Automated Target Selection and De Novo Drug Generation For Intrahepatic Cholangiocarcinomas

Priestly Adejo Student Number: 21001232
[1] Department of Mechanical Engineering, University College London,
Torrington Place, London, WC1E 7JE, UK

Abstract

Cholangiocarcinoma (CCA), particularly its intrahepatic form (ICC), is often diagnosed in its advanced stages due to its typically asymptomatic progression, posing significant challenges in oncology. This study introduces an innovative automated drug discovery pipeline that utilizes a conditional variational autoencoder (ConditionalVAE) alongside a protein multimodal network (PMN) to generate and optimize novel ligands. This method is inspired by the success of numerous computational drug discovery companies and targeted FGFR2 inhibitor therapies, including FDA-approved drugs like ivosidenib and pemigatinib. The pipeline enhances drug discovery by integrating extensive molecular property predictions within its binding affinity model, leveraging advanced tools from Volkamer Lab's TeachOpenCadd for rapid molecular optimization and evaluation against extensive databases such as PubChem. The approach has led to the generation and optimization of several promising compounds, significantly expediting the drug development process for CCA and potentially other oncological targets. Its modular design facilitates customization and scalability, making it a foundational tool for researchers developing their own drug discovery pipelines or APIs. Ongoing enhancements include the integration of reinforcement learning models to further refine drug candidate selection. For more details and to contribute to this evolving project, visit our GitHub repository at https://github.com/PriestlyAdejo/drug_discovery_cholangiocarcinoma.

Keywords: Binding Affinity Prediction, Reinforcement Learning, Computational Drug Discovery, Intrahepatic Cholangiocarcinoma, Variational Autoencoders.

Nomenclature

AI Artificial Intelligence
ConditionalVAE Conditional Variational Encoder
FGFR2 Fibroblast Growth Factor Receptor 2
GENTRL Generative Tensorial Reinforcement Learning
ICC Intrahepatic Cholangiocarcinoma
logP Logarithm of the partition coefficient between n-octanol and water
ML Machine Learning
MolWt Molecular Weight
NumHAcceptors Number of Hydrogen Acceptors
NumHDonors Number of Hydrogen Donors
NumRotBonds Number of Rotatable Bonds
PDB Protein Data Bank
QED Quantitative Estimate of Drug-likeness
SMILES Simplified Molecular Input Line Entry System
TargetVAE Target-aware Variational Auto-encoder
TeachOpenCadd Platform That Provides Workflows for a Variety of Computer Aided Drug Design Tasks
TPSA Topological Polar Surface Area
UniProt Universal Protein Resource Database

1 Introduction

Cholangiocarcinoma (CCA), especially the intrahepatic type (ICC), poses significant diagnostic and therapeutic challenges due to its asymptomatic nature in early stages and complex anatomical considerations. The intricate microstructure and development of the

biliary tract significantly complicate early detection and effective management of the disease, often leading to a poor prognosis with advanced CCA frequently being inoperable, as documented by Nakanuma et al. [1]. The heterogeneity in presentation and progression, reviewed by Banales et al. [2], further emphasizes the complexity of this malignancy, affecting treatment outcomes across different patient populations.

Additionally, epidemiological data highlighted by Kamsa-Ard et al. [3] indicate variable incidence rates globally, with certain regions showing higher prevalence due to specific genetic and environmental factors. Recent statistics from the National Cancer Center Japan [4] show an increasing trend in CCA cases, highlighting the urgent need for advancements in therapeutic strategies. Despite developments in surgical techniques and localized treatments such as radiotherapy [5] and selective internal radiation therapy [6], the overall management of CCA often remains palliative, with systemic therapies providing limited survival benefits [3].

Recent clinical guidelines by Nagino et al. [4] and therapeutic advancements discussed by Oh et al. [7] point towards a paradigm shift in the treatment of advanced biliary tract cancers with the incorporation of targeted therapies and immunotherapies. These approaches exploit molecular targets specific to CCA, which are often identified through extensive genetic profiling and molecular studies like those conducted by Nakamura et al. [8] and Jain et al. [9]. Such targeted therapies, including FGFR2 inhibitors and IDH1 inhibitors, have shown promise in clinical trials, offering a new hope for patients with specific mutational profiles [10].

Despite significant advancements, there remains a notable deficiency in effective treatments for patients, particularly those with nonspecific, rare, or multiple genetic variations as Emma Lodi et al. [11] documented in the "Summary of Molecular Signaling Pathways and Targets in CCA" table in the appendix. This highlights the urgent need for comprehensive approaches that not only identify new therapeutic targets but also develops medications tailored to the complex molecular profile of CCA. The proposed automated drug discovery pipeline aims to address this gap by utilizing advanced computational techniques and machine learning to accelerate the generation and optimization of potential therapeutic agents, regardless of the variations in target expression among cases. By integrating sophisticated algorithms with systematic evaluations of molecular targets, this pipeline promises to enhance the personalization of treatment, thereby transforming the therapeutic landscape for cholangiocarcinoma.

2 Literature Review

2.1 Background of the Problem

The initial stage of drug discovery involves the understanding of complex biological pathways and identifying potential therapeutic targets. This pre-discovery phase is critical, as it lays the groundwork for all subsequent stages of drug development. Researchers like Gashaw et al. [12], Hughes et al. [13], and Mohs and Greig [14] delve deep into the molecular mechanisms of diseases, often integrating vast datasets from genomics, proteomics, and other omics technologies to hypothesize potential targets. This stage is particularly challenging due to the intricate nature of biological systems, which are often not fully understood. Moreover, the identification of a valid target is compounded by the need to distinguish between causative and merely associative factors in disease pathology. According to Natesh et al. [15], novel target identification is thus a critical bottleneck, requiring a nuanced understanding of disease mechanisms within to propose viable targets for therapeutic intervention.

2.2 Current Processes in Drug Discovery

Once potential targets are identified, the process advances towards the meticulous and iterative tasks of screening and optimizing drug candidates, as described by Hughes et al. [13]. High-throughput screening is employed across vast chemical libraries, exploring the immense chemical space estimated to range between 10^{30} to 10^{60} potential drug-like molecules. This stage aims to discover molecules that effectively modulate these targets, focusing not just on finding any modulator but on discovering one that offers specificity, efficacy, and safety—a challenging trifecta rarely easily achieved. The structural and chemical diversity of these libraries compounds the complexity of this task, requiring sophisticated screening methodologies and substantial computational resources to manage and interpret the data generated. Further, as highlighted by researchers like DiMasi [16] and Paul et al. [17], each potential candidate undergoes rigorous optimization to enhance its pharmacological properties, ensuring its safety and efficacy for clinical use. This stage is crucial and resource-intensive, often involving numerous modifications to the molecular structure of compounds to enhance their drug-like properties and minimize toxicity [18].

2.3 Challenges and the Role of Deep Learning

The traditional drug discovery process faces significant challenges, including high costs, lengthy timelines, and a high rate of attrition. DiMasi [16] highlights these challenges, noting the inefficiencies that plague conventional approaches. In response, deep-learning-based approaches (DLBA) have emerged as transformative forces in drug discovery. DLBA, as developed and applied by researchers such as Paul [17], Schneider [19], and Jayatunga [20], are increasingly applied to predict drug-receptor interactions, optimize drug candidates, and simulate clinical trial outcomes, potentially reducing both time and expenditure. These technologies are capable of analyzing complex datasets much faster than human researchers, providing insights that can guide decision-making processes. Furthermore, DLBA in drug discovery range from designing molecules, as discussed by Kontoyianni [21], to predicting the success of drug candidates in clinical trials, illustrating a shift towards more data-driven, predictive approaches. Multiple researchers, including Brogi and Calderone et al., Ruffolo et al., Jayatunga et al., Sadybekov et al. and Kontoyianni et al. [22, 22, 23, 20, 24, 21] seem to strongly agree that the integration of DLBA not only accelerates various phases of drug discovery but also enhances the precision with which these tasks are executed, thereby increasing the likelihood of success in the later stages of drug development.

2.4 Emerging Technologies and Trends in Generative Drug Discovery

Deep learning (DL) has profoundly transformed generative drug discovery, evidenced by companies like Insilico Medicine, which utilizes its PandaOmics platform integrating autoencoder technologies and reinforcement learning to reduce drug discovery times from 3–5 years to 12–18 months [25]. This platform, leveraging variational autoencoders (VAEs) and reinforcement learning, significantly accelerates the exploration of chemical space [26], generating optimized molecular structures with desirable properties such as binding affinity. Meanwhile, Exscientia and Recursion Pharmaceuticals employ graph neural networks and data-intensive approaches respectively, which generally result in slower progression [27, 28]. Such technologies underscore a shift towards more efficient drug discovery processes as DL systems set new industry benchmarks [29]. Additionally, the integration of deep learning-based approaches from target discovery to preclinical trials highlights a significant evolution in the drug development pipeline, overcoming challenges like data biases and the need for empirical validation [30]. These advancements are expected to not only speed up the discovery process but also reduce costs and enhance the efficacy and safety of drugs, heralding a new era of efficiency in pharmaceutical development [31].

2.5 Aims of the Project

This research project aims to establish a comprehensive pipeline for drug discovery, focusing on a mutated target commonly expressed in intrahepatic cholangiocarcinoma (ICC) (FGFR-2) by leveraging and making improvements to the models used in the breakthroughs made by Ngo and Hy et al. [32]. This approach allows for the efficient exploration of chemical space, significantly reducing the search space for viable drug candidates.

The pipeline will integrate the generative capabilities of TargetVAE within a structured workflow adapted from Volkamer Lab's TeachOpenCADD platform, which offers a variety of tools for hit identification and lead optimization [33]. These tools include modules for binding site detection, molecular docking, and the evaluation of ligand efficiency and drug-likeness, crucial for optimizing compounds for better efficacy and safety profiles. This integration aims to enhance the predictability and efficiency of the drug development process, particularly in the stages of hit to lead and lead optimization, by automating the selection and refinement of potential drug molecules.

2.6 Summary of Work

This paper reviews the integration of deep learning (DL) technologies in drug discovery, emphasizing the substantial reduction in development times achieved through variational autoencoders (VAEs) and reinforcement learning. It explores how these technologies optimize chemical space to generate drug candidates with enhanced binding affinity and selectivity. Notably, platforms like Insilico Medicine have significantly shortened preclinical drug discovery phases. The paper also considers the adaptation of the future prospoects of using reinforcement learning models to refine chemical properties, thereby improving drug efficacy and safety.

3 Methodology

3.1 Target Protein Selection: FGFR-2

Intrahepatic cholangiocarcinoma (ICC) is characterized by diverse genetic alterations, prominently including mutations in IDH1, ARID1A, TP53, and FGFR2 (mutated targets). The selection of FGFR-2 was particularly influenced by its critical role in ICC pathogenesis and the prevalence of actionable FGFR2 fusions in 11-14% of ICC cases with Kuwat et al. [34]. This focus aligns with evidence suggesting that targeting FGFR2 could improve locoregional therapy efficacy and patient survival rates [35]. Typically, target selection in drug discovery involves a comprehensive evaluation of a target's role in disease based on extensive molecular profiling and is validated through various stages, including clinical trials [36]. While modern target discovery systems like Insilico Medicine's PandaOmics automate target discovery by analyzing vast datasets with AI/ML, this project uses manual selection to refine focus on FGFR-2, enabling detailed comparisons and benchmarking against established therapies [37].

3.2 Computational Representation of Target Protein

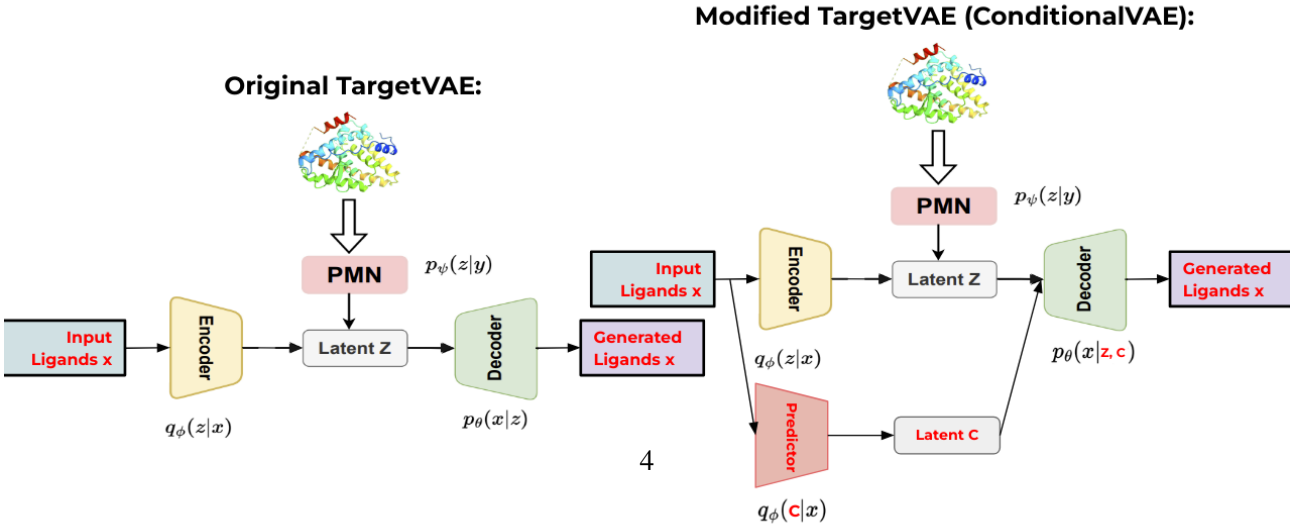
The Target-aware Variational Auto-encoders (Target VAE) study, by Ngo and Hy [32], introduces a Protein Modal Network (PMN) designed to effectively capture and integrate multimodal protein representations for ligand generation [38]. This PMN uniquely amalgamates textual protein sequences with three-dimensional structural data, employing a combination of transformer architectures and geometric deep learning techniques. The textual sequence data is processed using self-attention mechanisms within a transformer model to extract contextual embeddings [39]. Simultaneously, the 3D structural information of amino acids, crucial for understanding protein functionality, is represented through a Geometric Vector Perceptron (GVP). This component is essential for handling the rotational invariant features of protein structures, thereby maintaining the biological relevance of the spatial configuration.

The Geometric Vector Perceptron (GVP) operates by transforming protein residue features, categorized into scalar ($s \in \mathbb{R}^n$) and vector ($V \in \mathbb{R}^{\mu \times 3}$) components, which are respectively invariant and equivariant under Euclidean transformations. In the GVP module, the input tuple of scalar and vector features is transformed into a new tuple through a sequence of specialized linear and nonlinear transformations. Initially, the vector features undergo a linear transformation ($\mathrm{W}_h V$) to enhance dimensional control and facilitate the extraction of rotation-invariant information, followed by a non-linear activation function (σ_+) applied element-wise. Simultaneously, the scalar features are enriched by concatenating with the L2-norm of the transformed vector features, followed by their own linear transformation (W_m) and non-linear activation (σ). These processes are captured in the equations:

$$V' = \sigma_+(\mathrm{W}_\mu(\|\mathrm{W}_h V\|_2)) \odot \mathrm{W}_\mu(\mathrm{W}_h V),$$
$$s' = \sigma(\mathrm{W}_m(\|V\|_2, s) + b),$$

where σ and σ_+ represent the activation functions tailored for scalar and vector outputs respectively. This dual-pathway transformation within the GVP allows for the effective propagation of information from vector to scalar channels, ensuring that the PMN can harness both local and global structural details crucial for precise protein-ligand interaction modeling, while maintaining the necessary properties of invariance and equivariance [40].

3.3 Drug Molecule Generation with Modified TargetVAE



Within the standard TargetVAE model from Ngo and Hy et al. [32], we have a generative model and an inference model. The generative model, defined by a probabilistic decoder $p_\theta(x|z)$ and a prior $p_\psi(z)$, forms a joint distribution $p_{\theta,\psi}(x,z) = p_\theta(x|z)p_\psi(z)$ between latent variables z and data x. The inference model uses an encoder $q_\phi(z|x)$ to approximate the posterior distribution of the latent space. The objective is to maximize the evidence lower bound (ELBO) for each training sample x:

$$\begin{aligned}\log p_{\theta,\psi}(x) \geq \mathbb{E}_{q_\phi}[\log p_{\theta,\psi}(x|z)] \\ - \mathrm{KL}[q_\phi(z|x)||p_\psi(z)],\end{aligned} \tag{1}$$

where KL denotes the Kullback-Leibler divergence [41].

To improve the generative capabilities for specific targets, we use a conditional VAE (ConditionalVAE) as described by Tong et al. [42], which incorporates auxiliary covariates y. The generative model now defines a conditional joint distribution $p_{\theta,\psi}(x,z|y) = p_\theta(x|y,z)p_\psi(z|y)$, and the encoder is modified to $q_\phi(z|x,y)$. The conditional ELBO is maximized as follows:

$$\begin{aligned}\log p_{\theta,\psi}(x|y) \geq \mathcal{L}_{\text{cond}} = \mathbb{E}_{q_\phi}[\log p_{\theta,\psi}(x|y,z)] \\ - \mathrm{KL}[q_\phi(z|x,y)||p_\psi(z|y)].\end{aligned} \tag{2}$$

To further enhance the model, we introduce a predictor network for conditions that cannot be directly labeled, such as biological activity against specific targets. This modification results in a hybrid ConditionalVAE, where the condition vector c is treated as a latent variable inferred from the predictor. The objective function for unlabeled molecules is then modified to:

$$\begin{aligned}\mathcal{L}_{\text{hybrid}} = \mathbb{E}_{q_\phi}[\log p_{\theta,\psi}(x|z,c)] \\ - \mathrm{KL}[q_\phi(z|x)||p_\psi(z|c)] \\ - \mathrm{KL}[q_\phi(c|x)||p(c)].\end{aligned} \tag{3}$$

The modified framework thus incorporates a more flexible conditioning mechanism, enhancing the diversity and validity of generated molecules [42]. This approach aligns with the overarching goal of leveraging advanced machine learning techniques to facilitate the discovery of novel drug candidates with high efficacy and specificity against targeted proteins.

3.3.1 Experimental Setup for ConditionalVAE

The experimental setup for the ConditionalVAE model involved a sophisticated computational framework utilizing Python and PyTorch Geometric for graph creation and manipulation. We utilized the KIBA dataset, which includes 229 proteins and 2,111 ligands, leading to 118,254 protein-ligand pairs [43]. This dataset was augmented with 3D structural data, where the proteins are split into 90% for training and 10% for testing. The model's performance was evaluated against various metrics to assess its efficiency in generating valid, diverse, and chemically accurate ligands (appendix). The full implementation and codebase are made available on our project's GitHub repository, ensuring transparency and reproducibility of the results.

3.3.2 Evaluation Metrics for Drug-Likeness

In modifying the evaluation process, traditional metrics such as the Fréchet ChemNet Distance (FCD) [44] were eschewed in favor of direct assessments of drug-likeness and molecular properties. This includes measures like molecular weight, hydrogen bond donors and acceptors, logP, topological polar surface area (TPSA), and the number of rotatable bonds. Additionally, customized drug-likeness scores from Volkamler Lab's TeachOpenCaDD were computed using Lipinski's rule of five [45, 46] and a novel custom scoring function designed to assess the synthetic accessibility and overall quality of generated molecules:

$$\begin{aligned}\text{Drug-Likeness Score} = \frac{1}{7}\Big(&\text{molWt\_score} \\ &+ \text{numHAcceptors\_score} \\ &+ \text{numHDonors\_score} \\ &+ \text{molLogP\_score} \\ &+ \text{TPSA\_score} \\ &+ \text{numRotBonds\_score} \\ &+ \text{saturation\_score}\Big)\end{aligned} \tag{4}$$

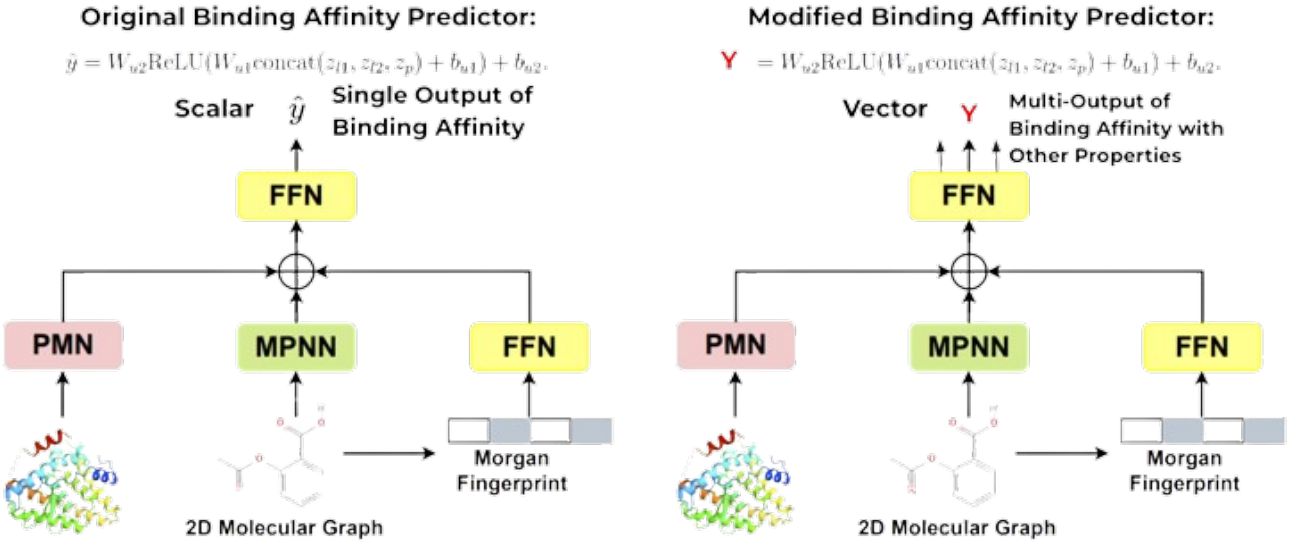
This scoring formula integrates multiple aspects of a molecule to provide a comprehensive drug-likeness profile, enhancing the utility of our predictions for pharmaceutical applications.

3.3.3 Experimental Execution and Outcomes

The experimental setup was limited by time and computational resources, and therefore, conducted using a single data fold on a laptop equipped with an Nvidia RTX 3070 GPU, spanning 130 epochs. Despite these limitations, valuable insights were gained into the performance of the modified ConditionalVAE [42]. Notably, we observed variations in the drug-likeness scores among the generated molecules, with a focus on properties such as druggability, uniqueness, and novelty. We provided visual representations of these variations to illustrate the impact of our modifications on both druggable and non-druggable ligands, demonstrating their potential validity for subsequent stages

of the drug discovery process. Comparative analyses with existing models highlighted the improvements made by our approach as seen in the appendix. This was especially seen in balancing drug-like properties with high binding affinities. This showcases the adaptability and potential of our ConditionalVAE to generate therapeutically relevant molecules.

3.4 Binding Affinity Prediction with Modified Binding Affinity Predictor



(a) Comparison of the original binding affinity predictor model from Ngo and Hy et al. [32] to the modified binding affinity predictor model.

In continuation of the work put forward by Ngo and Hy, their original binding affinity predictor was designed to compute the affinity between a generated ligand and a target protein structure [32]. The original model utilizes a combination of geometric and relational features extracted from protein-ligand interactions. This model integrates features using a protein multimodal network (PMN) and feed-forward networks of various layers (FFNs) to predict the binding affinity score. The network operates by first transforming the 2D molecular graph of a ligand into a vector representation using a message-passing neural network (MPNN). Simultaneously, ligand features such as Morgan Fingerprints are processed via a feed-forward network to produce another vector representation [47]. These representations, alongside the encoded protein structure from the PMN, are combined and fed into a final FFN that predicts the binding affinity as a scalar value:

$$z_{l1} = \text{MPNN}(G_l), \tag{5}$$
$$z_{l2} = W_{m2} \cdot \text{ReLU}(W_{m1} \cdot v_m + b_{m1}) + b_{m2}, \tag{6}$$
$$z_p = \text{PMN}(p), \tag{7}$$
$$y = \text{FFN}(\text{concat}(z_{l1}, z_{l2}, z_p)). \tag{8}$$

Modifications were made to the original model to enhance its predictive capabilities and output additional qualities about the binding interaction. The modified network now not only predicts the binding affinity in kcal/mol but also estimates the distance from the best mode root-mean-square deviation (RMSD) for both lower and upper bounds. This expanded output provides a more comprehensive understanding of the ligand's fit within the binding site. Additionally, the model now assumes a binding mode of 1 [48] for all input ligand-target protein pairs, simplifying the complexity of predicting multiple binding modes [49]. The governing equations remain similar, but the final output layer has been adapted to produce multiple outputs:

$$Y = \text{FFN}_{\text{modified}}(\text{concat}(z_{l1}, z_{l2}, z_p)), \tag{9}$$

where Y is a vector that includes not only the binding affinity but also distances indicating the quality of the binding pose. These modifications aim to provide a richer set of data for evaluating potential drug candidates, improving the utility of the model in real-world drug discovery applications and lays the foundation for predicting additional properties of ligand-protein interactions [50].

3.4.1 Experimental Setup and Data Processing

The experimental setup for the binding affinity prediction involves extensive data processing and computational modeling. We utilize the DAVIS and KIBA datasets, which contain information on protein-ligand interactions, with the binding affinities expressed as K_D constants and KIBA scores, respectively [43, 51]. The datasets include 30,056 and 118,254 protein-ligand pairs for DAVIS and KIBA respectively, enriched with 3D protein structures generated by AlphaFold [52]. Data preprocessing was performed in Python, utilizing PyTorch and PyTorch Geometric for graph-based neural network implementations. The experiments were conducted using a train-test split as specified in previous studies, ensuring fair comparison. The performance of the models were evaluated using metrics like Mean Squared Error (MSE), concordance index (CI), and the modified squared correlation coefficient r_m^2, providing a comprehensive assessment of their predictive accuracy.

3.4.2 Evaluation Metrics

To quantitatively assess the performance of our modified binding affinity predictor, we employ several statistical metrics cited by Ngo et al. [32]. These metrics include Mean Squared Error (MSE), Concordance Index (CI), and the modified squared correlation coefficient (r_m^2), defined as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2, \tag{10}$$

where n is the number of samples, y_i is the observed value, and $\hat{y}_i$ is the predicted value.

$$\text{CI} = \frac{1}{Z}\sum_{\delta_i > \delta_j} h(\hat{y}_i - \hat{y}_j), \tag{11}$$

where $\delta_i > \delta_j$ indicates that the affinity δ_i is greater than δ_j, Z is the normalization constant, and $h(x)$ is the Heaviside step function defined by:

$$h(x) = \begin{cases} 1 & \text{if } x > 0, \\ 0.5 & \text{if } x = 0, \\ 0 & \text{if } x < 0. \end{cases}$$

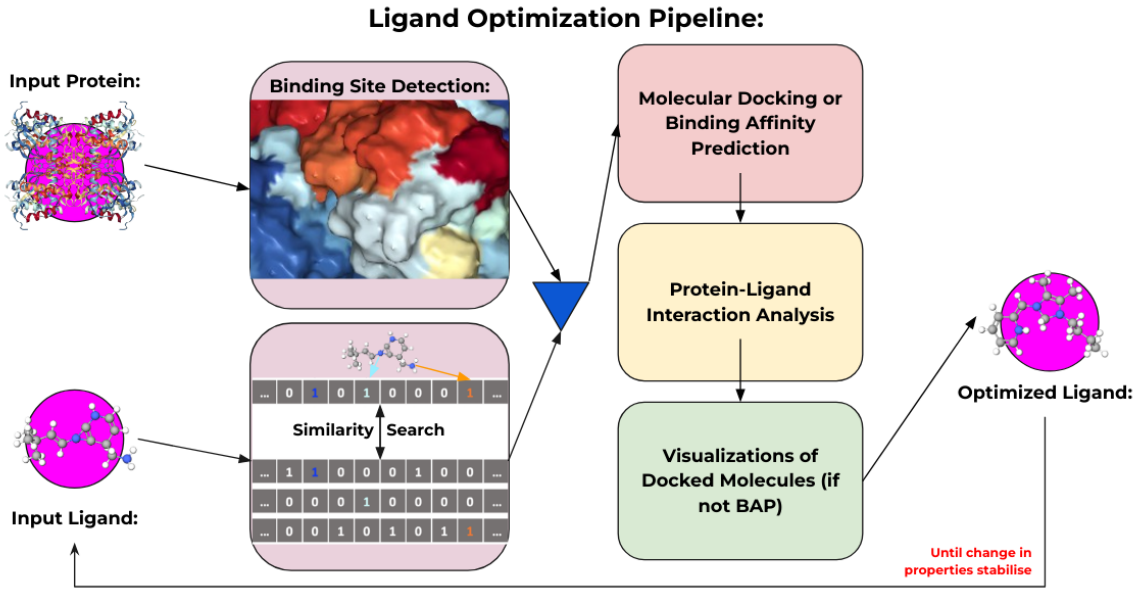
$$r_m^2 = r^2\left(1 - \sqrt{\frac{r^2 - r_0^2}{1 - r_0^2}}\right), \tag{12}$$

where r^2 and r_0^2 are the squared correlation coefficients with and without intercepts, respectively.

3.4.3 Computational Experiments and Results

The binding affinity prediction experiments were conducted on a single fold due to constraints related to time and computational resources, specifically using a laptop Nvidia RTX 3070 GPU for training. The network was trained for 100 epochs, utilizing both PyTorch and PyTorch Geometric to handle graph-based operations. The full code implementation and detailed setup are available on the project's GitHub repository. This setup allowed for the generation of comprehensive loss plots and correlation metrics comparing the ground truth data with predicted values. Additionally, comparative analyses against baseline models were performed to validate the enhancements made in the modified binding affinity predictor.

3.5 Optimization Pipeline Integration with TeachOpenCADD



(a) Diagram of the modified ligand optimization pipeline from TeachOpenCADD [46] with modifications implemented.

The optimization pipeline for drug discovery involves several computational and bioinformatics tools aimed at refining the efficacy and selectivity of potential drug candidates (see appendix). Utilizing the workflows suggested by Volkamer Lab's TeachOpenCADD [33], we adapt a streamlined approach to automate the structure-based virtual screening process. The primary adaptation involves replacing traditional docking calculations with a modified binding affinity predictor network, which significantly reduces computational costs and accelerates throughput [53].

Our process begins with the extraction of SMILES strings of druggable ligands from the pool of generated molecules. These SMILES strings are then processed to identify and prioritize compounds that exhibit potential for high binding affinity and optimal drug-like properties.

The optimization pipeline involves several key stages, starting with Binding Site Detection, where the most druggable binding sites on the target protein are identified using geometric and chemical features of the protein structure using DoGSiteScorer [54, 48]. Following this, Ligand Modification and Selection takes place, where structural analogs of the initial hit or lead compound are computationally generated and filtered based on their drug-likeness scores and molecular properties [55]. In the Binding Affinity Prediction stage, traditional molecular docking is replaced by the modified binding affinity predictor. This network utilizes the features of both the ligand and the protein to predict binding affinities directly, as expressed in the equation:

$$Y = \text{FFN}_{\text{modified}}(\text{concat}(z_{l1}, z_{l2}, z_p)), \quad (13)$$

where Y outputs not only the binding affinity but also distances that reflect the quality of the binding pose, providing a comprehensive view of the interaction dynamics. Finally, during the Analysis and Selection phase, the predicted interactions and binding modes are analyzed to select ligands that likely exhibit the best therapeutic profiles [50, 56].

This optimized pipeline is fully compatible with Python-based computational platforms, enhancing the utilization of chemical informatics libraries like RDKit, and deep learning frameworks such as TensorFlow and PyTorch. This integration moves beyond traditional command-line docking programs like AutoDock GPU or Smina, although these remain supported within the suggested workflow [49]. The outputs generated by this pipeline include ligand structures refined for enhanced affinity and selectivity, accompanied by comprehensive interaction profiles. These outputs significantly contribute to the meticulous refinement of potential drug candidates, advancing their development through detailed biochemical interactions analysis [57].

4 Results and Discussion

4.1 Overview of the Automated System

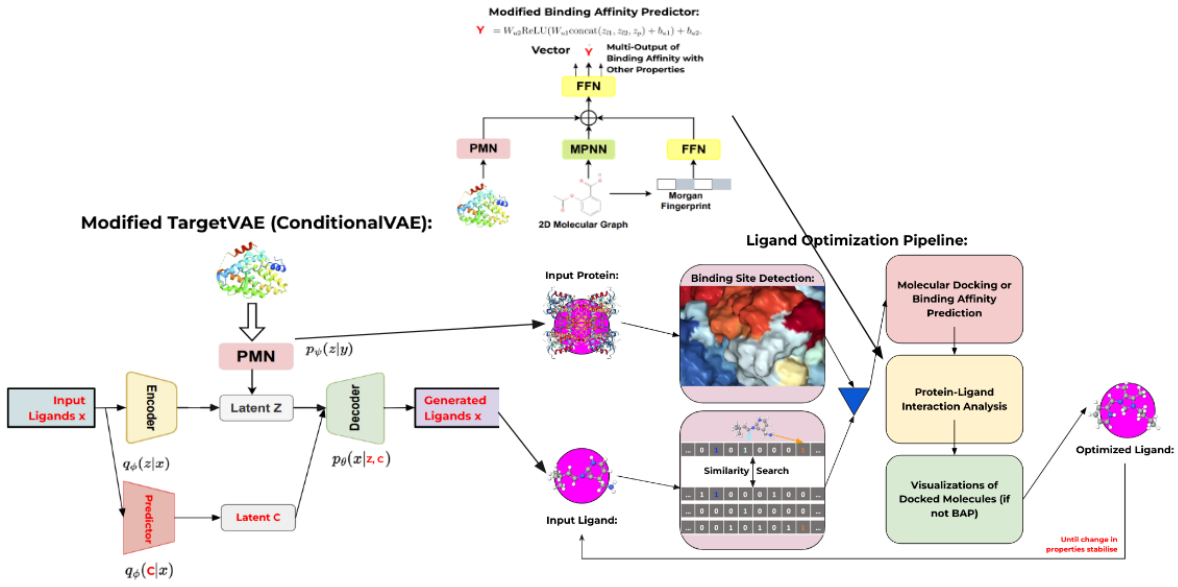


Figure 4: Schematic representation of the automated drug discovery system. This diagram highlights the integration of different computational stages from target selection, through ligand generation, to the final optimization pipeline.

4.2 Target Selection and Computational Representation

In the initial phase of our drug discovery pipeline, the target protein FGFR-2 is identified as a crucial element for further analysis [10]. The process involves extracting information from the Protein Data Bank (PDB) using a modified script that maps the UniProt ID to its corresponding PDB entry [58], followed by downloading and processing the PDB file. The function `fetch_and_save_pdb_file(mapped_protein_id, protein_pdb_path)`, handles the retrieval and storage of protein data. Once acquired, the protein data is transformed into a graph representation using the function `featurize_as_graph(protein_df, device=self.args.device)`.

4.3 Analysis of Generated Druggable vs. Non-Druggable Ligands

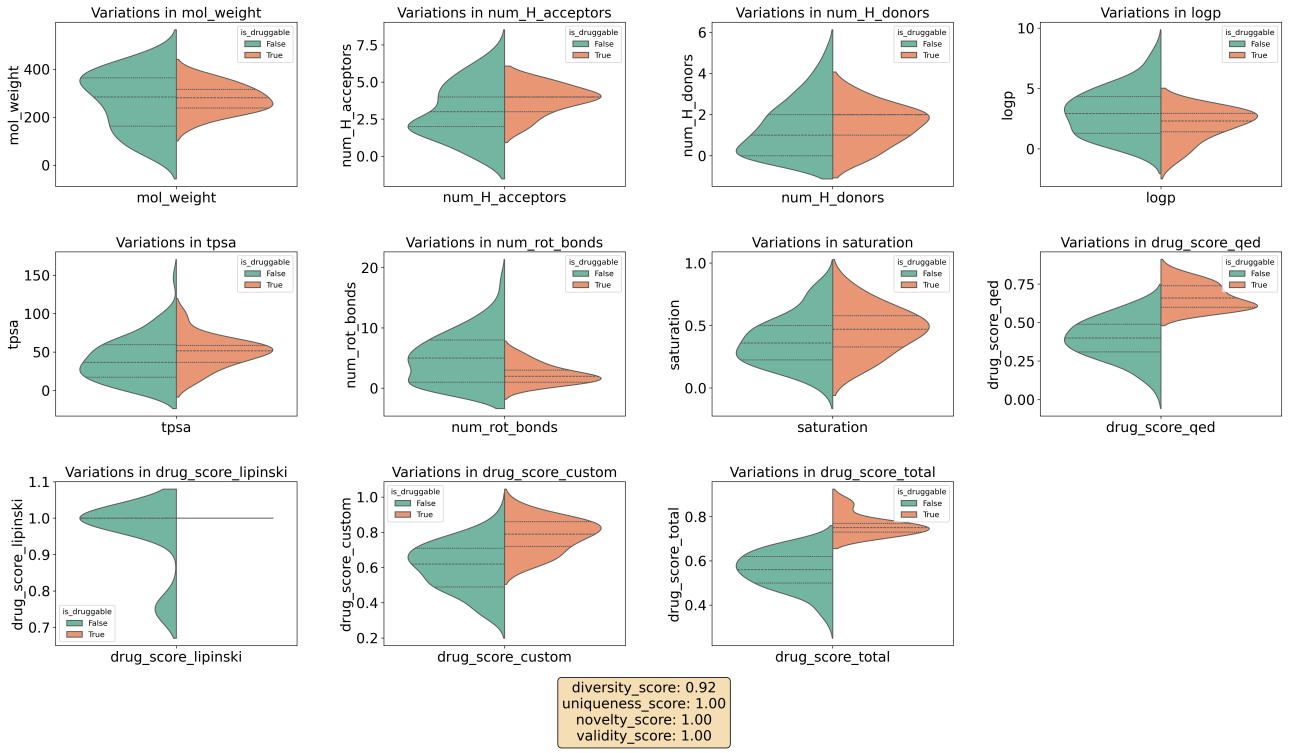


Figure 5: Distribution of molecular properties including Molecular Weight (MolWt), Number of Hydrogen Acceptors (NumHAcceptors), Number of Hydrogen Donors (NumHDonors), LogP (logP), Topological Polar Surface Area (TPSA), Number of Rotatable Bonds (NumRotBonds), and Saturation for druggable and non-druggable molecules.

Upon generating 1000 molecular entities via the ConditionalVAE model, we determine the druggability of the output molecules based on their calculated properties (inferred by their molecular structures). The properties evaluated included molecular weight ($MolWt$), number of hydrogen acceptors ($NumHAcceptors$), donors ($NumHDonors$), logP ($logP$), topological polar surface area ($TPSA$), number of rotatable bonds ($NumRotBonds$), and saturation ($Saturation$). Each property was calculated using the function `calculate_druglikeness(mol_obj)`, ensuring a comprehensive assessment of each molecule's potential as a drug candidate. The distribution of these properties was then plotted, highlighting the variations between molecules deemed druggable and non-druggable based on established thresholds and scores such as QED and Lipinski's rule of five [45].

The molecular diversity within the generated set was quantified using various metrics, focusing particularly on drug-likeness (QED), Lipinski, and custom drug scores. Druggable molecules exhibited a narrower distribution in molecular weight, ranging from 150 to 400 Daltons, compared to non-druggable ones, which showed a broader range up to 575 Daltons. This pattern was mirrored across other properties, indicating a distinct profile for molecules that potentially qualify as drug candidates. The diversity, uniqueness, and validity scores for these molecules were computed, yielding high values, which underscores the model's ability to generate structurally diverse and valid molecular structures.

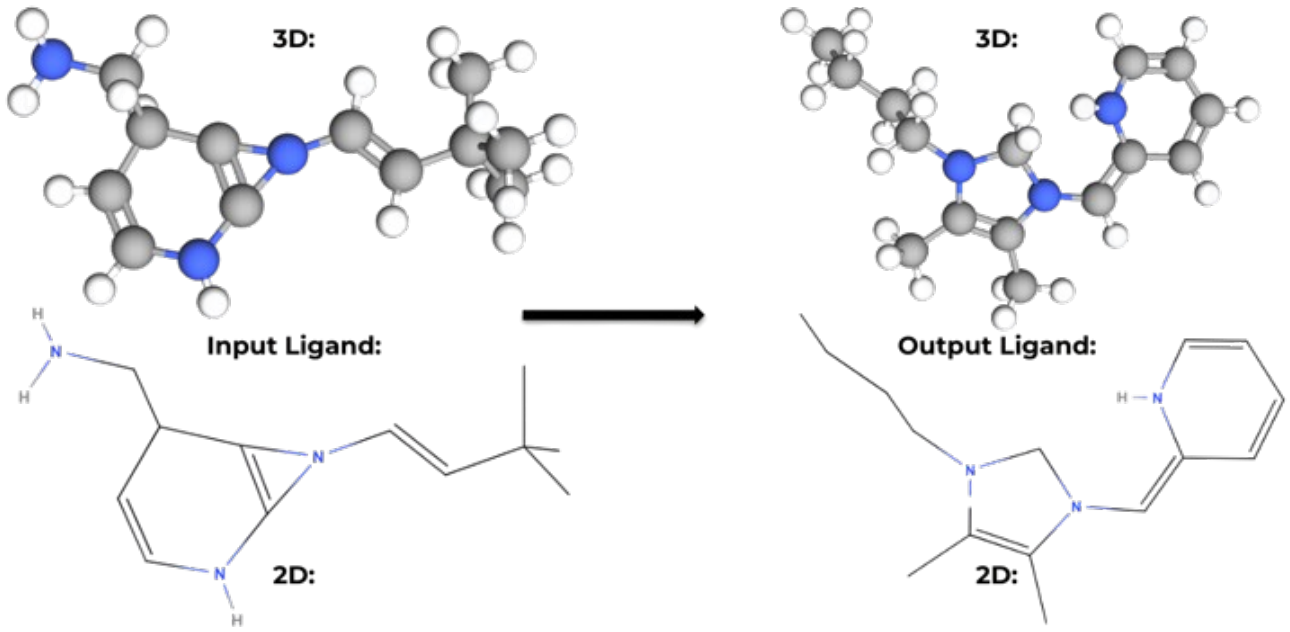
4.4 Optimizing Generated Ligands Using Modified TeachOpenCADD Pipeline

Upon generating druggable molecules from the ConditionalVAE, we transition to optimizing these ligands using a streamlined TeachOpenCADD pipeline [46], adapted for FGFR-2 as the target protein. The first stage involves the conversion of SMILES strings into PyMOL objects for ligands. These objects are then featurized using the function `featurize_as_graph` and subsequently processed through a series of computational steps aimed at enhancing their therapeutic potential.

During the pipeline, a `LeadOptimizationPipeline` class orchestrates the workflow, beginning with `BindingSiteDetection`, which identifies the most druggable sites on the protein surface [59]. This is followed by `LigandSimilaritySearch` and `Docking` processes, where ligands are compared against known databases and docked to the protein to predict binding affinities. Importantly, this process integrates a binding affinity prediction model that replaces the traditional, computationally expensive docking simulations, offering significant computational savings [53].

The optimization continues with `InteractionAnalysis`, which evaluates the interactions between the ligands and the protein at the predicted binding sites, followed by selecting the best analogs based on these interactions. Each ligand's performance is quantitatively assessed using various metrics, such as binding affinity and the total number of interactions, aiding in the selection of the most promising candidates for further development.

The improved structure of the input generated molecule is shown comparatively below, and it is contrasted with the optimized ligand. The tables detailing the properties, predicted binding affinities, and molecular characteristics of each molecule are included in the appendix. The enhanced molecular properties and improved binding affinity confirm the success of our pipeline. This comparison highlights similar atomic components but a significantly longer chain in the FDA-approved drug. Further training or a focus on model parameterization might have yielded different outcomes.



(a) Molecular representation of the input generated ligand and the output ligand of the pipeline. Input molecular smiles string: CC(C)(C)C=CN1C2=C1NC=CC2CN output molecular smiles string CCCCCN1CN(C(=C1C)C)C=C2C=CC=CN2.

4.5 Discussion

The modifications introduced to the original TargetVAE model and the TeachOpenCADD pipeline have showcased nuanced impacts on the drug discovery process. Although the integration of a ConditionalVAE aimed to enhance the specificity and relevance of generated molecules, the proportion of druggable molecules remained relatively low. This suggests that the training and inference processes could benefit from incorporating target-specific strategies, such as those used in the GENTRL paper, which actively leverages reinforcement learning to optimize its latent space, potentially refining the molecular generation process to

ensure a higher yield of candidate molecules with desired properties [60].

While the pipeline demonstrated the capability to generate and optimize ligands, the drugs produced did not match the efficacy of current FDA-approved drugs. This was apparent from the structural differences observed between the optimized ligand, CCCCCN1CN(C(=C1C)C)C=C2C=CC=CN2, and the FDA-approved drug Infigratinib, CCN1CCN(CC1)C2=CC=C(C=C2)NC3=CC(=NC=N3)N(C)C(=O)NC4=C(C(=CC(=C4Cl)OC)OC)Cl. A key observation was the simplicity in the bond structures and lower complexity in heteroatoms of the optimized ligand compared to Infigratinib, which features multiple aromatic rings and heteroatoms essential for higher binding efficacy [61]. Effective parameterization, such as adjusting the model to predict more complex molecular frameworks with diverse functional groups, could have significantly improved outcomes. This includes refining parameters that govern bond angle predictions, aromaticity, and the presence of pharmacophore features [62] which are critical for drug-receptor interactions.

Future iterations could benefit from integrating insights from advanced structural prediction models like AlphaFold, developed by DeepMind [52], to better predict how complex molecular structures fold and interact with biological targets. By adjusting model parameters to enhance the generation of molecules with specific bond types, angles, and necessary chemical groups, it's possible to bridge the gap between the generated molecules and high-efficacy drugs like Infigratinib. These changes could lead to the generation of molecules not only structurally similar to FDA-approved drugs but also functionally capable of binding effectively to target proteins, reflecting a deeper understanding of molecular dynamics and interactions critical for drug development.

On the scalability of these approaches, the modified pipeline exhibits a high degree of adaptability and potential applicability across various targets and diseases. This adaptability is crucial in the context of deep learning-driven drug discovery, where the ability to generalize across different biological contexts can significantly expedite the development process. However, scalability is contingent upon overcoming substantial computational bottlenecks encountered during analog search and feature extraction processes, suggesting a potential improvement could be the local caching of compound databases or employing a more robust server-side computation scheme to handle these tasks efficiently [63].

The primary challenge was the integration and computational efficiency of searching for analogs within extensive chemical databases like PubChem, which was both time-consuming and resource-intensive. Future iterations of the project could benefit from streamlined database interactions or enhanced computational strategies, possibly incorporating more efficient algorithms or hardware solutions. Despite these challenges, the project's success in identifying a candidate for FGFR-2 interaction highlights the potential of deep learning approaches in revolutionizing drug discovery, supporting continued development and optimization of these computational tools for broader applications in biomedicine.

5 Conclusions and Future Directions

The exploration and development of FGFR-2 targeting drug discovery using a modified TeachOpenCADD pipeline and ConditionalVAE have shown significant potential and identified areas for enhancement. While the project has effectively demonstrated the integration of machine learning to generate and optimize ligands, the complexity and efficacy of the resultant molecules still fall short of FDA-approved standards like Infigratinib. Key future directions include refining model parameterization to predict more intricate molecular architectures and employing advanced structural prediction techniques like AlphaFold to better model interactions with biological targets. Furthermore, addressing computational bottlenecks, such as through local caching of compound databases or utilizing high-performance computing solutions, is essential for improving scalability and efficiency. Continuous evolution of AI techniques promises to revolutionize traditional drug discovery processes, expediting the development of new therapies and broadening the application of these computational tools across various biomedical domains.

6 Declarations

Acknowledgements

I express my sincere gratitude to Professor Emad Moeendarbary for his invaluable guidance on the structure of this work and for providing advice for refining the presentation of our research.

Use of AI Tools

AI tools GPT-4 and GPT-3.5 (from OpenAI) were used for rewording of sentences and providing initial structure to sections to aide brainstorming.

References

[1] Y. Nakanuma et al. “Microstructure and Development of the Normal and Pathologic Biliary Tract in Humans, Including Blood Supply”. In: (). DOI: 10.1002/(SICI)1097-0029(19970915)38:6<552::AID-JEMT2>3.0.CO;2-H.

[2] Jesus M. Banales et al. “Cholangiocarcinoma 2020: The next Horizon in Mechanisms and Management”. In: (). DOI: 10.1038/s41575-020-0310-z.

[3] Supot Kamsa-ard et al. “Cholangiocarcinoma Trends, Incidence, and Relative Survival in Khon Kaen, Thailand From 1989 Through 2013: A Population-Based Cancer Registry Study”. In: (). DOI: 10.2188/jea.JE20180007.

[4] Masato Nagino et al. “Clinical Practice Guidelines for the Management of Biliary Tract Cancers 2019: The 3rd English Edition”. In: (). DOI: 10.1002/jhbp.870.

[5] Ningyu Wang et al. “Progress in Radiotherapy for Cholangiocarcinoma”. In: (). DOI: 10.3389/fonc.2022.868034.

[6] Sedighe Hosseini Shabanan et al. “Selective Internal Radiation Therapy with Yttrium-90 for Intrahepatic Cholangiocarcinoma: A Systematic Review on Post-Treatment Dosimetry and Concomitant Chemotherapy”. In: (6). DOI: 10.3390/curroncol29060306.

[7] Do-Youn Oh et al. “Durvalumab plus Gemcitabine and Cisplatin in Advanced Biliary Tract Cancer”. In: (). DOI: 10.1056/EVIDoa2200015.

[8] Hiromi Nakamura et al. “Genomic Spectra of Biliary Tract Cancer”. In: (). DOI: 10.1038/ng.3375.

[9] Apurva Jain and Milind Javle. “Molecular Profiling of Biliary Tract Cancer: A Target Rich Disease”. In: (). DOI: 10.21037/jgo.2016.09.01.

[10] Yuta Maruki et al. “Molecular Detection and Clinicopathological Characteristics of Advanced/Recurrent Biliary Tract Carcinomas Harboring the FGFR2 Rearrangements: A Prospective Observational Study (PRELUDE Study)”. In: (). DOI: 10.1007/s00535-020-01735-2.

[11] Emma Lodl et al. “Updates in the Use of Targeted Therapies for the Treatment of Cholangiocarcinoma”. In: (). DOI: 10.1177/10781552231171079.

[12] Isabella Gashaw et al. “What Makes a Good Drug Target?” In: (). DOI: 10.1016/j.drudis.2011.12.008.

[13] J. P. Hughes et al. “Principles of Early Drug Discovery”. In: (). DOI: 10.1111/j.1476-5381.2010.01127.x.

[14] Richard C. Mohs and Nigel H. Greig. “Drug Discovery and Development: Role of Basic Biological Research”. In: (). DOI: 10.1016/j.trci.2017.10.005.

[15] Natesh Singh et al. “Drug Discovery and Development: Introduction to the General Public and Patient Groups”. In: (). DOI: 10.3389/fddsv.2023.1201419.

[16] Joseph A. DiMasi. “Research and Development Costs of New Drugs”. In: (). DOI: 10.1001/jama.2020.8648.

[17] Debleena Paul et al. “Artificial Intelligence in Drug Discovery and Development”. In: (). DOI: 10.1016/j.drudis.2020.10.010.

[18] Francois Pognan et al. “The Evolving Role of Investigative Toxicology in the Pharmaceutical Industry”. In: (). DOI: 10.1038/s41573-022-00633-x.

[19] Petra Schneider et al. “Rethinking Drug Design in the Artificial Intelligence Era”. In: (). DOI: 10.1038/s41573-019-0050-3.

[20] Madura K. P. Jayatunga et al. “AI in Small-Molecule Drug Discovery: A Coming Wave?” In: (). DOI: 10.1038/d41573-022-00025-1.

[21] Maria Kontoyianni. “Library Size in Virtual Screening: Is It Truly a Number’s Game?” In: (). DOI: 10.1080/17460441.2022.2130244.

[22] Simone Brogi and Vincenzo Calderone. “Artificial Intelligence in Translational Medicine”. In: (3). DOI: 10.3390/ijtm1030016.

[23] Jeffrey A. Ruffolo, Jeremias Sulam, and Jeffrey J. Gray. “Antibody Structure Prediction Using Interpretable Deep Learning”. In: (). DOI: 10.1016/j.patter.2021.100406.

[24] Anastasiia V. Sadybekov and Vsevolod Katritch. “Computational Approaches Streamlining Drug Discovery”. In: (). DOI: 10.1038/s41586-023-05905-z.

[25] Jiho Yoo et al. “Industrializing AI/ML during the End-to-End Drug Discovery Process”. In: (). DOI: 10.1016/j.sbi.2023.102528.

[26] Yunjiang Zhang et al. “Universal Approach to De Novo Drug Design for Target Proteins Using Deep Reinforcement Learning”. In: (). DOI: 10.1021/acsomega.2c06653.

[27] Naitik Jariwala et al. “Intriguing of Pharmaceutical Product Development Processes with the Help of Artificial Intelligence and Deep/Machine Learning or Artificial Neural Network”. In: (). DOI: 10.1016/j.jddst.2023.104751.

[28] Joseph Chon et al. *Automation and Machine Learning: A Look into Recursion Pharmaceuticals*. DOI: 10.13140/RG.2.2.19921.43368.

[29] Ying Yang et al. “Efficient Exploration of Chemical Space with Docking and Deep Learning”. In: (). DOI: 10.1021/acs.jctc.1c00810.

[30] Brian Goldman et al. “Defining Levels of Automated Chemical Design”. In: (). DOI: 10.1021/acs.jmedchem.2c00334.

[31] *Combining Cloud-Based Free-Energy Calculations, Synthetically Aware Enumerations, and Goal-Directed Generative Machine Learning for Rapid Large-Scale Chemical Exploration and Optimization | Journal of Chemical Information and Modeling*.

[32] Nhat Khang Ngo and Truong Son Hy. *Target-Aware Variational Auto-encoders for Ligand Generation with Multimodal Protein Representation Learning*. DOI: 10.1101/2023.08.10.552868.

[33] Dominique Sydow et al. *TeachOpenCADD 2021: Open Source and FAIR Python Pipelines to Assist in Structural Bioinformatics and Cheminformatics Research*. DOI: 10.26434/chemrxiv-2021-8x13n.

[34] Masaki Kuwatani and Naoya Sakamoto. “Promising Highly Targeted Therapies for Cholangiocarcinoma: A Review and Future Perspectives”. In: (14). DOI: 10.3390/cancers15143686.

[35] Ranish K. Patel and Flavio G. Rocha. “Role of Genomics in Intrahepatic Cholangiocarcinoma for Liver-Directed Therapy”. In: (). DOI: 10.1016/j.surg.2023.04.009.

[36] Andrea Cercek et al. “Assessment of Hepatic Arterial Infusion of Floxuridine in Combination With Systemic Gemcitabine and Oxaliplatin in Patients With Unresectable Intrahepatic Cholangiocarcinoma: A Phase 2 Clinical Trial”. In: (). DOI: 10.1001/jamaoncol.2019.3718.

[37] Frank W. Pun et al. “Identification of Therapeutic Targets for Amyotrophic Lateral Sclerosis Using PandaOmics – An AI-Enabled Biological Target Discovery Platform”. In: (). DOI: 10.3389/fnagi.2022.914017.

[38] Justin Gilmer et al. *Neural Message Passing for Quantum Chemistry*. DOI: 10.48550/arXiv.1704.01212.

[39] Mahmood Kalemati, Mojtaba Zamani Emani, and Somayyeh Koohi. “BiComp-DTA: Drug-target Binding Affinity Prediction through Complementary Biological-Related and Compression-Based Featurization Approach”. In: (). DOI: 10.1371/journal.pcbi.1011036.

[40] Jiaqi Guan et al. *3D Equivariant Diffusion for Target-Aware Molecule Generation and Affinity Prediction*. DOI: 10.48550/arXiv.2303.03543.

[41] Tim van Erven and Peter Harremoës. “R\’enyi Divergence and Kullback-Leibler Divergence”. In: (). DOI: 10.1109/TIT.2014.2320500.

[42] Xiaochu Tong et al. “Generative Models for De Novo Drug Design”. In: (). DOI: 10.1021/acs.jmedchem.1c00927.

[43] *KiBA - a benchmark dataset for drug target prediction*. Helsingin yliopisto.

[44] Kristina Preuer et al. “Fréchet ChemNet Distance: A Metric for Generative Models for Molecules in Drug Discovery”. In: (). DOI: 10.1021/acs.jcim.8b00234.

[45] *Lipinski’s Rule of Five - an Overview | ScienceDirect Topics*.

[46] *T018 · Automated Pipeline for Lead Optimization*. TeachOpenCADD.

[47] Bowen Jing et al. *Learning from Protein Structure with Geometric Vector Perceptrons*. DOI: 10.48550/arXiv.2009.01411.

[48] Gregory L. Warren et al. “A Critical Assessment of Docking Programs and Scoring Functions”. In: (). DOI: 10.1021/jm050362n.

[49] Nataraj S. Pagadala, Khajamohiddin Syed, and Jack Tuszynski. “Software for Molecular Docking: A Review”. In: (). DOI: 10.1007/s12551-016-0247-1.

[50] Sebastian Salentin et al. “PLIP: Fully Automated Protein–Ligand Interaction Profiler”. In: (). DOI: 10.1093/nar/gkv315.

[51] Hakime Öztürk, Elif Ozkirimli, and Arzucan Özgür. *WideDTA: Prediction of Drug-Target Binding Affinity*. DOI: 10.48550/arXiv.1902.04166.

[52] John Jumper et al. “Highly Accurate Protein Structure Prediction with AlphaFold”. In: (). DOI: 10.1038/s41586-021-03819-2.

[53] J. S. Smith, O. Isayev, and A. E. Roitberg. “ANI-1: An Extensible Neural Network Potential with DFT Accuracy at Force Field Computational Cost”. In: (). DOI: 10.1039/C6SC05720A.

[54] Andrea Volkamer et al. “Combining Global and Local Measures for Structure-Based Druggability Predictions”. In: (). DOI: 10.1021/ci200454v.

[55] Adrià Cereto-Massagué et al. “Molecular Fingerprint Similarity Search in Virtual Screening”. In: Virtual Screening (). DOI: 10.1016/j.ymeth.2014.08.005.

[56] Renato Ferreira de Freitas and Matthieu Schapira. “A Systematic Analysis of Atomic Protein–Ligand Interactions in the PDB”. In: (). DOI: 10.1039/C7MD00381A.

[57] Xing Du et al. “Insights into Protein–Ligand Interactions: Mechanisms, Models, and Methods”. In: (2). DOI: 10.3390/ijms17020144.

[58] *Programmatic Access - Retrieving Entries via Queries | UniProt Help | UniProt*.

[59] Andrea Volkamer et al. “Prediction, Analysis, and Comparison of Active Sites”. In: *Applied Chemoinformatics*. John Wiley & Sons, Ltd. ISBN: 978-3-527-80653-9. DOI: 10.1002/9783527806539.ch6g.

[60] Alex Zhavoronkov et al. “Deep Learning Enables Rapid Identification of Potent DDR1 Kinase Inhibitors”. In: (). DOI: 10.1038/s41587-019-0224-x.

[61] Suleiman A. Khan et al. “Identification of Structural Features in Chemicals Associated with Cancer Drug Response: A Systematic Data-Driven Analysis”. In: (). DOI: 10.1093/bioinformatics/btu456.

[62] Deborah Giordano et al. “Drug Design by Pharmacophore and Virtual Screening Approach”. In: (). DOI: 10.3390/ph15050646.

[63] *Computational Approaches Streamlining Drug Discovery | Nature*.

Appendix for Automated Target Selection and De Novo Drug Generation For Intrahepatic Cholangiocarcinomas

Priestly Adejo Student Number: 21001232
[1] Department of Mechanical Engineering, University College London,
Torrington Place, London, WC1E 7JE, UK

1 Introduction

For more details and to contribute to this evolving project, visit our GitHub repository at `https://github.com/PriestlyAdejo/drug_discovery_cholangiocarcinoma`. All the code used can be found here!

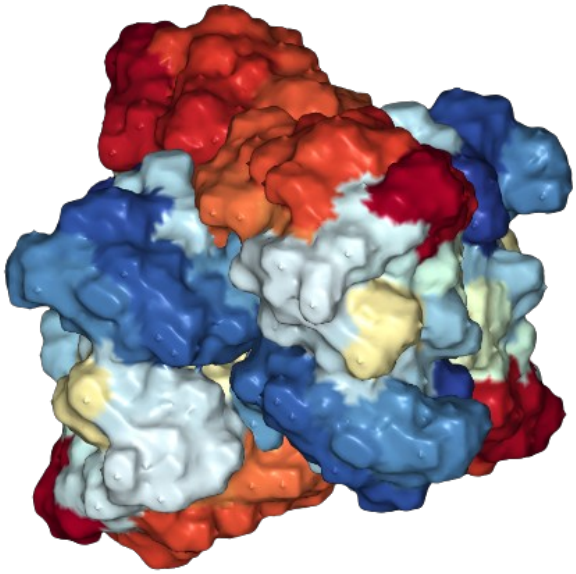


Figure 2: Molecular graphics of FGFR-2 showing binding sites and important functional domains essential for its activity in intrahepatic cholangiocarcinoma.

2 Methodology

2.1 Target Protein Selection: FGFR-2

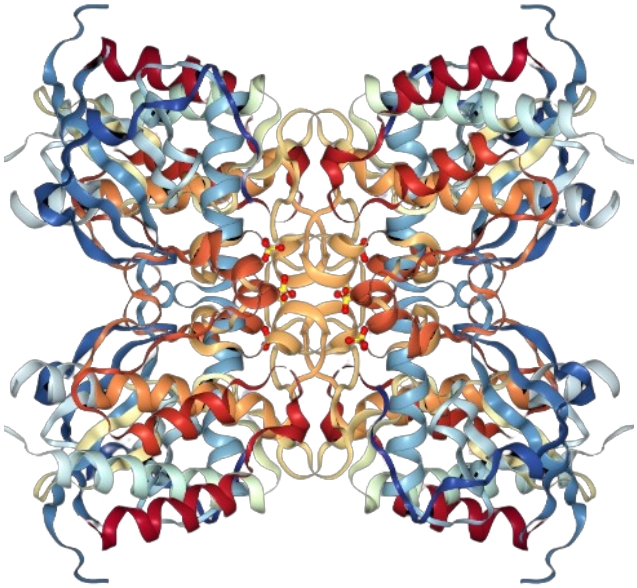


Figure 1: 3D structural ribbon representations of FGFR-2, highlighting key molecular features and domains.

2.2 Computational Representation of Target Protein

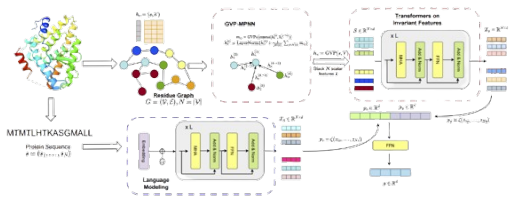


Figure 3: Architecture of the Protein Modal Network (PMN) as proposed in the Target VAE study by Ngo and Hy.

2.3 Binding Affinity Prediction with Modified Binding Affinity Predictor

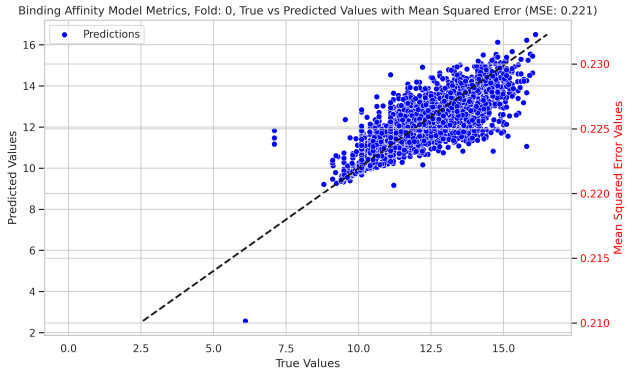


Figure 4: Plot of mean squared error and correlation in predicting different binding affinity values after training the model for 100 epochs.

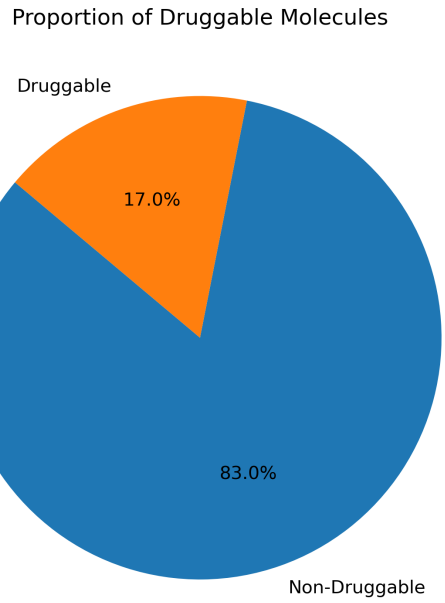


Figure 5: Pie chart showing proportion of molecules that were actually druggable according to aforementioned characteristics of druggable molecules.

3 Results and Discussion

3.1 Analysis of Generated Druggable vs. Non-Druggable Ligands

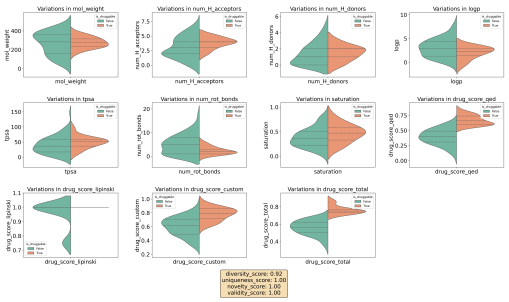


Figure 6: Distribution of molecular properties including Molecular Weight (MolWt), Number of Hydrogen Acceptors (NumHAcceptors), Number of Hydrogen Donors (NumHDonors), LogP (logP), Topological Polar Surface Area (TPSA), Number of Rotatable Bonds (NumRotBonds), and Saturation for druggable and non-druggable molecules. This visualization highlights the contrast in key molecular descriptors that define potential drug candidates versus those less likely to be viable drugs, based on established thresholds and scores such as QED and Lipinski's rule of five.

3.2 Optimizing Generated Ligands Using Modified TeachOpenCADD Pipeline

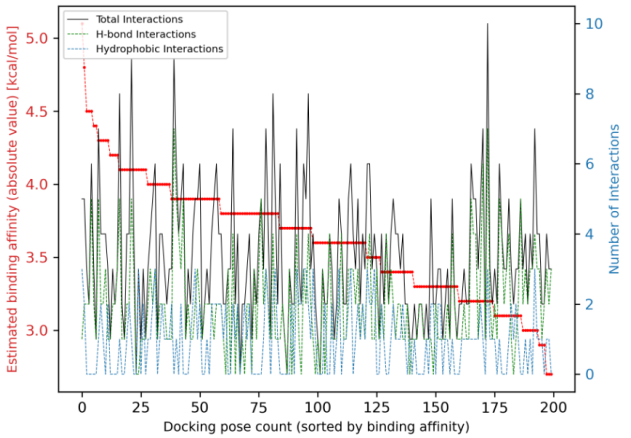


Figure 7: Interaction analysis of the optimized ligands with the FGFR-2 protein target. Visual representations of ligand-protein interactions highlighting key binding sites and interaction strengths, which were crucial for selecting the most promising drug candidates.

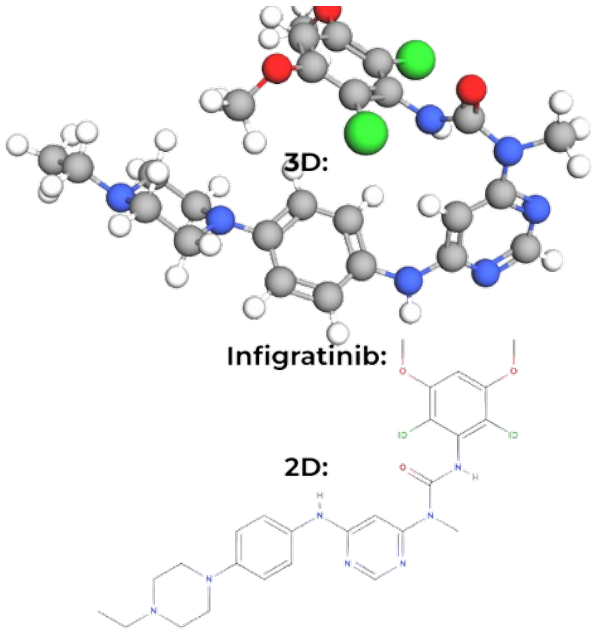


Figure 8: Molecular representation of existing FGFR-2 inhibitor, Infigratinib, showcasing its molecular structure, with a smiles string of CCN1CCN(CC1)C2=CC=C(C=C2)NC3=CC(=NC=N3)N(C)C(=O)NC4=C(C(=CC(=C4Cl)OC)OC)Cl.

Table 1: Binding Affinity and RMSD Metrics for Input Ligand

Mode	Affinity (kcal/mol)	Dist from Best Mode rmsd_l.b	Dist from Best Mode rmsd_u.b
1	-4.8	0.000	0.000
2	-4.6	1.362	1.895
3	-4.6	1.304	2.580
4	-4.3	2.120	6.058
5	-3.9	2.418	6.407

Table 2: Binding Affinity and RMSD Metrics for Optimized Ligand

Mode	Affinity (kcal/mol)	Dist from Best Mode rmsd_l.b	Dist from Best Mode rmsd_u.b
1	-5.1	0.000	0.000
2	-4.8	2.680	5.897
3	-4.5	1.996	6.627
4	-4.5	2.078	6.016
5	-4.4	2.280	6.583

Table 3: Comparative Molecular Properties between Input and Optimized Ligands

Property	**Input Ligand**	**Optimized Ligand**
Smiles	CC(C)(C)C=CN1C2=C1NC=CC2CN	CCCCCN1CN(C(=C1C)C)C=C2C=CC=CN2
Num. Rot. Bonds	2	4
Num. H Donors	2	1
Num. H Acceptors	3	3
Mol. Weight	205.305	245.37
LogP	1.72	3.13
Drug Score Total	0.8	0.86
Drug Score QED	0.72	0.82
Drug Score Lipinski	1.0	1.0
Drug Score Custom	0.82	0.85
Saturation	0.5	0.47
TPSA	41.06	18.51
Druggable Smiles	True	True