaceur[1], Salma Fadel[2] and Rim Kaanane[3]

[1] *IBM Maroc, Morocco*

[2] *University of Abdelmalik Esaâdi, Morocco*

[3] *University of Mohamed V, Morocco*

Abstract: Effective inhibitors for the epidermal growth factor receptor (EGFR) still present a major obstacle in the development of cancer drugs. In this work, we create new EGFR inhibitors using generative artificial intelligence (AI) by fine-tuning a GPT-2 model on a dataset comprising around 500,000 molecules from the ChEMBL database. We evaluate our method against a Long Short-Term Memory (LSTM) network trained on the same dataset to create benchmarks. Before being filtered depending on Lipinski's rule of five, synthetic accessibility, and drug-likeness scores, the produced compounds were evaluated for validity, distinctiveness, and novelty. To assess their binding affinities, chosen candidates underwent molecular docking experiments against EGFR (PDB ID: 1M17). Our results show that although GPT-2 excels in producing structurally varied molecules, the LSTM model generates a larger fraction of chemically valid compounds. Many produced candidates showed good binding interactions with EGFR, therefore highlighting the possibilities of deep learning-based generative models in hastening the early phases of therapeutic development. This work presents a scalable method for creating tailored cancer treatments by showing the synergy between AI and in silico drug design.

Keywords: EGFR, GPT-2, LSTM, protein–ligand docking, in silico drug discovery

1. Introduction

The process of drug discovery has consistently been a laborious endeavor that spans several years. Despite advancements in high-throughput screening and combinatorial chemistry, the task of discovering new, effective, and safe compounds continues to be a significant challenge, primarily because of the high costs associated with experimental validation [1]. In recent years, the process has been progressively redefined through the growing use of artificial intelligence (AI). In particular, generative models have shown promise in proposing entirely new chemical structures that meet predefined criteria such as drug-likeness, synthetic accessibility, and target binding affinity [2]. Generative deep learning approaches offer powerful frameworks for automated molecular design, though careful evaluation of novelty, validity, and synthesizability remains critical [3].

Central to this shift is the adoption of SMILES strings. Representing "Simplified Molecular Input Line Entry System," SMILES provides a text-based method for describing chemical structures. It translates a molecule's connectivity and stereochemistry into a streamlined sequence of characters, allowing computers to effortlessly interpret and recreate it as 2D or 3D models [4]. This notation has established itself as a benchmark for sharing chemical data across databases and software platforms, as it offers an efficient and compatible representation. In its canonical form, SMILES provides a near-bijective mapping between molecular structures and their string representations (Figure 1), ensuring that each molecule corresponds to a unique string, which reduces redundancies and improves the consistency of the training.

Their structured yet flexible nature makes SMILES well-suited for deep learning, where molecules are treated as sequences and learned much like language; they are now widely used in generative models for exploring and designing novel compounds [5, 6].

Our research utilizes the Protein Data Bank (PDB) for the macromolecule 1M17 (Figure 2), on which the docking simulation of the generated candidate compounds will be performed [7]. This database provides experimentally validated 3D structural data essential for accurately modeling protein–ligand interactions and assessing binding affinity.

ChEMBL is a large-scale, manually curated bioactivity database that features chemical structures of drug-like compounds, as well as detailed information on bioactivity and targets [8]. The extensive dataset provided by ChEMBL, which encompasses binding affinities, assay outcomes, and pharmacological

*Corresponding author: Omar Dasser, IBM Maroc, Morocco. Email: omar.dasser1@ibm.com

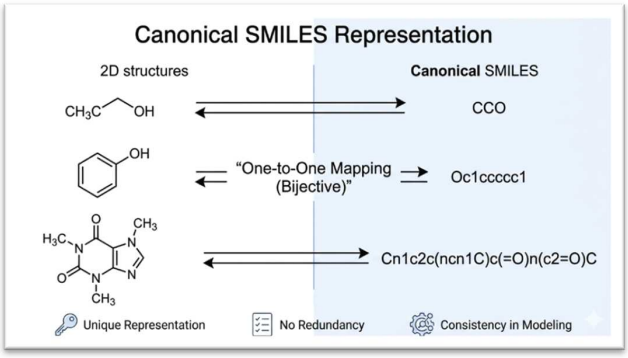

Figure 1. Depiction of canonical SMILES as a bijective molecular representation

Figure 2. Crystal structure of 1M17

profiles, is an essential resource for training machine learning models in the field of drug discovery.

One notable advancement is the adaptation of language models to the field of chemistry. Models such as GPT-2, initially created for natural language processing, have been retrained on SMILES strings, enabling them to effectively "speak" the language of molecules. Thanks to their Transformer architecture, these models can capture long-range dependencies in sequential data, which is crucial for generating innovative and diverse molecular designs [9]. In our research, we fine-tuned a GPT-2 as well as a Long Short-Term Memory (LSTM) model using the ChEMBL dataset, allowing them to generate SMILES strings that represent chemically plausible and synthetically accessible molecules. Following the generation process, we apply established filters based on Lipinski's rules and synthetic accessibility scores (SASs) to identify candidates with favorable pharmacokinetic profiles [10, 11].

This study focuses on the epidermal growth factor receptor (EGFR), a crucial protein associated with several cancers, including non-small cell lung cancer (NSCLC), colorectal cancer, and glioblastoma [12, 13]. For validation purposes, we utilize the crystal structure of EGFR (PDB ID: 1M17) as the docking target and employ PyRx to simulate binding interactions [14, 15]. By combining AI-driven molecule generation with thorough filtering and docking analysis, our approach seeks to enhance the efficiency of the initial stages of lead compound identification.

In the past few years, advances in deep generative modeling have revolutionized the field of de novo drug design. Early investigations using architectures such as generative adversarial networks and variational autoencoders (VAEs) demonstrated that deep learning can capture complex chemical representations and propose innovative therapeutic candidates that overcome the constraints of traditional combinatorial chemistry [16]. For instance, methods based on recurrent neural networks (RNNs) have been shown to generate diverse molecular libraries, often outperforming conventional techniques in terms of structural variety [2]. Deep learning enables exploration of novel chemical space and structure-property relationships; however, challenges related to model generalization and interpretability remain [17].

Graph neural networks are emerging as promising alternatives to SMILES-based models in molecular generation and property prediction; however, these approaches still exhibit notable limitations [18].

Subsequent evaluations of various generative strategies—including both VAEs and Transformer-based approaches—revealed that Transformer models excel at producing chemically valid and diverse structures. In this context, the GPT-2 model, originally designed for natural language processing, has been adapted to generate SMILES strings, a compact text representation of chemical compounds. By fine-tuning GPT-2 on extensive chemical datasets, researchers have successfully generated novel drug-like molecules. Further refinement through post-generation filtering using criteria such as Lipinski's rule of five and synthetic accessibility metrics ensures that the candidate compounds are optimized for practical drug discovery applications [9].

Another study introduced a deep transfer-learning strategy utilizing an LSTM architecture to design inhibitors targeting the SARS-CoV-2 main protease [19, 20]. This work demonstrated that such models can rapidly produce compounds with promising antiviral properties, underscoring their value during public health emergencies.

Meanwhile, molecular docking remains an indispensable tool in drug discovery. Platforms such as PyRx, which integrate AutoDock Vina [21], facilitate high-throughput docking simulations to assess the binding interactions between candidate molecules and their biological targets. Improvements in docking algorithms have led to faster and more accurate binding affinity predictions, creating a robust screening pipeline when combined with generative models. For example, docking studies have been effectively employed to validate AI-generated inhibitors for targets such as EGFR, a key oncogenic protein in cancer therapy.

Furthermore, the advent of high-accuracy protein structure prediction tools like AlphaFold has significantly enhanced the drug design process [22]. AlphaFold's capability to predict previously uncharacterized protein structures with exceptional precision provides reliable receptor models that improve docking studies. Its applications to targets such as the SARS-CoV-2 spike protein and mutated forms of EGFR have yielded critical structural insights that accelerate both vaccine development and the design of targeted inhibitors.

Collectively, these developments—from Transformer-based molecule generation and transfer learning to advanced docking simulations and state-of-the-art structure prediction—illustrate that the integration of modern AI techniques steadily bridges the gap between computational predictions and experimental outcomes in drug discovery.

Unlike previous EGFR molecular generation studies, this work directly compares Transformer-based (GPT-2) and recurrent (LSTM) architectures under identical training conditions while incorporating an iterative transfer-learning strategy driven by molecular docking feedback. This enables simultaneous evaluation of validity, uniqueness, novelty, drug-likeness, synthetic accessibility, and docking performance.

2. Materials and Methods

2.1. Data collection

The main format used in this work for presenting chemical compounds was SMILES (Simplified Molecular Input Line Entry System) [23, 24]. SMILES encodes complex chemical structures into a linear text format, letting sequential and generative models treat molecular design as a sequence-to-sequence task and rapidly generate and optimize novel drug-like candidates [25, 26]. Its textual nature integrates smoothly into deep learning architectures, easing the generation, validation, and transformation of molecular data relative to graphical or three-dimensional representations [27]. The compact SMILES format is also efficient for processing vast chemical libraries and compatible with cheminformatics tools such as molecular docking and drug-likeness assessment, and its canonical form ensures a single, consistent representation of each molecule during training and testing. In this work, we aimed to create SMILES strings representing compounds targeting EGFR, a protein linked to several malignancies, especially NSCLC [28]. EGFR was chosen for its central role in controlling differentiation, survival, and proliferation; its overexpression or mutation drives uncontrolled cell growth, and small-molecule inhibitors that block its signaling pathways have shown great success in many trials [29, 30]. Because it is often mutated in cancer patients and mutations like L858R and T790M create resistance against first-generation inhibitors [31], the EGFR protein is also a difficult target [32]. EGFR is therefore a promising target for new inhibitors able to overcome resistance mechanisms. Representing EGFR-blocking compounds as SMILES enables fast virtual screening of candidates, and randomized SMILES augmentation further improves generative-model learning by exposing alternative molecular representations during training [33].

2.2. Ligands

SMILES encodes molecules as compact, human-readable linear strings (e.g., ethanol is written "CCO"), and its character-based form is easily tokenized and fed to sequential models such as GPT-2 and LSTM for forecasting and synthesizing novel compounds [34]. Its wide adoption in cheminformatics databases such as ChEMBL [8], DrugBank [35], and PubChem [36] confirms its significance, and its adaptability captures stereochemistry and diverse bond types. This work derived a subset of almost 500,000 ligands with corresponding SMILES representations. The choice criteria guaranteed relevance to therapeutic uses by including compounds with claimed action against recognized pharmacological targets. To ensure compatibility with the GPT-2 and LSTM models, the SMILES dataset strings underwent rigorous preprocessing. This process included standardizing the SMILES format, removing salts and counterions, and filtering out molecules that did not meet fundamental chemical validity criteria, such as those with disconnected structures or incorrect syntax. After tokenizing the preprocessed SMILES strings, each character or collection of characters—tokens—represents an atomic or molecular substructure. The model was able to build appropriate chemical structures by means of this tokenizing approach. Several utility functions were developed: methods to translate molecules to SMILES and translate SMILES into an SDF file usable by PyRx.

2.3. Macromolecule

Here are some key structures of the EGFR available in the PDB, which are relevant for research:

1) PDB ID: 1M17 – This is the structure of the EGFR kinase domain bound to an inhibitor. It is one of the most studied EGFR structures and is frequently used in molecular docking and drug design studies.
2) PDB ID: 2ITN – This structure represents the EGFR kinase domain in complex with Erlotinib, a well-known tyrosine kinase inhibitor used for cancer treatment.
3) PDB ID: 4HJO – This structure shows the EGFR kinase domain with the inhibitor Afatinib, another important drug for EGFR-targeted cancer therapies.
4) PDB ID: 5UG9 – This is a structure of the EGFR T790M mutant bound to an irreversible inhibitor. The T790M mutation is associated with drug resistance in NSCLC.
5) PDB ID: 3W2S – This structure depicts the EGFR extracellular domain bound to an antibody fragment, providing insights into the receptor's conformation and its interactions with therapeutic antibodies.

These sites can provide excellent structural details that are crucial for understanding EGFR's interaction with various inhibitors and for conducting molecular docking studies. For our study, we used 1M17 [7].

The EGFR, a member of the ErbB family of receptor tyrosine kinases, controls proliferation, differentiation, and survival. It is especially important in oncology because its overexpression or mutation in malignancies such as glioblastoma, colorectal cancer, and NSCLC drives uncontrollable cell division [28].

Often the focus of medication creation is the EGFR kinase domain, which drives phosphorylation of downstream signaling proteins. Mutations in the kinase domain such as the well-documented L858R or T790M mutations are linked to both sensitivity to tyrosine kinase inhibitors and treatment resistance [37]. Small chemical targeting of the kinase domain helps to efficiently disrupt the downstream signaling cascades, therefore limiting tumor growth and proliferation.

PDB ID 1M17, a well-studied model of the EGFR kinase domain bound to an inhibitor, exposes the active conformation of the kinase and thus how small-molecule inhibitors bind and reduce its activity. Several EGFR-targeted treatments, including reversible inhibitors like those in Figure 2, have been modeled from this structure. Both 1M17, the crystal structure of EGFR, and the inhibitors Gefitinib and Erlotinib have been shown to significantly improve outcomes in EGFR-mutant NSCLC patients [28].

1M17 also highlights the key residues of the ATP-binding site, essential for designing inhibitors that outcompete ATP and

block EGFR phosphorylation. Structural insights from this model have guided first-, second-, and third-generation inhibitors aimed at overcoming resistance mutations such as T790M [12]. Modern drug research initiatives are still guided by the thorough awareness of EGFR's binding pocket and conformational flexibility given by the 1M17 structure.

3. Model Background

This pipeline architecture (Figure 3) serves as the foundation for progressively improving the precision of the generated molecules that meet our criteria for selection [30]. For both models.

GPT-2 was selected as a representative Transformer-based language model, whereas LSTM represents a well-established recurrent architecture for molecular generation, allowing a direct comparison between attention-based and recurrent sequence modeling.

The first goal depicted by the architecture above is to train a model that can generate SMILES with high or satisfactory validity and uniqueness. After that, we generate a sample of 10,000 SMILES that will be filtered using validity, uniqueness, Lipinski's rule of five, and, given that the generated molecules are novel, SAS. The filtered molecules are passed to PyRx to determine the binding affinity score, and then the top 100 molecules plus 15 from the general models to promote some randomness were used in transfer learning to fine-tune the models. The proposed framework does not perform a single generation step. Instead, the highest-scoring docked molecules are recursively used for transfer learning, progressively biasing subsequent generations toward regions of chemical space associated with improved EGFR affinity.

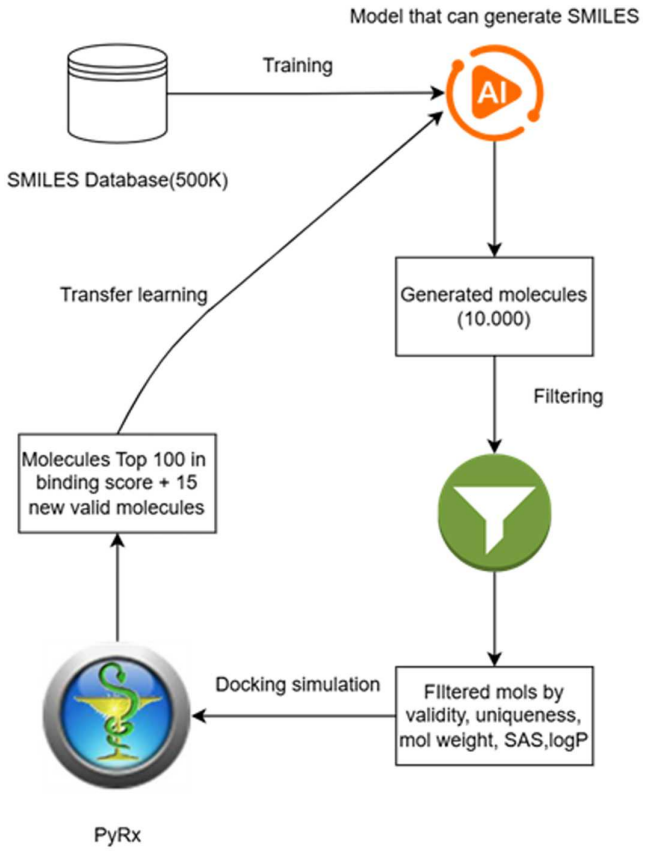

Figure 3. Process architecture

3.1. GPT-2 architecture and Transformer mechanism

Employed in this work to generate unique SMILES strings, the GPT-2 model is based on the Transformer architecture. Targeted compound generation through deep generative models enables more efficient lead optimization and scaffold hopping [38, 39]. Designed to analyze data sequences, the model excels at identifying long-range dependencies—essential for producing chemically correct SMILES. The original Transformer uses an encoder–decoder structure, whereas GPT models such as GPT-2 retain only the decoder stack and generate tokens autoregressively from previously generated tokens [33]. Self-attention and feed-forward neural networks abound in every layer of the Transformer.

1) Self-Attention Mechanism: At the heart of the Transformer model is the self-attention mechanism, which allows the model to weigh the importance of different tokens in the input sequence relative to each other. The attention mechanism is defined by Equations (1)–(4):

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{1}$$

where Q, K, and V represent the query, key, and value matrices, respectively. The denominator $\sqrt{d_k}$ is a scaling factor, where d_k is the dimensionality of the key vectors. This scaling prevents excessively large dot-product values that could cause the Softmax function to become saturated, thereby improving numerical stability and gradient flow during training.

2) Multi-Head Attention: The attention mechanism is applied multiple times in parallel, known as multi-head attention, which allows the model to capture different types of relationships between tokens [39]. Multi-head attention is defined as:

$$Multihead(Q, K, V) = Concat(head_1, head_2, \ldots head_h)W^O \tag{2}$$

where each head i is a separate attention mechanism with its own learned weight matrices and W^o is the output weight matrix.

3) Feed-Forward Networks: After the attention layers, each token is passed through a position-wise feed-forward network [40].

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \tag{3}$$

These feed-forward networks apply transformations to each token independently, helping to map the output of the attention layers into a richer, more expressive space.

4) Positional encoding: Since the Transformer architecture doesn't inherently capture the order of tokens, a positional encoding is added to the input embeddings [41]:

$$PE(pos, 2i) = sin\left(\frac{pos}{10000^{2i/d_{model}}}\right)$$

$$PE(pos, 2i+1) = cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{4}$$

Where:

Pos: The position of the token in the sequence (e.g., 0, 1, 2, ...).
i: The embedding dimension index.
d_{model}: The dimensionality of the embedding vector.

$10000^{\frac{i}{d}}$: A scaling factor that ensures the positional encodings operate over different wavelengths for different dimensions.

Sinusoidal functions allow the positional encodings to be periodic and continuous, enabling the model to generalize to unseen sequence lengths during inference. Using sine for even indices and cosine for odd indices ensures a unique encoding for each position and dimension.

3.2. GPT-2's adaptation for SMILES generation

GPT-2 is a unidirectional Transformer, which means that at each step, it generates the next token in a sequence based on the previously generated tokens. This autoregressive process is well-suited for generating SMILES strings [42], which require maintaining strict chemical rules and dependencies between atomic substructures. During training, GPT-2 learns to predict the next token in a sequence given the previous ones. In the case of SMILES generation, this corresponds to predicting the next atom or bond in a molecular structure. The model is trained using a maximum likelihood estimation (MLE) objective, which minimizes the difference between the predicted and actual sequences. The fine-tuning process employed in this study ensured that GPT-2 captured the nuanced relationships between atoms in SMILES strings [43]. Through its attention mechanism, the model learns to focus on chemically relevant parts of the sequence when generating new molecules. This is especially important for ensuring the chemical validity of the generated structures. The MLE objective is fundamental in training models like GPT-2, especially for sequence generation tasks such as generating SMILES strings. The goal of MLE is to maximize the likelihood that the model assigns to the correct sequence of tokens (e.g., atoms or molecular bonds in SMILES) in the training dataset. The training loss (LMLE) is defined in Equation (6):

For a given sequence of tokens:

$$X = (x_1, x_2 \ldots x_T) \tag{5}$$

where T denotes the length of the sequence and x_t represents the token at position t. The MLE objective maximizes the probability of the observed sequence under the model's learned parameters θ. This is done by minimizing the negative log-likelihood of the sequence:

$$LMLE = -\sum_{t=1}^{T} \log P(x_t \vee x_1, x_2, \ldots, x_{t-1}; \theta) \tag{6}$$

Where:

LMLE: The loss function (negative log-likelihood) to be minimized.
T: The length of the sequence.
x_t: Token at position t in the sequence.
$P(x_t|x_1, x_2, \ldots, x_{t-1}; \theta)$: The probability of the current token x_t, conditioned on all previous tokens in the sequence (x_1 to x_t), computed using the model with parameters θ.
θ: The parameters of the model (e.g., weights and biases of the neural network).

Log Probability: We use the logarithm of the conditional probability as it turns the product of probabilities into a sum, which is easier to compute and optimize [44].

Negative Log-Likelihood: We minimize the negative log-likelihood because maximizing the likelihood is equivalent to minimizing its negative, an approach more conventional in optimization routines [45].

During training, this loss function is optimized by adjusting the parameters θ using techniques like backpropagation and gradient descent to reduce the loss, improving the model's ability to predict the next token in a sequence.

In the context of SMILES generation, this means the model is learning to predict the next atom, bond, or molecular substructure in the string given all previous ones, thereby generating chemically valid molecules. Transformer-based language models, such as GPT architectures, are increasingly adopted for chemical space exploration due to their flexibility in capturing chemical grammar [46].

3.3. LSTM architecture

LSTM networks (Figure 4) are a specialized form of RNNs, designed to address the problem of long-term dependencies in sequential data. Unlike traditional RNNs, LSTMs are capable of learning from long sequences without suffering from the vanishing gradient problem, which often prevents RNNs from effectively capturing long-range dependencies [47]. LSTMs achieve this by using a series of gates—input, forget, and output gates—that regulate the flow of information into and out of the memory cells [48]. These gates allow the network to decide which information to keep, which to forget, and which to output at each time step, thus maintaining the important context across long sequences [49].

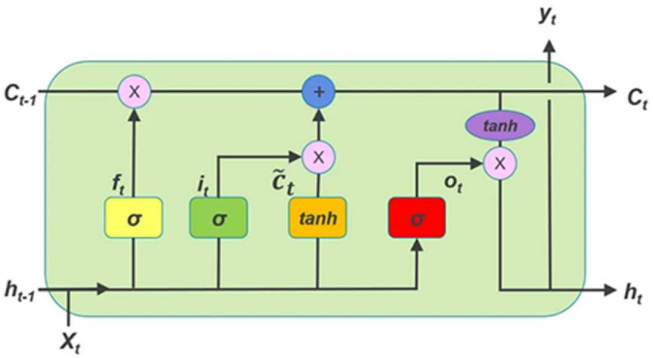

Figure 4. Architecture of an LSTM cell

Where:

- Forget Gate: $f_t = \sigma(W_f . [h_{t-1}, x_t] + b_f)$
  The forget gate determines which information from the previous state should be discarded.
- Input Gate: $i_t = \sigma(W_i . [h_{t-1}, x_t] + b_i)$
  The input gate controls the update of new information to the memory cell.
- Cell State Update: $Ct = ft \odot Ct - 1 + \odot \tanh(Wc \cdot xt + Uc \cdot ht - 1 + bc)$
  The cell state is updated based on the input and forget gates [48].
- Output Gate: $o_t = \sigma(W_o . [h_{t-1}, x_t] + b_o)$
  The output gate decides what part of the memory cell's state will be passed to the next hidden state.
- Hidden State Update: $h_t = o_t . \tanh(C_t)$
  The new hidden state is computed by combining the output gate's decision and the current cell state [47–49].
- And: $\tilde{C}_t = \tanh(W_c \cdot x_t + U_c \cdot h_{t-1} + b_c)$

3.4. SMILES generation and evaluation

SMILES Generation: Upon completion of training, the fine-tuned GPT-2 and LSTM models generated a batch of 1,000 novel SMILES strings by sampling from each model's output distribution, with temperature scaling applied to control the randomness of the generated sequences for both models.

Filtering Criteria: The generated molecules were subjected to a rigorous filtering process to ensure drug-likeness and synthetic feasibility:

Lipinski's Rule of Five: Each molecule was evaluated against Lipinski's rule of five, which stipulates that a drug-like molecule should have no more than five hydrogen bond donors, no more than 10 hydrogen bond acceptors, a molecular weight under 500 Daltons, and a logP (octanol-water partition coefficient) under 5 [9]. Beyond Lipinski's rule, additional physicochemical properties such as brain penetration and CNS permeability guide drug-likeness assessments; for example, for a drug to cross the blood-brain barrier, the molecular weight should usually be less than 300 Da among other conditions [50].

Drug-likeness: The drug-likeness of each molecule was assessed using QED (Quantitative Estimate of Drug-likeness), a metric that considers several physicochemical properties to quantify the likelihood that a molecule possesses favorable characteristics for drug development [51].

Synthetic Feasibility: Synthetic feasibility was evaluated using the SAS, which estimates the ease with which a molecule can be synthesized based on its structural complexity and the presence of common chemical fragments. Only molecules that met all three criteria were retained for further analysis. Machine-learned synthetic accessibility scoring such as RAscore enables rapid estimation of molecule synthesizability, complementing traditional SAS approaches [52].

In the context of SMILES generation, traditional metrics such as accuracy, precision, and Fβ score do not fully capture model performance; we need to employ new metrics like validity, uniqueness, novelty [28], and SAS.

These metrics are defined in Equations (7), (8), and (9):

$$Validity = \frac{|Valid(A)|}{|A|} \tag{7}$$

$$Uniqueness = \frac{|set(A)|}{|A|} \tag{8}$$

$$Novelty = \frac{|\{m \in Valid(A) : m \notin T\}|}{|Valid(A)|} \tag{9}$$

where A represents the generated molecules, Valid (A) refers to the valid molecules generated, Set (A) denotes the non-duplicate molecules generated, and T refers to the training set molecules. A conversion of the generated molecules into canonical SMILES is required before comparing them with the canonized training set.

The SAS is computed using the scoring functions defined in Equations (10)–(14):

$$Penalty = N_a^{1.005} - N_a \tag{10}$$

$$StereoPenalty = \log_{10}(N_c + 1) \tag{11}$$

$$SpiroPenalty = \log_{10}(N_s + 1) \tag{12}$$

$$BridgePenalty = \log_{10}(N_b + 1) \tag{13}$$

$$MacrocyclePenalty = \begin{cases} \log_{10}(2)\, if N_m > 0 \\ 0 if N_m = 0 \end{cases} \tag{14}$$

Where:

Na: Number of atoms
Nc: Number of chiral centers
Ns: Number of spiros
Nb: Number of bridgeheads
Nm: Number of macrocycles

3.5. Molecular docking studies

Target Macromolecule: The EGFR was selected as the docking target given its central role in cancer and its status as a validated drug target. The crystal structure of EGFR (PDB ID: 1M17) was downloaded from the PDB [7] and prepared for docking. This preparation included removing water molecules, adding polar hydrogens, and computing Gasteiger charges using AutoDockTools.

Docking Protocol: Molecular docking was performed using PyRx, a virtual screening tool that integrates AutoDock Vina for docking simulations. The prepared EGFR structure was used as the receptor, and the filtered SMILES strings were converted to PDBQT format for docking. Each ligand was docked into the active site of EGFR, with the docking grid centered on the binding site of the co-crystallized ligand in the 1M17 structure. The grid box dimensions were set to ensure coverage of the entire binding pocket, and the values are:

$$X = 100.20 \text{ Å} \quad Y = 73.89 \text{ Å} \quad Z = 59.38 \text{ Å}$$

Binding Affinity Calculation: The binding affinity of each ligand to EGFR was calculated using AutoDock Vina's scoring function, which estimates the free energy of binding. Ligands with the lowest binding affinity scores (i.e., most negative) were considered to have the strongest interactions with EGFR and were retained for further analysis and filtered further.

Statistical Analysis: The distribution of binding affinity scores was analyzed to identify outliers and trends among the top-scoring ligands. Descriptive statistics, such as mean, median, and standard deviation, were calculated to attempt to quantify the obtained results. Correlation analysis was conducted to explore the relationship between the binding affinity scores and the drug-likeness and synthetic feasibility metrics.

Selection of Lead Compounds: Based on the docking results, a subset of ligands with the most promising binding affinities, drug-likeness, and synthetic feasibility scores were selected as potential lead compounds. These leads were prioritized for further experimental validation and structure–activity relationship studies.

3.6. Docking protocol validation

Before docking the generated compounds, we validated (Figure 5) the protocol by redocking the co-crystallized ligand AQ4 into the prepared EGFR receptor (PDB ID: 1M17) under the same parameters used throughout the study. The search grid was centered on the crystallographic binding pocket at (22.0, 0.3, 52.8), with a box of 25 × 25 × 25 Å. Agreement between the redocked and crystallographic poses was

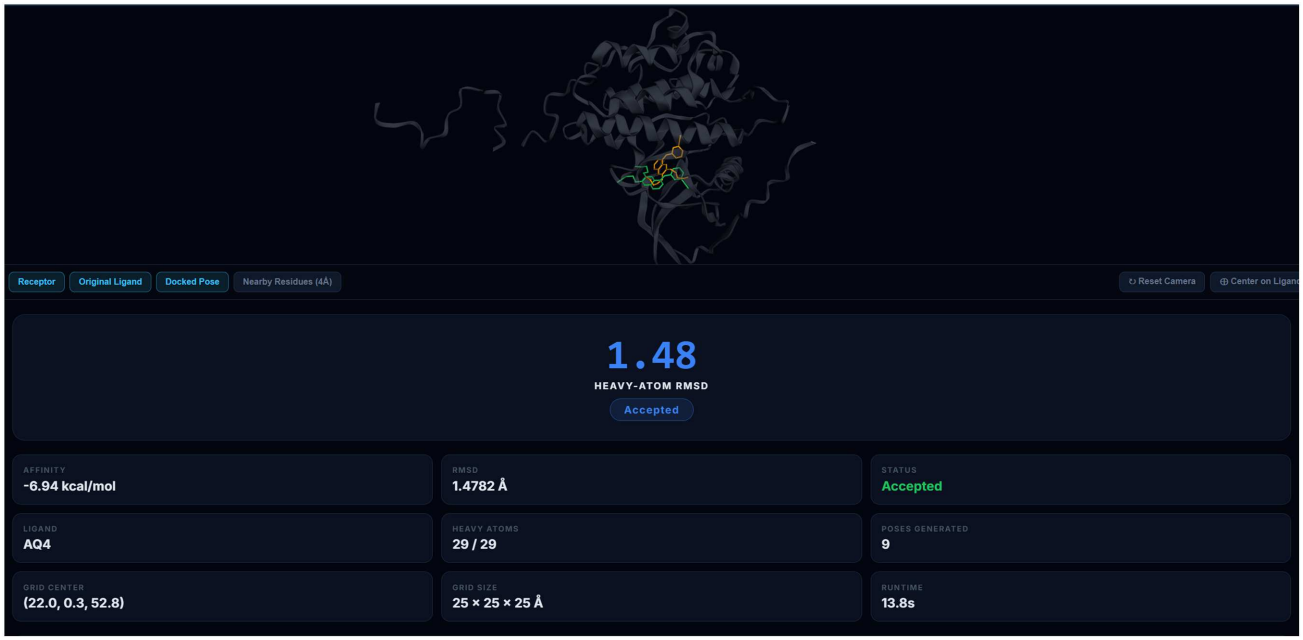

Figure 5. Redocking validation of the co-crystallized ligand AQ4 in EGFR (PDB ID: 1M17), showing a heavy-atom RMSD of 1.48 Å

measured as the heavy-atom root-mean-square deviation (RMSD), and an RMSD below 2.0 Å was taken to indicate a reliable protocol.

4. Results and Discussion

4.1. Training comparison

The training dynamics of the GPT-2 and LSTM models exhibit distinct behaviors that reflect the underlying architectural differences between Transformer-based and RNN approaches. These differences are evident in convergence speed, loss stability, and sensitivity to overfitting [53]. During training, the GPT-2 model showed a gradual reduction in loss over successive epochs, indicating steady learning of SMILES token distributions. However, the convergence process was relatively slower and exhibited higher variability compared to the LSTM model. As training progressed, continued optimization beyond a certain point did not lead to substantial performance gains and would likely increase the risk of overfitting. This behavior can be attributed to the high model capacity of GPT-2, which enables extensive exploration of chemical space but requires careful regularization and early stopping strategies to maintain generalization.

In contrast, the LSTM-based model demonstrated faster and more stable convergence during training. The loss decreased smoothly and reached a plateau after fewer epochs, suggesting efficient learning of sequential dependencies inherent to SMILES representations. The inclusion of batch normalization contributed to improved training stability and reduced sensitivity to fluctuations in gradient updates. Additionally, the gating mechanisms of the LSTM architecture facilitated effective retention of chemically relevant patterns across long sequences, limiting the generation of syntactically invalid structures during training [47, 48, 54].

Overall, the training phase highlights a fundamental trade-off between the two architectures. GPT-2 benefits from greater expressive capacity and flexibility but requires careful tuning to prevent overfitting, whereas the LSTM model offers a more stable and efficient training process for structured chemical sequences. These observations are consistent with the quantitative performance differences reported in subsequent sections and help explain the superior validity achieved by the LSTM model in this study.

We used the following LSTM parameters (Table 1).

The following table (Table 2) presents the hyperparameters for LSTM.

The parameters used for the GPT-2 model are as follows (Table 3).

Table 1. LSTM parameters

Parameter	Description	Value
Embedding Dim	Dimensionality of the character embedding layer	128 / 192 / 256 (search)
LSTM Layer 1 Units	Hidden units in the first LSTM layer	256 / 384 / 512 (search)
LSTM Layer 2 Units	Hidden units in the second LSTM layer	256 / 384 / 512 (search)
Dropout	Dropout rate applied after each BatchNorm	0.2 / 0.3 / 0.35 (search)
Recurrent Dropout	Dropout inside LSTM recurrent connections (set to 0 for cuDNN)	0
Vocab Size	Number of unique characters in the vocabulary (data-dependent)	Data dependent
Max Sequence Length	Clipped at 95th percentile of tagged SMILES lengths, range [32, 256]	Data dependent
Output Layer	TimeDistributed Dense layer over vocab size (logits, no softmax)	Vocabulary size

Table 2. LSTM hyperparameters

Hyperparameter	Description	Value
Max Epochs	Maximum training epochs (before early stopping)	100
Batch Size	Number of samples per gradient update	128
Learning Rate	Adam optimizer learning rate	3×10^{-4}
Optimizer	Optimization algorithm	Adam
Loss Function	Sparse categorical cross-entropy (from logits)	SparseCategorical Crossentropy
Validation Split	Fraction of data reserved for validation	0.1 (10%)
Shuffle Buffer	Buffer size for tf.data.Dataset.shuffle	10,000
Early Stopping Patience	Epochs to wait without val_loss improvement before stopping	6
Early Stopping Monitor	Metric monitored for early stopping	val_loss
Restore Best Weights	Restore model weights from the epoch with best val_loss	TRUE
Search Max Epochs	Epochs per candidate during architecture search	10
Random Seed	Seed for reproducibility (NumPy, TF, Python random)	42

Table 3. GPT-2 parameters

Parameter	Description	Value
Model Type	Pre-trained base model	gpt2 (GPT-2 Small)
Architecture	Model class	GPT2LMHeadModel
n_layer	Number of Transformer decoder layers	12
n_head	Number of attention heads per layer	12
n_embd	Hidden size/embedding dimensionality	768
n_ctx	Context window size (max tokens in a single forward pass)	1,024
n_positions	Maximum sequence length (positional embeddings)	1,024
n_inner	Inner dimensionality of the feed-forward layer (defaults to 4 × n_embd)	null (→ 3,072)
Vocab Size	Number of tokens in the vocabulary (after adding special tokens)	50,259
Activation Function	Nonlinearity in the feed-forward blocks	gelu_new
Layer Norm Epsilon	Epsilon for layer normalization stability	1×10^{-5}
Initializer Range	Std dev for weight initialization	0.02
Scale Attention Weights	Whether to scale attention scores by √(d_k)	TRUE
Scale Attn by Inverse Layer Idx	Additional per-layer attention scaling	FALSE
Reorder and Upcast Attn	Reorder attention ops and upcast to FP32	FALSE
Torch Dtype	Model weight precision	float32
Use Cache	Enable KV caching during generation	TRUE
Transformers Version	HuggingFace Transformers library version used	4.26.1

4.2. Results

We stopped fine-tuning our model after 20 epochs as we have noticed that the test loss begins to plateau at around epoch 10, indicating diminishing returns from further training. As seen in the loss curves (Figure 6), the training loss continues to decrease slightly, but the test loss shows little to no improvement after this point. This suggests that the model is no longer learning generalizable patterns from the data, and continued training would likely lead to overfitting, where the model improves on the training set at the expense of its ability to generalize the unseen data. Early stopping at this point ensures that the model achieves optimal performance without overfitting, balancing training efficiency with model accuracy.

The fine-tuned GPT-2 model demonstrated a strong capacity for generating structurally diverse SMILES representations of molecules. During evaluation, the model achieved a uniqueness rate of 100% (Table 4), indicating that all generated molecules were distinct and that the model was able to explore a broad region of chemical space. This high diversity highlights the effectiveness of the Transformer architecture in capturing long-range dependencies within SMILES sequences and avoiding repetitive molecular patterns.

In contrast, the chemical validity of the generated molecules reached 52.27%, revealing that while GPT-2 excels at diversity, it struggles to consistently produce chemically valid SMILES strings. This limitation is likely due to the strict syntactic and semantic constraints inherent to chemical representations, which

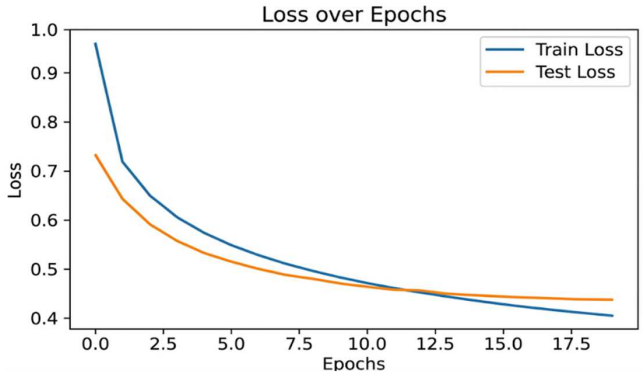

Figure 6. Loss over epochs for GPT-2

Table 4. Model generation metrics

Model	Uniqueness	Validity	Novelty
GPT-2	100%	52.27%	100%
LSTM	99.88%	90.98%	100%

are more challenging for autoregressive Transformer models to enforce without additional chemical constraints or post-processing steps. Nevertheless, the proportion of valid molecules remains sufficient to provide a meaningful pool of candidates for downstream filtering and docking analyses.

Temperature scaling played an important role in balancing diversity and coherence during generation. A temperature value of 0.7 was selected through a grid search, as it enabled sufficient exploration of novel structures while maintaining acceptable chemical consistency. Higher temperatures led to increased randomness and a further drop in validity, whereas lower temperatures reduced diversity without significantly improving correctness. Overall, these results demonstrate that GPT-2 is particularly well-suited for exploring novel regions of chemical space although additional optimization strategies would be required to improve chemical validity in future iterations [54].

Transformers such as GPT-2 often produce low-validity SMILES because they learn statistical patterns rather than strict chemical rules. SMILES requires exact syntax (e.g., ring closures, branching, valency), and even a small error can invalidate the molecule. While attention helps capture long-range dependencies, it does not enforce chemical correctness [55].

Character-level tokenization is an important factor, since generating SMILES one character at a time increases the risk of breaking syntax (e.g., unmatched parentheses or rings). However, it is not the only issue; the main limitation is the absence of explicit chemical constraints, meaning the model can still generate invalid molecules even with improved tokenization [1]. Future improvements could incorporate reinforcement learning to optimize the generative process using chemically informed reward functions, enabling the model to favor valid, synthesizable, and high-affinity molecules during generation. We trained the LSTM model for 100 epochs (Figure 7) using the same database mentioned before (approximately 500,000 SMILES strings) extracted from the ChEMBL database. The architecture featured a batch normalization LSTM designed to maintain long-term dependencies through its specialized gating mechanisms, including input, forget, and output gates, which regulate the flow of chemical information across the sequence. By implementing batch normalization and leveraging these gating units, the model achieved stable and efficient loss convergence, avoiding common issues like the vanishing gradient problem [47].

In comparison to GPT-2, the LSTM-based model exhibited substantially stronger performance in terms of chemical validity while maintaining a high degree of molecular diversity. The LSTM achieved a validity rate of 90.98%, significantly outperforming the GPT-2 model and indicating that the majority of generated SMILES strings corresponded to chemically valid structures. This result reflects the suitability of LSTM architectures for sequential data such as SMILES, where maintaining local and long-range dependencies between atoms and bonds is critical.

The uniqueness rate of the LSTM model reached 99.88%, which is comparable to GPT-2 and demonstrates that the improved validity did not come at the cost of excessive duplication. This balance between validity and diversity represents a desirable trade-off in de novo drug design, where generating chemically correct molecules is often prioritized over maximal structural novelty.

Overall, the LSTM model proved more reliable for generating chemically valid EGFR inhibitor candidates: while GPT-2 offers advantages in diversity and chemical space exploration, the LSTM achieved superior consistency, faster convergence, and

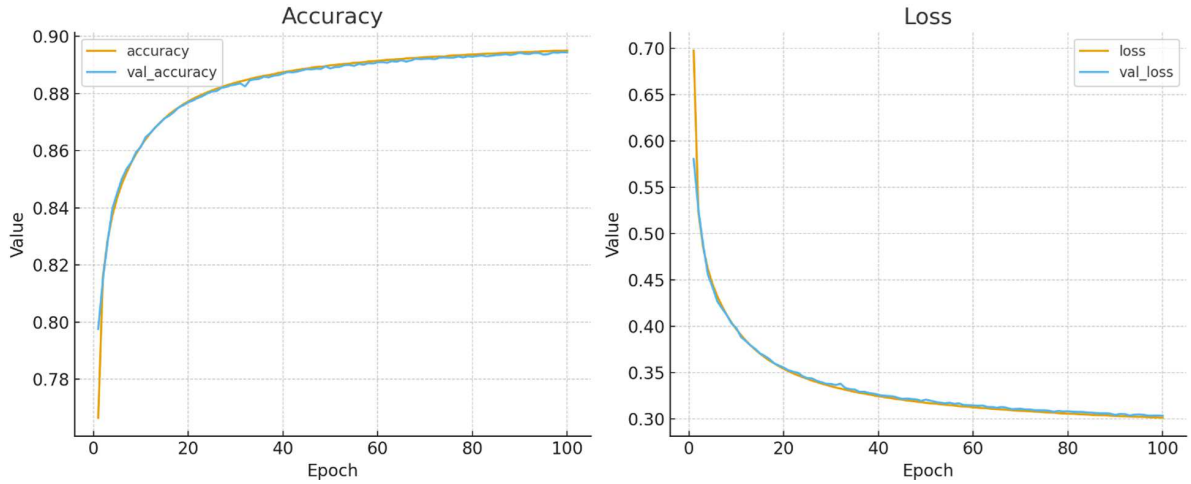

Figure 7. Evolution of training accuracy and loss of the LSTM model over 100 epochs

higher validity, making it more suitable for generating drug-like molecules in this study.

4.3. Iterative optimization across generations

To see if the iterative retraining was helping, we followed the docking scores over nine generations (Figure 8). The average score of the top 100 molecules kept improving, going from about −10.0 kcal/mol in the first generation to −11.45 kcal/mol by generation 9. In other words, each round was producing a stronger overall set of candidates. The best individual molecules improved as well, dropping from −11.8 kcal/mol at the start to a lowest score of −12.5 kcal/mol in generation 7. Overall, these results suggest that the pipeline was not just finding occasional strong hits but gradually improved the quality of the whole pool of generated molecules.

To further examine how the quality of candidates generated changed across generations, the full distribution of docking scores was analyzed for each iteration (Figure 9). As the generations progressed, the score distributions shifted toward more negative values, indicating a general improvement in predicted binding affinity rather than gains driven only by a few isolated molecules. This trend is also reflected in the gradual decrease of the median docking score from the early to the later generations. At the same time, a certain degree of variability was maintained within each generation, suggesting that the iterative retraining process continued to explore diverse regions of chemical space while enriching for stronger EGFR-binding candidates.

A Mann–Whitney U test showed the docking scores obtained after iterative optimization were significantly lower than those of Generation 1 ($p < 0.05$).

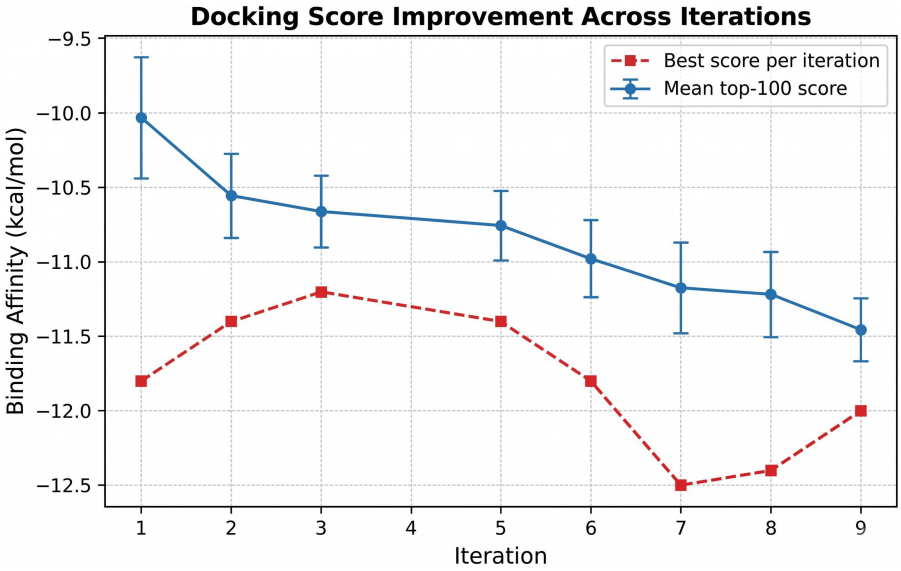

Figure 8. Progressive improvement in docking performance across nine iterative generations

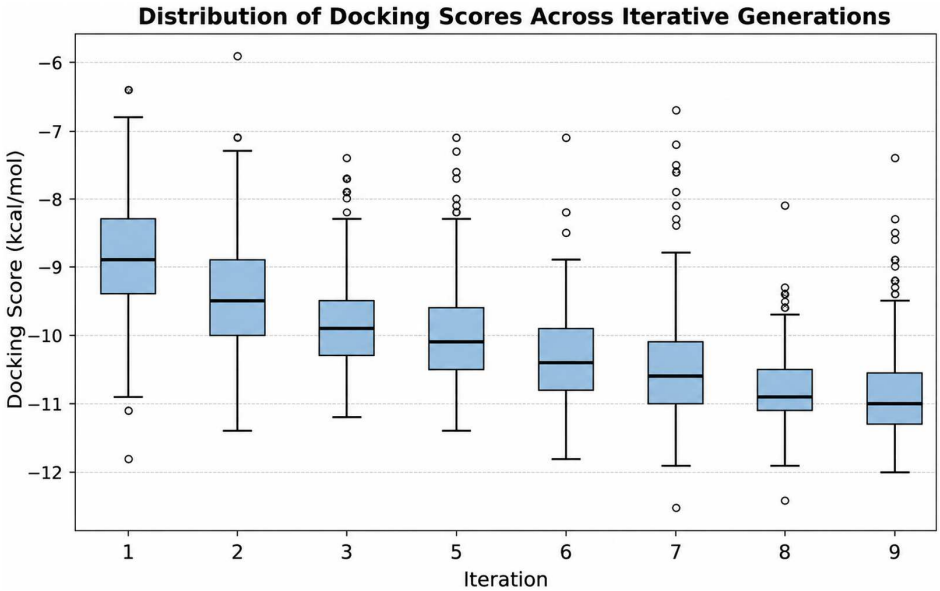

Figure 9. Boxplot distribution of docking scores across nine iterative generations

4.4. Benchmark EGFR inhibitors

Before presenting the generated molecules, we highlight current EGFR-targeting medications. These tyrosine kinase inhibitors block the EGFR signaling pathways that promote cancer cell growth and survival; well-known examples with their binding affinities and SMILES notations are listed below (Table 5).

These inhibitors are used to target EGFR mutations and block overactive signaling pathways that lead to uncontrolled cell proliferation. Erlotinib and Gefitinib are first-generation inhibitors, while Afatinib and Dacomitinib represent newer generations, designed to overcome resistance mutations like T790M [56].

In this section, we present (Table 6) promising molecules generated by the LSTM model, which exhibited strong binding affinities and favorable drug-like properties. These molecules were evaluated using multiple criteria, including binding affinities, molecular weights, hydrogen bond donors and acceptors, logP values, QED [51, 57, 58], and synthetic accessibility (SAS) scores. Below are some of the top candidates identified through this generative process.

This rigorous training approach allowed the model to effectively capture the complex chemical grammar of the dataset [16]. As demonstrated in the results, the trained LSTM model achieved a chemical validity rate of 90.98% and a uniqueness rate of 99.88%, confirming its superior reliability for generating valid drug-like molecules for EGFR inhibition compared to the GPT-2 baseline.

To evaluate robustness against clinically relevant resistance mechanisms, the top generated molecules were additionally docked against the EGFR T790M mutant structure (PDB ID: 5UG9), one of the most common mechanisms of acquired resistance to first-generation EGFR tyrosine kinase inhibitors (Table 7).

Table 5. Binding affinity score of existing EGFR inhibitors

Drug Name	Binding Affinity (Kcal/mol)	SMILES
Erlotinib	-7.5	COCCOC1=C(C=C2C(=C1)C(=NC=N2)NC3=CC=CC(=C3)C#C)OCCOC
Gefitinib	-7.6	COC1=C(C=C2C(=C1)N=CN=C2NC3=CC(=C(C=C3)F)Cl)OCCCN4CCOC C4
Afatinib	-8.7	CN(C)C/C=C/C(=O)NC1=C(C=C2C(=C1)C(=NC=N2)NC3=CC(=C(C=C3)F)Cl)O[C@H]4CCOC4
Dacomitinib	-8.7	COC1=C(C=C2C(=C1)N=CN=C2NC3=CC(=C(C=C3)F)Cl)NC(=O)/C=C/CN4CCCCC4

Table 6. Binding energy and pharmacokinetic properties of generated molecules

Binding Affinity (kcal/mol)	SMILES	Molecular Weight	H-Bond Donors	H-Bond Acceptors	LogP	QED	SAS
-9.4	Cc1nc(N)nc(Nc2ccc(C(=O)N3CCN(C(=O)N4CCCCC4)CC3)cc2)n1	424.509	2	7	1.8695	0.7714	2.3386
-10.1	Cc1ccc(C(=O)NC(=O)c2cccc(C3(C)NC(=O)c4ccccc43)c2)cc1	384.435	2	3	3.5721	0.678	2.7525
-9.7	Cc1ccc(-c2ccc(C(=O)N3CCN(C(=O)N4CCC(C)CC4)CC3)cn2)cc1	406.53	0	3	3.6667	0.7647	2.169
-10.4	Cc1ccc(-c2cc(NC(=O)c3cccc(C(=O)N4CCC(C(=O)N5CCCC5)CC4)c3)cnn2)cc1	497.60	1	5	4.1789	0.5706	2.3896
-10.1	CC(C)(C)c1cccc(-c2c[nH]c(=O)c3c2CN(C(=O)Nc2ccccc2)CC3)c1	401.51	2	2	4.9296	0.6394	2.4470
-10.1	Cc1ccc(CN2CCN(C(=O)Nc3cccc(C(F)(F)F)c3)CC2)nc1NC(=O)c1ccccc1	497.521	2	4	5.01	0.5196	2.3245

Table 7. Binding affinity score of generated molecules against EGFR T790M (PDB 5GU7)

Binding Affinity (Kcal/mol)	SMILES
-11.04	Cc1ccc(-c2cc(NC(=O)c3cccc(C(=O)N4CCC(C(=O)N5CCCC5)CC4)c3)cnn2)cc1
-10.47	Cc1ccc(-c2ccc(C(=O)N3CCN(C(=O)N4CCC(C)CC4)CC3)cn2)cc1
-10.45	Cc1ccc(CN2CCN(C(=O)Nc3cccc(C(F)(F)F)c3)CC2)nc1NC(=O)c1ccccc1
-10.34	CC(C)(C)c1cccc(-c2c[nH]c(=O)c3c2CN(C(=O)Nc2ccccc2)CC3)c1
-9.901	Cc1ccc(C(=O)NC(=O)c2cccc(C3(C)NC(=O)c4ccccc43)c2)cc1
-8.872	Cc1nc(N)nc(Nc2ccc(C(=O)N3CCN(C(=O)N4CCCCC4)CC3)cc2)n1

5. Conclusion

In this study, we demonstrated the potential of generative AI in drug development by enhancing GPT-2 and LSTM models to produce novel EGFR inhibitors. The LSTM design was better than GPT-2 in terms of validity (90.98%) and overall performance. It made compounds with strong binding affinities and good pharmacokinetic properties. These results show how helpful it is to use deep generative modeling, Lipinski-based filtering, and molecular docking together to find leads quickly. Future efforts will concentrate on the integration of other dataset types like SELFIES (Self-Referencing Embedded Strings), which were shown to yield a better validity rate, and also refining the model architecture; Like the emergent usage of eXtended LSTM [59, 60], or other models that have shown great potential recently like Hierarchical Language Models [61] or Markov-chain-based diffusion models and reinforcement learning optimization [62, 63], and also maybe some wet-lab validation to accelerate the progression of AI-driven cancer therapies.

Acknowledgment

The authors are grateful to IBM Maroc for allocating the time and resources for this study.

Ethical Statement

This study involved only in silico analyses using publicly available data and did not require ethical approval or informed consent.

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

Data Availability Statement

The data that support this work are available upon reasonable request to the corresponding author.

Author Contribution Statement

Omar Dasser: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration. **Othmane Filali benaceur:** Software, Visualization. **Salma Fadel:** Validation, Investigation, Writing – review & editing. **Rim Kaanane:** Validation, Formal analysis, Investigation, Writing – review & editing.

References

[1] Blanco-González, A., Cabezón, A., Seco-González, A., Conde-Torres, D., Antelo-Riveiro, P., Piñeiro, Á., & Garcia-Fandino, R. (2023). The role of AI in drug discovery: Challenges, opportunities, and strategies. *Pharmaceuticals*, *16*(6), 891. https://doi.org/10.3390/ph16060891

[2] Zeng, X., Wang, F., Luo, Y., Kang, S. G., Tang, J., Lightstone, F. C., . . . , & Cheng, F. (2022). Deep generative molecular design reshapes drug discovery. *Cell Reports Medicine*, *3*(12), 100794. https://doi.org/10.1016/j.xcrm.2022.100794

[3] Martinelli, D. D. (2022). Generative machine learning for de novo drug discovery: A systematic review. *Computers in Biology and Medicine*, *145*, 105403. https://doi.org/10.1016/j.compbiomed.2022.105403

[4] Nguyen-Vo, T. H., Teesdale-Spittle, P., Harvey, J. E., & Nguyen, B. P. (2024). Molecular representations in biocheminformatics. *Memetic Computing*, *16*(3), 519–536. https://doi.org/10.1007/s12293-024-00414-6

[5] Flores-Hernandez, H., & Martinez-Ledesma, E. (2024). A systematic review of deep learning chemical language models in recent era. *Journal of Cheminformatics*, *16*(1), 129. https://doi.org/10.1186/s13321-024-00916-y

[6] Nigam, A., Pollice, R., Krenn, M., Gomes, G. D. P., & Aspuru-Guzik, A. (2021). Beyond generative models: superfast traversal, optimization, novelty, exploration and discovery (Stoned) algorithm for molecules using SELFIES. *Chemical Science*, *12*(20), 7079–7090. https://doi.org/10.1039/d1sc00231g

[7] Truong, D. T., Ho, K., Nhi, H. T. Y., Nguyen, V. H., Dang, T. T., & Nguyen, M. T. (2024). Imidazole[1,5-a]pyridine derivatives as EGFR tyrosine kinase inhibitors unraveled by umbrella sampling and steered molecular dynamics simulations. *Scientific Reports*, *14*, 12218. https://doi.org/10.1038/s41598-024-62743-3

[8] Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., . . . , & Leach, A. R. (2024). The ChEMBL Database in 2023: A drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, *52*(D1), D1180–D1192. https://doi.org/10.1093/nar/gkad1004

[9] Tong, X., Liu, X., Tan, X., Li, X., Jiang, J., Xiong, Z., . . . , & Zheng, M. (2021). Generative models for de novo drug design. *Journal of Medicinal Chemistry*, *64*(19), 14011–14027. https://doi.org/10.1021/acs.jmedchem.1c00927

[10] Ivanenkov, Y., Zagribelnyy, B., Malyshev, A., Evteev, S., Terentiev, V., Kamya, P., . . . , & Zhavoronkov, A. (2023). The hitchhiker's guide to deep learning driven generative chemistry. *ACS Medicinal Chemistry Letters*, *14*(7), 901–915. https://doi.org/10.1021/acsmedchemlett.3c00041

[11] Karami, T. K., Hailu, S., Feng, S., Graham, R., & Gukasyan, H. J. (2022). Eyes on Lipinski's rule of five: A New "rule of thumb" for physicochemical design space of ophthalmic drugs. *Journal of Ocular Pharmacology and Therapeutics*, *38*(1), 43–55. https://doi.org/10.1089/jop.2021.0069

[12] Alharbi, K. S., Shaikh, M. A. J., Afzal, O., Altamimi, A. S. A., Almalki, W. H., Alzarea, S. I., . . . , & Gupta, G. (2022). An overview of epithelial growth factor receptor (EGFR) inhibitors in cancer therapy. *Chemico-biological interactions*, *366*, 110108. https://doi.org/10.1016/j.cbi.2022.110108

[13] Sha, C., & Lee, P. C. (2024). EGFR-targeted therapies: A literature review. *Journal of Clinical Medicine*, *13*(21), 6391. https://doi.org/10.3390/jcm13216391

[14] Kondapuram, S. K., Sarvagalla, S., & Coumar, M. S. (2021). Docking-based virtual screening using PyRx tool: Autophagy target Vps34 as a case study. In M. S. Coumar (Ed.), *Molecular Docking for Computer-Aided Drug Design* (pp. 463–477). Academic Press. https://doi.org/10.1016/B978-0-12-822312-3.00019-9

[15] Eberhardt, J., Santos-Martins, D., Tillack, A. F., & Forli, S. (2021). AutoDock Vina 1.2.0: New docking methods, expanded force field, and Python bindings. *Journal of*

Chemical Information and Modeling, *61*(8), 3891–3898. https://doi.org/10.1021/acs.jcim.1c00203

[16] Wang, M., Wang, Z., Sun, H., Wang, J., Shen, C., Weng, G., . . . , & Hou, T. (2022). Deep learning approaches for de novo drug design: An overview. *Current Opinion in Structural Biology*, *72*, 135–144. https://doi.org/10.1016/j.sbi.2021.10.001

[17] Askr, H., Elgeldawi, E., Aboul Ella, H., Elshaier, Y. A. M. M., Gomaa, M. M., & Hassanien, A. E. (2023). Deep learning in drug discovery: An integrative review and future challenges. *Artificial Intelligence Review*, *56*, 5975–6037. https://doi.org/10.1007/s10462-022-10306-1

[18] Cavalli, A., Decherchi, S., & Abate, C. (2023). Graph neural networks for conditional de novo drug design. *WIREs Computational Molecular Science*, *13*(4), e1651. https://doi.org/10.1002/wcms.1651

[19] Chenthamarakshan, V., Hoffman, S. C., Owen, C. D., Lukacik, P., Strain-Damerell, C., Fearon, D., . . . , & Das, P. (2023). Accelerating drug target inhibitor discovery with a deep generative foundation model. *Science Advances*, *9*(25), eadg7865. https://doi.org/10.1126/sciadv.adg7865

[20] Dasser, O., Tahri, M., Kila, L., & Sekkaki, A. (2022). Designing potential drugs that can target SARS-CoV-2's main protease: A proactive deep transfer learning approach using LSTM architecture. *World Bulletin of Public Health*, *12*, 133–150.

[21] Pawar, R. P., & Rohane, S. H. (2021). Role of AutoDock Vina in PyRx molecular docking. *Asian Journal of Research in Chemistry*, *14*(2), 132–134. https://doi.org/10.5958/0974-4150.2021.00024.9

[22] Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., . . . , & Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*(7873), 583–589. https://doi.org/10.1038/s41586-021-03819-2

[23] Raghunathan, S., & Priyakumar, U. D. (2022). Molecular representations for machine learning applications in chemistry. *International Journal of Quantum Chemistry*, *122*(7), e26870. https://doi.org/10.1002/qua.26870

[24] Krenn, M., Ai, Q., Barthel, S., Carson, N., Frei, A., Frey, N. C., . . . , & Aspuru-Guzik, A. (2022). SELFIES and the future of molecular string representations. *Patterns*, *3*(10), 100588. https://doi.org/10.1016/j.patter.2022.100588

[25] Guo, H., Xing, X., Zhou, Y., Jiang, W., Chen, X., Wang, T., . . . , & Xu, J. (2025). A survey of large language model for drug research and development. *IEEE Access*, *13*, 51110–51129. https://doi.org/10.1109/access.2025.3552256

[26] Sadeghi, S., Bui, A., Forooghi, A., Lu, J., & Ngom, A. (2024). Can large language models understand molecules? *BMC Bioinformatics*, *25*(1), 225. https://doi.org/10.1186/s12859-024-05847-x

[27] Wigh, D. S., Goodman, J. M., & Lapkin, A. A. (2022). A review of molecular representation in the age of machine learning. *WIREs Computational Molecular Science*, *12*(5), e1603. https://doi.org/10.1002/wcms.1603

[28] Boch, T., Köhler, J., Janning, M., & Loges, S. (2022). Targeting the EGF receptor family in non-small cell lung cancer-increased complexity and future perspectives. *Cancer Biology & Medicine*, *19*(11), 1543–1564. https://doi.org/10.20892/j.issn.2095-3941.2022.0540

[29] Halder, S., Basu, S., Lall, S. P., Ganti, A. K., Batra, S. K., & Seshacharyulu, P. (2023). Targeting the EGFR signaling pathway in cancer therapy: What's new in 2023? *Expert Opinion on Therapeutic Targets*, *27*(4–5), 305–324. https://doi.org/10.1080/14728222.2023.2218613

[30] Korshunova, M., Huang, N., Capuzzi, S., Radchenko, D. S., Savych, O., Moroz, Y. S., . . . , & Isayev, O. (2022). Generative and reinforcement learning approaches for the automated de novo design of bioactive compounds. *Communications Chemistry*, *5*(1), 129. https://doi.org/10.1038/s42004-022-00733-0

[31] Corvaja, C., Passaro, A., Attili, I., Aliaga, P. T., Spitaleri, G., Del Signore, E., & de Marinis, F. (2024). Advancements in fourth-generation EGFR TKIs in EGFR-mutant NSCLC: Bridging biological insights and therapeutic development. *Cancer Treatment Reviews*, *130*, 102824. https://doi.org/10.1016/j.ctrv.2024.102824

[32] Passaro, A., Jänne, P. A., Mok, T., & Peters, S. (2021). Overcoming therapy resistance in EGFR-mutant lung cancer. *Nature Cancer*, *2*(4), 377–391. https://doi.org/10.1038/s43018-021-00195-8

[33] Mswahili, M. E., & Jeong, Y. -S. (2024). Transformer-based models for chemical SMILES representation: A comprehensive literature review. *Heliyon*, *10*(20), e39038. https://doi.org/10.1016/j.heliyon.2024.e39038

[34] Ucak, U. V., Ashyrmamatov, I., & Lee, J. (2023). Improving the quality of chemical language model outcomes with atom-in-SMILES tokenization. *Journal of Cheminformatics*, *15*(1), 55. https://doi.org/10.1186/s13321-023-00725-9

[35] Knox, C., Wilson, M., Klinger, C. M., Franklin, M., Oler, E., Wilson, A., . . . , & Wishart, D. S. (2024). DrugBank 6.0: The DrugBank Knowledgebase for 2024. *Nucleic Acids Research*, *52*(D1), D1265–D1275. https://doi.org/10.1093/nar/gkad976

[36] Kim, S., Chen, J., Cheng, T., Gindulyte, A., He, J., He, S., . . . , & Bolton, E. E. (2025). PubChem 2025 update. *Nucleic Acids Res*, *53*(D1), D1516–D1525. https://doi.org/10.1093/nar/gkae1059

[37] Todsaporn, D., Mahalapbutr, P., Poo-Arporn, R. P., Choowongkomon, K., & Rungrotmongkol, T. (2022). Structural dynamics and kinase inhibitory activity of three generations of tyrosine kinase inhibitors against wild-type, L858R/T790M, and L858R/T790M/C797S EGFR. *Computers in Biology and Medicine*, *147*, 105787. https://doi.org/10.1016/j.compbiomed.2022.105787

[38] Zheng, S., Lei, Z., Ai, H., Chen, H., Deng, D., & Yang, Y. (2021). Deep scaffold hopping with multimodal transformer neural networks. *Journal of Cheminformatics*, *13*(1), 87. https://doi.org/10.1186/s13321-021-00565-5

[39] Bagal, V., Aggarwal, R., Vinod, P. K., & Priyakumar, U. D. (2022). MolGPT: Molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, *62*(9), 2064–2076. https://doi.org/10.1021/acs.jcim.1c00600

[40] Zhang, Y., Liu, C., Liu, M., Liu, T., Lin, H., Huang, C. B., & Ning, L. (2024). Attention is all you need: Utilizing attention in AI-enabled drug discovery. *Briefings in Bioinformatics*, *25*(1), bbad467. https://doi.org/10.1093/bib/bbad467

[41] Tay, Y., Dehghani, M., Bahri, D., & Metzler, D. (2022). Efficient transformers: A survey. *ACM Computing Surveys*, *55*(6), 1–28. https://doi.org/10.1145/3530811

[42] Wu, K., Xia, Y., Deng, P., Liu, R., Zhang, Y., Guo, H., . . . , & Liu, T.-Y. (2024). TamGen: Drug design with target-aware molecule generation through a chemical language model. *Nature Communications*, *15*(1), 9360. https://doi.org/10.1038/s41467-024-53632-4

[43] Haroon, S., Hafsath, C. A., & Jereesh, A. S. (2023). Generative Pre-trained Transformer (GPT) based model with relative attention for de novo drug design. *Computational Biology and Chemistry*, *106*, 107911. https://doi.org/10.1016/j.compbiolchem.2023.107911

[44] Wei, L., Fu, N., Song, Y., Wang, Q., & Hu, J. (2023). Probabilistic generative transformer language models for generative design of molecules. *Journal of Cheminformatics*, *15*, 88. https://doi.org/10.1186/s13321-023-00759-z

[45] Bishop, C. M., & Bishop, H. (2024). *Deep Learning: Foundations and Concepts*. Springer International Publishing. https://doi.org/10.1007/978-3-031-45468-4

[46] Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., & Das, P. (2022). Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, *4*(12), 1256–1264. https://doi.org/10.1038/s42256-022-00580-7

[47] Garg, S., & Krishnamurthi, R. (2023). A survey of long short term memory and its associated models in sustainable wind energy predictive analytics. *Artificial Intelligence Review*, *56*(S1), 1149–1198. https://doi.org/10.1007/s10462-023-10554-9

[48] Krichen, M., & Mihoub, A. (2025). Long short-term memory networks: A comprehensive survey. *AI*, *6*(9), 215. https://doi.org/10.3390/ai6090215

[49] Mienye, I. D., Swart, T. G., & Obaido, G. (2024). Recurrent neural networks: A comprehensive review of architectures, variants, and applications. *Information*, *15*(9), 517. https://doi.org/10.3390/info15090517

[50] Janigro, D., Bailey, D. M., Lehmann, S., Badaut, J., O'Flynn, R., Hirtz, C., & Marchi, N. (2021). Peripheral blood and salivary biomarkers of blood–brain barrier permeability and neuronal damage: Clinical and applied concepts. *Frontiers in Neurology*, *11*, 577312. https://doi.org/10.3389/fneur.2020.577312

[51] Li, B., Wang, Z., Liu, Z., Tao, Y., Sha, C., He, M., & Li, X. (2024). DrugMetric: Quantitative drug-likeness scoring based on chemical space distance. *Briefings in Bioinformatics*, *25*(4), bbae321. https://doi.org/10.1093/bib/bbae321

[52] Thakkar, A., Chadimová, V., Bjerrum, E. J., Engkvist, O., & Reymond, J. L. (2021). Retrosynthetic accessibility score (RAscore) - rapid machine learned synthesizability classification from AI driven retrosynthetic planning. *Chemical Science*, *12*(9), 3339–3349. https://doi.org/10.1039/d0sc05401a

[53] Chen, Y., Wang, Z., Zeng, X., Li, Y., Li, P., Ye, X., & Sakurai, T. (2023). Molecular language models: RNNs or transformer? *Briefings in Functional Genomics*, *22*(4), 392–400. https://doi.org/10.1093/bfgp/elad012

[54] Grisoni, F. (2023). Chemical language models for de novo drug design: Challenges and opportunities. *Current Opinion in Structural Biology*, *79*, 102527. https://doi.org/10.1016/j.sbi.2023.102527

[55] Schoenmaker, L., Béquignon, O. J. M., Jespers, W., & van Westen, G. J. P. (2023). UnCorrupt SMILES: a novel approach to de novo design. *Journal of Cheminformatics*, *15*(1), 22. https://doi.org/10.1186/s13321-023-00696-x

[56] Duggirala, K. B., Lee, Y., & Lee, K. (2022). Chronicles of EGFR Tyrosine Kinase Inhibitors: Targeting EGFR C797S Containing Triple Mutations. *Biomolecules & Therapeutics*, *30*(1), 19–27. https://doi.org/10.4062/biomolther.2021.047

[57] Zhu, W., Wang, Y., Niu, Y., Zhang, L., & Liu, Z. (2023). Current trends and challenges in drug-likeness prediction: Are they generalizable and interpretable? *Health Data Science*, *3*, 0098. https://doi.org/10.34133/hds.0098

[58] Gu, Y., Wang, Y., Zhu, K., Li, W., Liu, G., & Tang, Y. (2024). DBPP-Predictor: A novel strategy for prediction of chemical drug-likeness based on property profiles. *Journal of Cheminformatics*, *16*(1), 4. https://doi.org/10.1186/s13321-024-00800-9

[59] Meng, X., Xu, S., Yu, Y., Zhu, Y., & Gao, F. (2025). Extended Long Short-Term Memory network for robust state-of-health estimation of lithium-ion batteries under diverse charging strategies. *International Journal of Electrical Power & Energy Systems*, *172*, 111146. https://doi.org/10.1016/j.ijepes.2025.111146

[60] Sun, Y., Lu, Y., Li, Y. Y., Jing, Z., Leung, C. K., & Hu, P. (2025). MolGraph-xLSTM as a graph-based dual-level xLSTM framework for enhanced molecular representation and interpretability. *Communications Chemistry*, *8*(1), 286. https://doi.org/10.1038/s42004-025-01683-z

[61] Wang, Y., Zhang, H., Liu, Z., & Zhou, Q. (2021). Hierarchical concept-driven language model. *ACM Transactions on Knowledge Discovery from Data*, *15*(6), 1–22. https://doi.org/10.1145/3451167

[62] Alakhdar, A., Poczos, B., & Washburn, N. (2024). Diffusion models in de novo drug design. *Journal of Chemical Information and Modeling*, *64*(19), 7238–7256. https://doi.org/10.1021/acs.jcim.4c01107

[63] Svensson, H., Tyrchan, Gummesson., Engkvist, C., O, & Haghir Chehreghani, M. (2024). Utilizing reinforcement learning for de novo drug design. *Machine Learning*, *113*(7), 4811–4843. https://doi.org/10.1007/s10994-024-06519-w

How to Cite: Dasser, O., benaceur, O. F., Fadel, S., & Kaanane, R. (2026). Generative AI for Drug Discovery: GPT-2 and LSTM-Based Models for Designing EGFR Inhibitors. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN62029449