UNIVERSITAT DE BARCELONA

FUNDAMENTAL PRINCIPLES OF DATA SCIENCE MASTER'S THESIS

Pocket-Aware Molecular Generation Through Learned Protein Representations

Author:
Clàudia VALVERDE

Supervisor:
Laura IGUAL
Alexis MOLINA

A thesis submitted in partial fulfillment of the requirements for the degree of MSc in Fundamental Principles of Data Science

in the

Facultat de Matemàtiques i Informàtica

June 30, 2025

UNIVERSITAT DE BARCELONA

Abstract

Facultat de Matemàtiques i Informàtica

MSc

Pocket-Aware Molecular Generation Through Learned Protein Representations

by Clàudia VALVERDE

Drug discovery is constrained not only by the immense chemical space but by the difficulty of efficiently exploring it and the high cost of traditional screening methods. This thesis introduces and evaluates a deep learning (DL) strategy for the de novo generation of small molecules designed to bind specific protein pockets, aiming to accelerate the identification of novel drug candidates. Our approach leverages pre-trained protein and pocket embeddings within a decoder-only Transformer architecture that learns to translate complex biological information into SMILES strings.

Given the early stage of conditional binder generation, this work emphasizes systematic experimentation and thorough performance evaluation. We explored various protein and pocket representation strategies, including global protein (ESM2), structural-aware protein (SaProt), pocket-specific (PickPocket), and integrated Drug-Target Interaction (TensorDTI) embeddings.

Our comprehensive evaluation pipeline assessed molecule validity, novelty, internal and cross-model diversity, physicochemical properties, and predicted drug-target interactions. Key findings include demonstrating that a high proportion of viral proteins in the training data does not bias generation, and that different input representations guide the model to explore distinct chemical spaces. While the models effectively generate diverse molecules with favorable drug-like properties, a notable limitation is their propensity to produce exact matches to the training set, indicating overfitting. Furthermore, despite the model's sensitivity to pocket information, case studies of two specific kinase proteins revealed a challenge in consistently generating truly pocket-specific molecules, likely because of data set characteristics such as promiscuous motifs. This work provides valuable insights into the capabilities and current limitations of pocket-aware generative models, laying a foundation for future advancements in targeted molecule design.

Acknowledgements

I would like to express my sincere gratitude to my university tutor, Laura Igual. I am also deeply thankful to Alexis Molina for his continuous guidance and insightful advice, which have been invaluable from start to finish, and to Manel Gil for his constant support and readiness to help whenever needed.

I would like to extend my appreciation to all my colleagues for their collaboration and support during this journey.

As always, I am deeply grateful to my family and friends. Their presence, patience, and unwavering belief in me have been a constant source of strength and motivation.

Contents

Chapter 1

Introduction

1.1 Biological Motivation

In the fields of medicine, biotechnology, and pharmacology, drug discovery is the process to identify new candidate molecules capable of modulating biological systems to treat diseases. This typically begins with the selection of a biological target, (a protein, enzyme or a receptor), implicated in a specific pathological condition (Hughes et al., 2011). The next critical step involves identifying small molecules that can bind to this target and influence its function, either by inhibiting or enhancing its activity depending on the therapeutic objective (Thangudu et al., 2012).

One of the central challenges in this process is the immense scope and complexity of what is known as the chemical space. Consider, for instance, a molecule composed of just 10 atoms, each potentially one of 10 different elements, this alone presents roughly 10 billion unique combinations. When we expand this to drug-like molecules, which typically contain 20 or more atoms and involve intricate aspects like stereochemistry, conformational flexibility, and diverse bond patterns, the count of possible candidates escalates dramatically, reaching into the trillions and beyond (Rasul et al., 2024; Chakraborty, Kayastha, and Ramakrishnan, 2019). This immense landscape means that comprehensive experimental screening is simply unfeasible, even with sophisticated high-throughput methods, researchers can only investigate a tiny fraction of the potential chemical compounds.

Simultaneously, proteins are not static. They can adopt multiple conformations and dynamic states, each potentially influencing binding affinity and specificity in different ways (Miller and Phillips, 2021). This conformational flexibility adds another layer of complexity to the already complex process of identifying effective protein-drug interactions. A compound that binds effectively to one conformation might not bind to another, making it challenging to predict binding based on a single static protein structure. Furthermore, the binding process itself is a dynamic event, involving induced fit and conformational changes in both the protein and the ligand, adding another layer of complexity to accurately model and predict interactions (Greives and Zhou, 2014).

Traditional drug discovery methods, which rely heavily on high-throughput screening and trial-and-error experimentation, are consequently both time-intensive and costly. The hit rates from these screens are often very low, necessitating the synthesis and testing of hundreds of thousands, or even millions, of compounds to find a handful of initial 'hits' that require extensive further optimization. In response to these limitations, computational approaches, and in particular Artificial Intelligence

(AI), have emerged as powerful tools to accelerate and refine early-stage drug discovery (Kant, Roy, et al., 2025).

AI enables the rapid exploration of vast chemical spaces, focusing on the most promising candidates. It can model protein-drug interactions, predict binding affinities (Singh et al., 2023; Gil-Sorribes, Ciudad Serrano, and Molina, 2025), and even generate novel molecules that may never have been synthesized before, all while considering constraints such as drug-likeness, toxicity, and selectivity (Sadybekov and Katritch, 2023). By leveraging AI to better understand and predict protein-drug interactions, we can make the drug discovery process more targeted, efficient, and innovative, ultimately accelerating the development of new therapies.

1.2 Methodology Motivation

Several powerful AI methodologies have emerged for generating novel chemical compounds. Some generative models including Recurrent Neural Networks (RNNs) (Bjerrum and Threlfall, 2017) can generate molecules character-by-character based on SMILES, string representations of molecules, and Variational Autoencoders (VAEs) (Liu et al., 2018) or Generative Adversarial Networks (GANs) (Lin, Lin, and Lane, 2020) learn latent representations of molecules and can sample new compounds from this learned space. Another prominent approach utilizes Graph Neural Networks (GNNs) (Zhou et al., 2020), where molecules are represented as graphs, with atoms as nodes and bonds as edges. GNNs can learn complex relationships within molecular structures and generate new graphs corresponding to valid chemical compounds. These methods have demonstrated impressive capabilities in exploring chemical space and generating molecules with desired properties, such as drug-likeness and synthetic accessibility (Mak, Wong, and Pichika, 2024).

While these generative tools are powerful, a critical limitation for drug discovery is ensuring the functionality of the generated molecules, specifically their ability to bind to a protein of interest. Generating chemically valid and novel molecules is a necessary first step, but without targeted binding, these molecules are unlikely to modulate a biological process or treat a disease. The ultimate goal in rational drug design is to create molecules that not only exist but also exhibit a high affinity and specificity for a pre-defined biological target (Chen et al., 2023). Therefore, the challenge shifts from merely generating molecules to generating functional binders.

To bridge this gap, it is essential to integrate protein target information directly into the molecule generation process (Creanza et al., 2025). This is where Protein Language Models (PLMs) offer a transformative advantage (Lin et al., 2023; Su et al., 2023). PLMs, are deep learning models, often based on transformer architectures (similar to those used in natural language processing), that learn intricate patterns, relationships, and evolutionary constraints within protein sequences. Through pre-training on large protein databases, these models develop a sophisticated representation of protein structure and function.

Beyond full-sequence embeddings, more targeted representations, such as embeddings derived from protein binding pockets, can offer increased precision. These might be obtained using models that integrate both the protein's sequence and structural information (Zhang et al., 2023), allowing the generative model to focus on regions critical for molecular interaction and improving the likelihood of generating biologically active compounds that modulate the protein inhibition in a desired way.

1.3 Objectives

This work aims to develop and rigorously evaluate a deep learning strategy for the *de novo* generation of small molecules specifically designed to bind to target protein pockets. Our primary objectives are:

1. To develop a decoder-only model capable of generating SMILES strings, conditioned on comprehensive protein and pocket embedding information derived from various pre-trained sources.
2. To systematically benchmark the influence of dataset composition, specifically evaluating the potential bias introduced by a high proportion of viral protein families, on the molecule generation process.
3. To comprehensively assess the efficacy of different protein and pocket input representation types, including global protein embeddings (ESM2), structure-aware protein embeddings (SaProt), pocket-specific embeddings (PickPocket), and integrated Drug-Target Interaction (DTI) embeddings (TensorDTI).
4. To establish and apply a pipeline for the post-generation analysis of molecules, evaluating their validity, novelty relative to the training set, physicochemical properties, and predicted interaction with their corresponding target pockets.
5. To qualitatively assess the model's ability to generate distinct molecules when conditioned on different types of binding pockets (e.g., active vs. cryptic sites) in retrospective case studies.

1.3.1 Contributions

This thesis makes several key contributions to the field of conditional molecule generation:

1. **Development of a Conditional Generative Model:** We successfully implemented a Transformer decoder-only model capable of *de novo* SMILES generation, effectively conditioned by diverse pre-trained protein and pocket representations.
2. **Benchmark of Dataset Bias:** We conducted a systematic benchmarking study demonstrating that a disproportionate presence of viral protein families in the training dataset does not introduce a significant bias in the model's generative performance, ensuring broad applicability.
3. **Comparative Analysis of Input Representations:** Our work provides a comprehensive comparison of different protein and pocket embedding modalities (ESM2, SaProt, PickPocket, and TensorDTI), offering insights into their respective strengths and the impact of their fusion on the characteristics of generated molecules. We show that different modalities lead to diverse generated chemical spaces.
4. **Robust Evaluation Pipeline:** We established and applied a multi-faceted analytical pipeline to thoroughly characterize generated molecules, including

assessments of validity, novelty (against the training set), internal and cross-model diversity, physicochemical properties (including SA score), and predicted drug-target interactions.

5. **Assessment of Pocket-Specific Generation:** We conducted detailed retrospective case studies on proteins with distinct active and cryptic binding sites (e.g., CDK2 and RET) to assess the model's ability to generate molecules distinctively for specific pocket conditioning and their predicted interaction with these pockets.

1.3.2 Contents

This thesis is structured to provide a comprehensive exploration of conditional molecule generation using deep learning.

The Introduction first establishes the biological motivation, highlighting the critical need for novel pocket-specific molecule binders to address challenges in drug discovery. This is followed by the methodological motivation, which underscores the availability of advanced AI tools for molecule generation but emphasizes the necessity for systematic benchmarking when leveraging diverse pre-trained models.

The Background section provides foundational knowledge, starting with an introduction to proteins and small molecules for readers unfamiliar with these biological entities. It then delves into how these complex biological and chemical data are processed within Deep Learning (DL) algorithms, detailing their numerical representations. This section also reviews established molecule generation techniques and relevant Drug-Target Interaction (DTI) prediction models, setting the stage for the proposed methodology.

The Experimental Setup describes the practical implementation of our research. It covers the characteristics of the dataset utilized, including key observations from its initial analysis. Following this, the section elaborates on the preparation of input representations, the architecture of the custom deep learning model developed for this thesis, and the precise methodology employed for the molecule generation process. Finally, it outlines the inference setup and the comprehensive evaluation pipeline used to assess the quality and relevance of the generated compounds.

The Results section is dedicated to presenting and interpreting the findings derived from the experiments, covering the benchmarks and analyses performed. This leads to the Conclusions, which summarize the main outcomes of the thesis and discuss their implications for future research in conditional drug design.

The GitHub repository of this thesis can be found in here.

Chapter 2

Background

2.1 Proteins Fundamentals

Amino acids are often referred to as the building blocks of life, as they are the fundamental components of proteins. Proteins are formed through the linear polymerization of amino acids in a specific sequence, which is determined by the genetic code (Lopez and Mohiuddin, 2024). This precise sequence dictates how the protein folds into a unique three-dimensional structure, which in turn determines the protein's physicochemical properties and biological function.

The folding process creates distinct structural features on the protein surface, including binding pockets. These are specific cavities or grooves typically formed by the arrangement of several amino acids. Crucially, these pockets often serve as active or binding sites where other molecules, such as potential drug candidates, can interact with the protein to modulate its function (Stevenson et al., 2023).

To study these molecular interactions and understand protein structure in detail, researchers often rely on structural data. The Protein Data Bank (PDB) (Berman et al., 2000) is a globally recognized repository providing high-resolution 3D models of experimentally determined protein structures, including protein-ligand complexes. It enables the identification and analysis of binding pockets, interaction patterns, and conformational changes, all of which are critical for advancing drug discovery and protein function analysis. Complementing experimental efforts, AlphaFold (Jumper et al., 2021) provides a computational method that can regularly predict protein structures with atomic accuracy. This is achieved through a novel machine learning approach, incorporating physical and biological knowledge about protein structure and leveraging multi-sequence alignments within its deep learning algorithm. This AI-powered tool has significantly expanded the accessible structural landscape, offering valuable insights into proteins for which experimental structures are not yet available, thereby accelerating research into their function and potential drug targets.

Beyond the static structural information found in the PDB, comprehensive sequence and functional data are equally essential for understanding protein behavior. UniProt ("UniProt: the Universal protein knowledgebase in 2025" 2025) fufills the role serving as a comprehensive, high-quality, and freely accessible central database of protein sequences and functional annotations. It integrates data from numerous sources, including large-scale genomic sequencing projects and scientific literature, offering detailed information on protein function, cellular localization, post-translational modifications, and evolutionary relationships. UniProt is an indispensable resource for researchers seeking to understand the vast diversity of proteins and

their roles in biological systems.

The sheer number of known proteins, cataloged in different databases, necessitates systematic organization. To manage these, proteins are organized into protein families, which are groups of proteins sharing a common evolutionary origin, reflected in their similar amino acid sequences, structures, or functions. This classification helps researchers predict protein function, understand evolutionary relationships, and identify conserved features like motifs or domains critical for function. By grouping proteins, insights from one family member can be applied to others, streamlining research and drug design efforts.

2.1.1 Protein Representation in Deep Learning

For DL algorithms to effectively process and learn from proteins, these complex biological macromolecules must be transformed into a numerical format, or representation. Traditional methods often rely on features derived from sequence properties such as amino acid composition or physicochemical properties (Wu et al., 2022), or from structural information, including contact maps or secondary structure elements. While these representations have proven useful in various contexts, they may not always capture the complete biological complexity and evolutionary information intrinsic to protein sequences.

More recently, Protein Language Models (PLMs) have emerged as a significant advancement in protein representations. These DL models, often based on transformer architectures (similar to those widely used in natural language processing), are pre-trained on vast datasets of raw protein sequences, often comprising billions of sequences from resources like UniProt. During this pre-training phase, PLMs learn intricate patterns, relationships, and evolutionary constraints directly from the sequences. This process allows them to develop a sophisticated internal model of protein 'grammar', by predicting masked amino acids or learning contextual dependencies. Consequently, they effectively internalize principles of protein structure, function, and evolution. The output for a given protein sequence is a set of dense numerical vectors, known as protein embeddings, which encapsulate rich semantic and functional information. These embeddings are highly informative, capable of reflecting evolutionary relationships, functional annotations, and even implicitly, structural characteristics.

Among the most prominent and widely adopted PLMs are models like ProtTrans (Elnaggar et al., 2021), ProtGPT2 (Ferruz, Schmidt, and Höcker, 2022), and ESM-2 (Evolutionary Scale Modeling, version 2) (Lin et al., 2022). It is a large-scale transformer model trained on an extensive dataset of over 250 million protein sequences. A key atrribute of ESM2 is its capacity to generate high-quality protein embeddings that capture subtle evolutionary and structural signals. These embeddings have demonstrated effectiveness in diverse applications, including predicting protein structure, identifying functional sites, classifying protein families, even guiding de novo protein design, protein-protein interactions and protein-drug interactions.

While PLMs excel at capturing information from sequences, some approaches have integrated structural information to further enrich protein representations. These methods aim to combine the strengths of sequence-based language models with the explicit spatial relationships found in protein structures. For instance, models like SaProt (Su et al., 2023) incorporate structure-aware vocabularies using the Foldseek

(Van Kempen et al., 2024) representation, effectively creating sequence and structural tokens that allow the model to learn from both modalities simultaneously. This hybrid approach enables a more comprehensive understanding of protein characteristics, moving beyond purely sequence-derived features (Su et al., 2023).

Beyond representing the entire protein, accurately representing specific functional regions, such as protein binding pockets, is increasingly crucial, especially for drug discovery. An example of such a model is PickPocket (Tarasi, Malo, and Molina, 2025). PickPocket's architecture strategically combines ESM-2 with GearNet (Zhang et al., 2022). ESM-2 provides the sequence-derived evolutionary embeddings, which then serve as node features for GearNet. GearNet, in turn, models the protein structure as a graph, capturing spatial relationships between residues. By integrating the outputs of these two models, PickPocket generates a comprehensive representation that considers both the sequence context and the structural information of the protein. These combined embeddings are subsequently used by a two-layer Multi-Layer Perceptron (MLP) classifier to predict per-residue binding probabilities, thereby identifying potential binding pockets.

2.2 Small Molecules

At their core, molecules are stable arrangements of two or more atoms held together by chemical bonds. They represent the smallest identifiable unit of a substance that still retains its chemical properties. Within this broad category, small molecules are a specific class characterized by their relatively low molecular weight. Many pharmaceutical drugs, for example, are small molecules because their size allows them to easily interact with biological targets within cells to exert a therapeutic effect. Small molecules often act as ligands, interacting with target proteins, nucleic acids, or other biomolecules to modulate their activity. (Li and Kang, 2020).

In the field of drug discovery, a crucial challenge lies not only in identifying compounds with desired biological activity but also in ensuring that these compounds possess physicochemical properties conducive to oral bioavailability. Many promising drug candidates fail in clinical development due to poor absorption, distribution, metabolism, and excretion (ADME) properties. To address this, Christopher Lipinski and colleagues formulated the "Rule of Five" in 1997 (Lipinski et al., 1997, based on observations from a large dataset of orally active drugs. This rule provides a set of guidelines for predicting the oral absorption and permeability of a compound, helping to filter out molecules that are unlikely to be successful drug candidates early in the discovery process.

For a small molecule to be a successful drug, beyond its ability to bind to a target, it must possess suitable ADMET properties. ADMET is an acronym representing key pharmacokinetics and safety considerations: 1) Absorption, how the drug enters the bloodstream; 2) Distribution, how it spreads through the body 3) Metabolism, how the body chemically transforms it 4) Excretion, how it is eliminated and 5) Toxicity, its potential for harmful effects. Optimizing these properties is crucial in drug discovery, as poor ADMET characteristics are a primary reason drug candidates fail, even if they show string activity in early experiments. Therefore, understanding a drug's ADMET profile early in development is vital for designing safe and effective therapies (Yi et al., 2024).

2.2.1 Small Molecules Representation in Deep Learning

For DL algorithms to process molecules effectively, these chemical structures must be translated into a numerical or textual format. SMILES (Simplified Molecular Input Line Entry System) (Weininger, 1988) strings are a linear textual representation of molecular structures, designed to encode atoms and bonds as sequences of characters. To make these representations suitable for DL architectures, SMILES sequences must first be converted into a numerical format. This is typically achieved through one-hot encoding, where each character in the SMILES alphabet is represented as a binary vector indicating the presence (1) or absence (0) of that character at a given position.

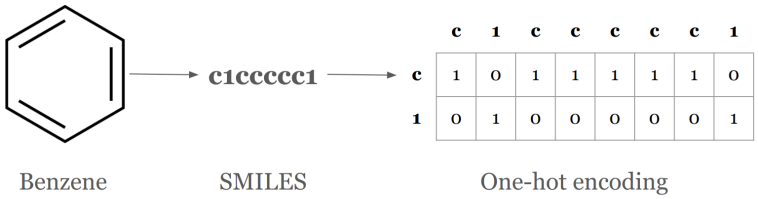


FIGURE 2.1: The SMILES representation and one-hot encoding for benzene. For purposes of illustration, only the characters present in benzene are shown in the one-hot encoding. In practice there is a column for each character in the SMILES alphabet.

Formally, a one-hot vector is a binary vector of length equal to the size of the SMILES alphabet. For example, given an alphabet size of $N = 34$, the character "N" (assigned ID 0) is encoded as $[1, 0, 0, \ldots, 0]$, while "-" (assigned ID 1) becomes $[0, 1, 0, \ldots, 0]$. Each SMILES string of length L is thus represented as a matrix of shape (L, N), where each row is the one-hot vector of a character at a specific position.

This discrete representation enables models such as RNNs and Transformers to process molecular data analogously to natural language. These models can learn to predict molecular properties like solubility, toxicity, or protein-binding affinity(David et al., 2020; Mswahili and Jeong, 2024). Furthermore, generative variants of these architectures are capable of producing novel SMILES strings that correspond to chemically valid molecules, offering a powerful tool for molecular discovery and design. The learned embeddings from such models capture underlying chemical relationships, allowing for structural clustering and similarity-based compound retrieval.

2.2.2 Molecule Generation Models

As mentioned in Introduction (1), there are lots of AI algorithms to generate molecules. While earlier methods utilized architectures such as RNNs for sequential SMILES string generation, VAEs, GANs for latent space manipulation, and Graph Neural Networks GNNs for direct graph generation, more recent advancements have introduced increasingly sophisticated paradigms, particularly Transformer models and Diffusion models (Mouchlis et al., 2021).

Transformer models have proven highly effective for molecule generation, especially when molecules are represented as sequential SMILES strings. Their core innovation lies in the self-attention mechanism, which allows the model to weigh

the importance of different parts of the input sequence when processing each element. This capability is particularly advantageous for capturing long-range dependencies within molecular structures, a limitation often encountered by RNNs. By learning complex contextual relationships in the 'molecular language' of SMILES, Transformer models can generate chemically valid and diverse compounds, often with improved control over specific properties (Luong and Singh, 2024).

More recently, Diffusion models (Alakhdar, Poczos, and Washburn, 2024) have emerged as a powerful generative paradigm, demonstrating remarkable success in various domains, including image synthesis, and are now being effectively applied to molecular generation. The fundamental idea behind diffusion models involves a two-step process: a 'forward diffusion' process progressively adds noise to data until it becomes pure noise, while a 'reverse diffusion' process learns to gradually denoise the data, reconstructing original samples from random input. For molecule generation, this translates to learning to reverse the corruption of molecular structures (e.g., noisy molecular graphs or latent representations) to synthesize novel, valid, and diverse molecules.

A significant advantage of both Transformer and Diffusion models is their inherent capacity to readily incorporate additional sources of information. This allows the generative process to be guided by external context, moving beyond mere de novo generation to conditional generation. Specifically, these models can effectively leverage protein representations, such as protein embeddings derived from PLMs. By feeding these protein embeddings alongside the molecular data, the models learn to generate molecules that are not just chemically plausible but are also tailored to the specific characteristics of a target protein. This conditioning mechanism allows the model to capture the nuanced relationship between a protein's features and the chemical properties required for effective binding.

2.3 Small Molecule - Protein Interactions

The ability of small molecules to modulate biological systems relies on their specific interactions with protein targets. These protein-drug interactions are governed by a combination of non-covalent forces and precise molecular recognition [cite]. Small molecules can bind selectively to specific pockets of a protein through non-covalent interactions like hydrogen bonds, van der Waals forces, and hydrophobic effects. This binding event, known as a protein-ligand interaction, can subsequently modulate the protein's function.

Usually, small molecule ligands bind to a specific pocket located within the protein three-dimensional structure. The strength of these interactions is quantitatively assessed by their binding affinity, commonly expressed as a dissociation constant (K_D or K_i). This constant reflects the equilibrium between the bound and unbound states of the drug-protein complex. A lower K_D or K_i value indicates higher affinity, generally signifying strong and specific binding of the ligand to its target.

Understanding these intricate molecular interactions is crucial for rational drug design. To support this, comprehensive databases like ChEMBL (Zdrazil et al., 2024) play a vital role. ChEMBL is a freely accessible, manually curated database that focuses on bioactive molecules with drug-like properties. It integrates chemical structures, quantitative bioactivity data against various protein targets, cellular and organism-level effects, as well as ADMET properties. Since its launch in 2009,

ChEMBL has grown significantly in size and scope, now containing over 2.2 million compounds and more than 18 million bioactivity records.

2.3.1 Deep Learning for Drug-Target Interaction (DTI) prediction

Tensor-DTI (Gil-Sorribes, Ciudad Serrano, and Molina, 2025) is a contrastive learning framework for drug–target interaction prediction that embeds proteins, binding pockets, and small molecules into a shared latent space using a dual-encoder architecture. It supports two main configurations: one using only whole-protein representations, and another incorporating explicit binding pocket embeddings derived from structural data, forming a protein and pocket–drug interaction setup. The pocket-based variant is trained on the PLINDER dataset, which provides high-quality residue-level binding site annotations and is specifically curated to minimize data leakage. The original Tensor-DTI framework includes multiple datasets covering diverse interaction scenarios; among them, two are especially relevant to this work: SMPBind-I, a large-scale dataset of experimentally validated protein–ligand interactions used for training protein-only models, and BindingDB (Gilson et al., 2016, which served for additional protein-level evaluations. In both configurations, Tensor-DTI is trained using a contrastive loss. This objective brings interacting pairs, whether protein–drug or protein + pocket–drug, closer in latent space, while separating non-interacting ones.

2.3.2 Deep Learning for Molecular Generation Conditioned on Protein Targets

DTI prediciton tools can serve to predict already existing molecules with possible targets or *de novo* generated ligands, as they can check if the generated molecule will interact with a chosen protein target. To bridge this gap and generate truly functional molecules, it is essential to integrate information about the protein target directly into the molecule generation process. This is where PLMs offer a transformative advantage.

By leveraging these embeddings, generative DL models can be conditioned to produce molecules tailored to a specific protein target. This approach can be conceptualized as framing drug design as a machine translation problem between two distinct 'languages': the amino acid language of proteins (represented by PLM embeddings) and the SMILES language of small molecules. In this paradigm, the model learns to "translate" the characteristics of a target protein into the chemical structure of a potential binder, thereby directly generating molecules that are predicted to bind to that protein. This enables the *de novo* creation of binders directly from the information encoded within pre-trained protein language models (Creanza et al., 2025).

Chapter 3

Experimental Setup

3.1 Data

3.1.1 Plinder

Our work leverages PLINDER (Protein Ligand INteractions Dataset and Evaluation Resource) (Durairaj et al., 2024) database critical for training and evaluating computational methods in protein-ligand interaction prediction with pocket information.

PLINDER aggregates data from key sources including PDB for experimentally determined complexes, AlphaFold2 for predicted structures where experimental apo forms are lacking, and SCOP for domain-level annotations. Foldseek and MMseqs2 are used to process and score structural similarities It computes over 20 billion similarity scores using 14 distinct metrics such as interaction fingerprints, binding pocket similarity, ligand similarity, and various sequence and structural comparisons (e.g., RMSD, Tanimoto scores).

The database offers predefined train/validation/test splits to ensure minimal data leakage and promote generalization to unseen proteins, ligands, and interactions. These splits are tailored for machine learning applications and are fully customizable depending on the task. Additionally, PLINDER includes a standardized evaluation framework that supports CASP-CAPRI-compatible metrics like DockQ and RMSD. It allows for performance benchmarking using both experimental (holo and apo) and AlphaFold2-predicted structures, facilitating fair comparisons of state-of-the-art computational models.

3.1.2 Data Preprocessing and Filtering

We implemented a filter for ligand size, retaining only those that did not exceed a length of 100 characters, representing the number of atoms. The filtered entries from PLINDER's original 'train' and 'val' splits were then combined to form what we refer to as our 'train split' for subsequent analyses.

In this types of works important to analyze the distribution of protein types within the dataset, as biases can significantly impact model performance. We plotted the distribution of protein families present in our filtered PLINDER dataset. Across the entire filtered dataset, we identified 2121 unique protein families. The train split alone contained 1685 of these, while the test split had 733.

Upon closer inspection, we realized a significant imbalance within our filtered PLINDER train split (Fig. 3.1), approximately 10% of the proteins corresponded

to the Betacoronavirus family, and in fact, the top five most frequent families were all viral. This highly disproportional representation raised concerns about potential bias in molecule generation. To address this, our first experiment involved analyzing whether training with these highly frequent families had an impact on the generated molecules. Consequently, we created two distinct training sets:

- **"With Virus" Training Set:** This set includes all filtered interactions, comprising 124,104 protein-pocket-drug interactions, thereby retaining the observed viral family overrepresentation.
- **"Without Top 5 Virus" Training Set:** To mitigate the identified possible bias, this set excludes interactions corresponding to the top five most frequent viral families, resulting in 85,164 protein-pocket-drug interactions. This allows us to directly assess the impact of these prevalent families. See distribution in Fig. A.1.

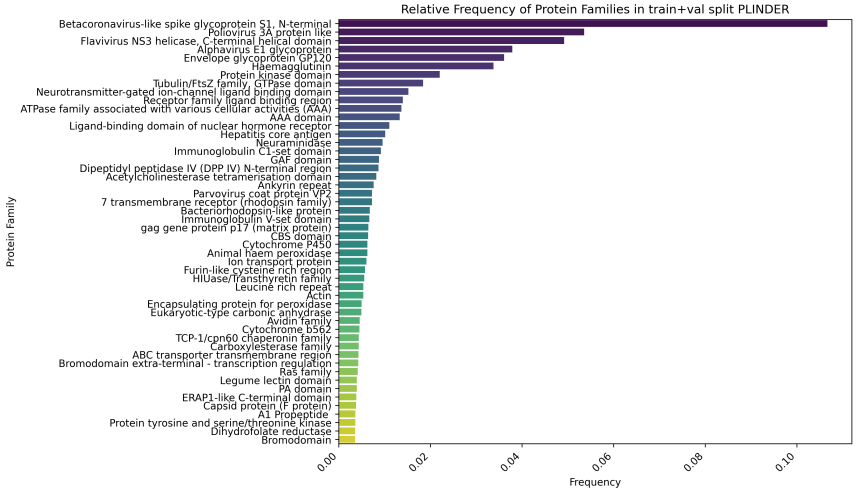


FIGURE 3.1: Proportion of top 50 most frequent families in train split with top 5 viral families.

For the test split (Fig. 3.2), we did not observe a similar disproportional distribution of viral families. Thus, we decided to keep it as is. Since these interactions correspond to proteins and ligands the model had not encountered during training, this test set served as our out-of-domain test set, for evaluating true generalization.

We also analyzed the types of molecules present in the dataset, more specifically, We analyzed the presence of PAINS (Pan-Assay Interference Compounds) across protein families in the PLINDER dataset. These molecules can cause false positives and compromise assay reliability. For each family, we counted the number of ligands flagged as PAINS. While most families had few or none, others, such as bromodomains and kinase domains, showed a disproportionately high number of flagged compounds (Fig. 3.3). This underlines the importance of accounting for PAINS distribution, as it may bias generative models toward producing unsuitable molecules.

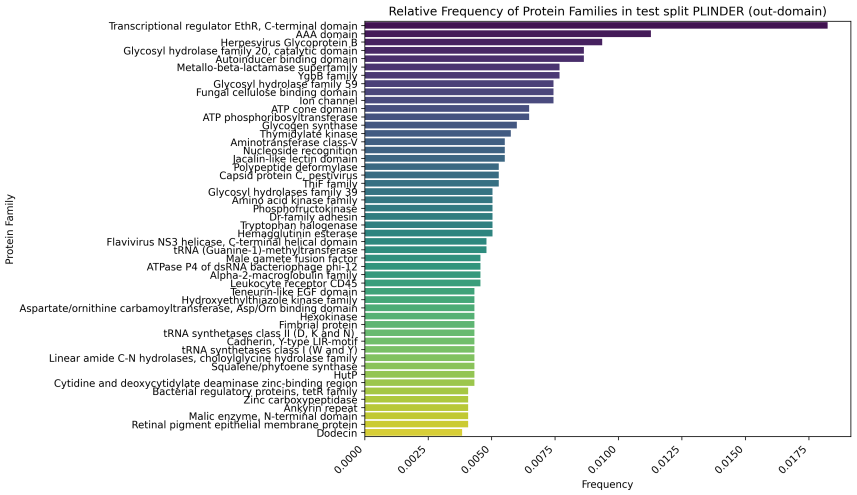


FIGURE 3.2: Proportion of top 50 most frequent families in test split (out-domain)

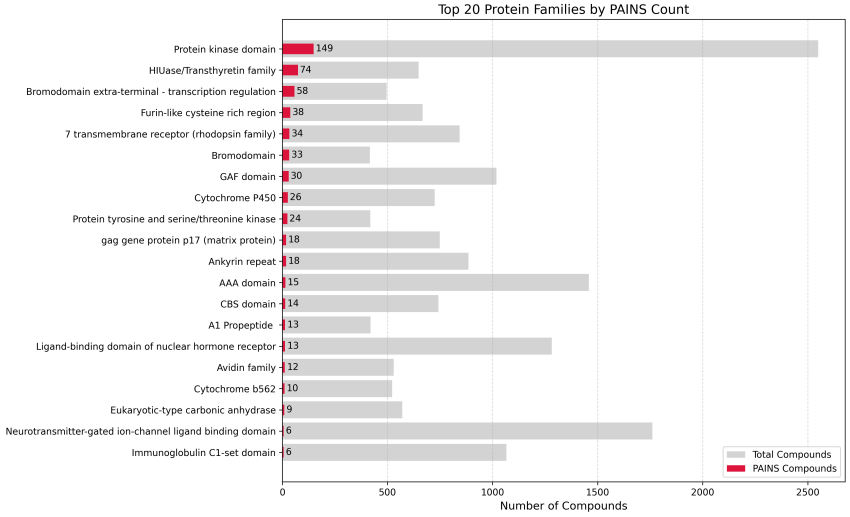


FIGURE 3.3: Top 20 protein families with the most PAINS compounds.

3.2 Input Representations

As previously introduced, effectively conditioning molecule generation on target protein and pocket information is crucial for rational drug design. This thesis benchmarks various pre-trained input representations for both proteins and pockets to investigate which provides the most informative conditioning signal for our generative model.

We explored three main categories of embeddings: 1) protein-only representations, 2) pocket-specific representations, and 3) integrated protein-pocket representations from a drug-target interaction (DTI) framework. For all mentioned algorithms, representations were extracted following their respective GitHub instructions.

It is important to mention a practical consideration because we explored different experiments with various types of input representations, each specific model ultimately ended up being trained with a slightly different final dataset. This can occur due to factors like minor data leakage during representation creation or limitations where some models are unable to create embeddings for all provided proteins or pockets due to technical reasons. Any such model-specific dataset variations will be detailed in their respective experimental sections.

3.2.1 Protein Embeddings

For protein representations, we used embeddings from two distinct protein language models: ESM2 and SaProt. ESM2 embeddings, extracted from the pre-trained model `esm2_t36_650M_UR50D()`, provide a comprehensive, global understanding of the target protein's identity and properties, derived from billions of protein sequences. In contrast, SaProt embeddings, obtained from the pre-trained `SaProt_650M_AF2` model, are designed with explicit structural awareness, allowing us to investigate whether providing a generative model with inherently more structure-aware representations improves molecule generation.

3.2.2 Pocket Embeddings

For pocket representations, we utilized PickPocket embeddings. These embeddings, derived from an in-house model (Tarasi, Malo, and Molina, 2025), offer a localized and detailed view of the target site. These embeddings provide the generative model with crucial information about the immediate environment where a ligand is expected to bind. This allows the model to condition molecule generation on precise pocket characteristics.

3.2.3 Tensor-DTI Embeddings

We also used Tensor-DTI, pretrained on the PLINDER dataset, to extract embeddings of proteins and pockets after contrastive training. These embeddings, learned through a contrastive objective, are not only grounded in the original representations from pretrained protein language models but also enriched with knowledge derived from drug–target interaction patterns. By training to distinguish binding from non-binding pairs, Tensor-DTI encourages the encoder to learn representations that emphasize features critical for molecular recognition and interaction. Conditioning the generative model on these embeddings allows it to benefit from this biologically informed signal, improving its ability to generate molecules that are not only structurally valid but also more likely to bind the intended target.

To obtain these embeddings, we leveraged internal checkpoints from the Tensor-DTI model trained on PLINDER. Specifically, the protein and pocket representations were extracted by forwarding the respective PDB inputs through the encoder modules of the trained model.

3.2.4 Experimental Setups for Input Combinations

To assess the impact of these various input representations, we investigated four distinct experimental setups for training our generative model:

1. **ESM2 Protein Only**: This setup evaluates the generative model's performance when conditioned solely on the global protein context provided by ESM2 embeddings.
2. **ESM2 Protein + PickPocket Pocket**: This setup adds local information of the binding site, enabling the generation of molecules to be more pocket specific.
3. **SaProt Protein + PickPocket Pocket**: This configuration investigates the synergy between structure-aware protein embeddings (SaProt) and explicit pocket representations (PickPocket), aiming to provide both global structural context and local binding site details.
4. **Tensor-DTI Protein Embeddings + Tensor-DTI Pocket Embeddings**: This setup leverages the integrated, interaction-aware embeddings derived from the Tensor-DTI framework for both protein and pocket inputs.

These experiments allow for a systematic comparison of how different levels and types of pre-trained biological information influence the efficacy of the conditional molecule generation process.

3.2.5 Creating SMILES Numerical Representations

The other side of the interaction corresponds to the ligand. To enable the models to process molecular structures, SMILES strings were transformed into a numerical, fixed-length representation.

Initially, raw SMILES strings were extracted from the dataset described previously. A custom tokenization procedure was applied to accommodate specific chemical elements: "Cl" (Chlorine) was replaced with "D", and "Br" (Bromine) was replaced with "E". This approach ensures that these two-character elements are treated as single tokens within the defined alphabet. Additionally, each preprocessed SMILES string was enclosed with special 'begining-of-sequence' (BOS) and 'end-of-sequence' (EOS) tokens ("è" and "§" respectively) to explicitly mark sequence boundaries.

The length of the SMILES strings in the dataset ranged from 25 to 100 characters. To achieve a uniform input length suitable for the model, all SMILES strings were padded to match the maximum sequence length observed ($\approx$ 100 characters). Padding was accomplished by appending a special character ("£") to the end of shorter SMILES strings until they reached the required length. This step is essential for enabling batch processing in neural networks, which typically require inputs of consistent dimensions.

The core of the numerical representation is a one-hot encoding scheme. A predefined alphabet of unique SMILES characters was established. This alphabet, containing 34 unique characters, includes standard SMILES characters, the custom substitutions (D, E), and the special tokens ("è", "§" and "£"). Each character in the alphabet was assigned a unique integer ID, and then a corresponding one-hot vector.

Finally, each SMILES string was converted into a three-dimensional numerical array. For a batch containing M SMILES strings, each of length L (where $L \approx 100$), and an alphabet size of N, the resulting representation is a NumPy array of shape (M, L, N). Each character in a SMILES string is replaced by its corresponding N-dimensional one-hot encoded vector, producing a dense numerical representation suitable for input into the deep learning model.

3.3 Model Architecture

The core of our approach is a custom deep learning model, built upon the Transformer architecture principles, specifically adapted for molecule generation conditioned on protein and pocket information, the overview architecure can be seen in Fig. 3.3. The model is implemented using the PyTorch framework.

3.3.1 Input Representation

The model recieved three primary inputs: protein features, pocket features, and the target SMILES sequence. When input embeddings were not already 256-dimensional, they were projected to this size using a forzen linear layer to ensure consistency accross modalities.

- **Protein Features**: These were 256-dimensional embeddings representing the entire protein sequence. These embeddings were extracted from pre-trained PLMs, either 1) ESM-2, 2) SaProt both originally 1280-dimensional reduced to 256 or 3) the protein passed through TensorDTI encoder which already produced 256-dimensional embeddings.
- **Pocket Features**: These represent the binding pocket of the protein. Two sources where used 1) PickPocket or 2) the pocket embeeding passed thorugh TensorDTI, providing localized structural and chemical information about the binding site. PickPocket og dimensions are 4352 and is passed thourgh a frozen linear layer to reduce it to 256 before starting training.
- **SMILES Target Sequence**: This is the SMILES string of the ligand to be generated, represented as a sequence of one-hot encoded characters, as described in Section 3.2.5. During training, the model receives the SMILES up to the current character and attempts to predict the next character.

All input features are batched, with dimensions $[B, \text{sequence\_length}, \text{feat\_dimension}]$ or $[B, \text{feat\_dimension}]$ depending on the input type, where B is the batch size. For protein and pocket features, if they are single vectors per example, they are unsqueezed to $[B, 1, \text{feat\_dimension}]$ for consistent processing.

3.3.2 Memory Fusion: Combining Protein and Pocket Information

The protein and pocket features are combined to form a contextual "memory" that guides the SMILES generation process. This fusion occurs within the decoder-only model and was done using concatenation procedure. The protein and pocket feature vectors were concatenated along their feature dimension. If protein features are

$[B, 1, \text{input\_size\_prot}]$ and pocket features are $[B, 1, \text{input\_size\_pocket}]$, the resulting tensor will have a shape of $[B, 1, \text{input\_size\_prot} + \text{input\_size\_pocket}]$. With both reduced to 256 dimensions, this gave a combined vector of 512 dimensions. This method allows the model to learn combined representations by placing the information side-by-side.

The resulting tensor was normalized using LayerNorm before being used as context in the Transformer decoder.

3.3.3 Transformer Decoder Structure

The model followed a Transformer Decoder architecture, well-suited for autoregressive sequence generation tasks.

- **Target Embedding**: One-hot encoded SMILES inputs were projected to a shared embedding space of size `d_model` using a linear layer. Positional encoding were added to preserve sequence order, and dropout was applied for regularization.
- **Decoder Layers**: The decoder consisted of four stacked layers, each containing masked multi-head self-attention (to prevent access to future tokens), multi-head cross-attention over the trg tensor and a feed-forward network for non-linear transformation.

3.3.4 Output Layer

The decoder output, shaped $[B, \text{trg\_len}, \text{d\_model}]$, was passed through a final layer prokecting to the size of the smiles vocabulary. The resulting logits $[B, \text{trg\_len}, \text{vocab\_size}]$ represented unnormalized probabilities over possible characters. During training, these were passed to a Cross-Entropy Loss function, which internally applies softmax.

3.3.5 Training Objective and Optimization

The model was trained to predict the next character in the SMILES sequence given the previous characters and the protein-pocket context.

- **Loss Function**: Cross-Entropy Loss (`nn.CrossEntropyLoss`) was used, which is standard for multi-class classification tasks like character prediction. It compares the model's predicted logits with the true next character (represented as integer indices).
- **Optimizer**: The AdamW optimizer (`optim.AdamW`) was employed for weight updates.
- **Learning Rate Scheduler**: A custom learning rate scheduler was implemented. This scheduler uses a warmup phase for a specified number of epochs (5), during which the learning rate linearly increases, followed by a cosine annealing schedule where the learning rate gradually decreases. This strategy helps in stabilizing training and achieving better convergence.

The training loop iterates through batches of protein-pocket-SMILES data. For each batch, the model performed a forward pass, the loss was computed, gradients were backpropagated, and the optimizer updated the model's weights. Training was distributed across multiple GPUs using PyTorch's Distributed Data Parallel (DDP) to accelerate the process.

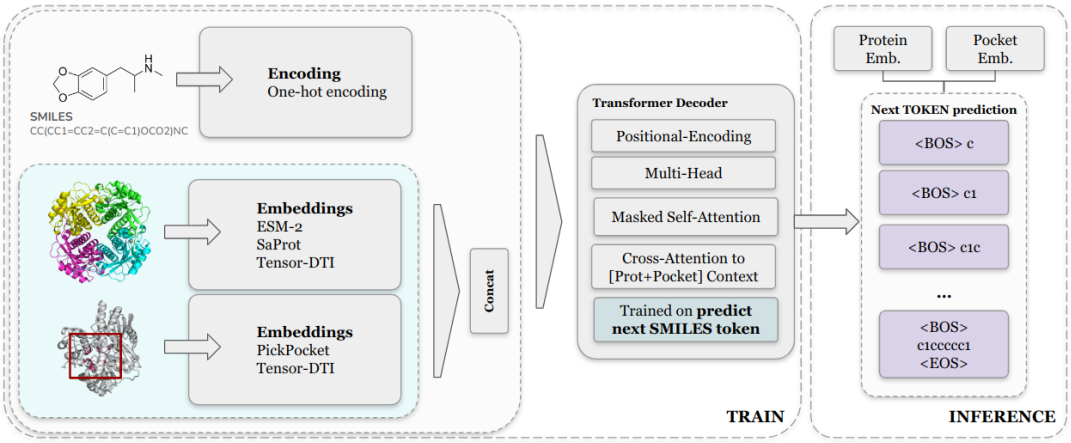


FIGURE 3.4: Simplified Diagram of the Model Architecture and Inference. The protein shown is PDBID: 5ISX.

3.4 Molecule Generation Process

Once the model was trained, it was used to generate novel SMILES strings conditioned on specific protein and pocket inputs. The generation process employed an autoregressive, character-by-character sampling approach, in Fig. 3.3 a simplified inference process can be seen.

3.4.1 Autoregressive Sampling

The model generated SMILES strings sequentially, one character at a time, until an end-of-sequence token was predicted or a maximum length was reached. This process was autoregressive, meaning each subsequent character prediction was conditioned on the characters generated thus far, in addition to the initial protein and pocket context.

3.4.2 Initialization and Iterative Generation

For each protein-pocket pair for which a molecule was to be generated, the process began by initializing the target input sequence with the special BOS token ("è"). The protein and pocket embeddings for the target pair were fed into the model as conditioning information.

In each generation step (t):

1. The current partial SMILES sequence was processed by the model along with its positional encodings. A causal mask was applied to ensure that the model only attended to previously generated characters.

2. The conditioned Transformer Decoder predicted a distribution of probabilities (logits) for the next possible character.
3. Instead of simply taking the most probable character (greedy decoding), Top-K sampling was applied. This method considered only the k most probable next characters and sampled one from this reduced set, introducing a controlled degree of randomness and diversity into the generated molecules. In our setup, k was set to 5.
4. The sampled character was appended to the current partial SMILES sequence.
5. The process continued iteratively until the model predicted a special EOS token ("S") or the generated SMILES reached a predefined maximum length, by default 120 characters.

3.4.3 Post-Generation Processing and Validation

After a SMILES string was generated, a post-processing step converted the custom tokens back to standard chemical notation: "D" was replaced with "Cl" (Chlorine) and "E" was replaced with "Br" (Bromine). Each generated SMILES string was then validated using the RDKit library to check for chemical validity. Only valid SMILES strings were typically considered for subsequent analysis.

The entire generation process was repeated 1000 times for each protein-pocket pair to produce multiple candidate molecules. Results, including the generated SMILES and their validity status, were saved to a TSV file. Distributed Data Parallel (DDP) was used during generation to parallelize the sampling across multiple GPUs, accelerating the process.

3.5 Model Inference

To assess our model's generative capabilities, particularly its ability to create new molecules for both familiar and entirely novel proteins, we designed three distinct types of test sets for inference:

- **In-Domain Test Set (with virus):** This set reflects the distribution of proteins found in our "with virus" training regimen.
- **In-Domain Test Set (without top 5 virus):** This set reflects the distribution of proteins from our "without top 5 virus" training regimen.
- **Out-of-Domain Test Set:** This set utilizes the unaltered PLINDER test split, providing a challenge with proteins unseen during any training.

For each of these test sets, we carefully selected a single protein for each of the top 50 most frequent families within that specific test set's distribution. This yielded test sets consisting of 50 proteins each. During inference, our strategy was to generate 1000 molecules for each protein-pocket target, meaning we generated a total of 50,000 molecules for each test set (50 proteins * 1000 molecules/protein).

3.6 Evaluation

3.6.1 Drug Similarity Metrics

Molecular similarity metrics provide quantitative measures of the resemblance between two molecules based on their structural, physicochemical, or biological properties. These metrics are crucial for several reasons: they enable the identification of known compounds similar to a generated candidate, facilitate the exploration of chemical space around promising hits, and help to ensure that generated molecules are diverse yet retain desirable drug-like characteristics or scaffold.

Our analysis relies on Tanimoto similarity to quantify the relationships between molecules, for more theoretical background about this metric refer to Appendix C.1. When assessing novelty against the training set, we categorize generated molecules based on their highest Tanimoto similarity score to any molecule in the training data:

- **Tanimoto $<$ 0.5**: These molecules are considered unique and structurally distinct from the training set, representing genuinely novel candidates.
- **0.5 $\leq$ Tanimoto $<$ 1.0**: Molecules in this range are considered to have moderate similarity to the training set. While not identical, they share significant structural features with known compounds.
- **Tanimoto = 1.0**: These are exact matches to molecules present in the training set.

For internal diversity and cross-dataset diversity, Tanimoto similarity is calculated for all-to-all pairs within or between generated sets. The same Tanimoto thresholds (0.5 and 1.0) are applied to categorize the similarity of these pairs.

3.6.2 Interaction Evaluation of Generated Molecules

Following SMILES generation, we assessed the likelihood of interaction between each molecule and its target using pretrained Tensor-DTI classifiers. As pocket information was available for all targets in our setup, the primary evaluation was performed using the PLINDER-based Tensor-DTI model, which explicitly models protein + pocket–drug interactions. In addition, we evaluated the same generated molecules using Tensor-DTI models trained on SMPBind-I and BindingDB, which operate solely at the whole-protein level and do not incorporate pocket information. By comparing pocket-informed and protein-only models, we were able to assess how much additional value the pocket context provides during molecule evaluation.

3.7 Physicochemical Properties

For generated molecules to be considered "drug-like" and increase their likelihood of progressing through the drug discovery pipeline, they must possess favorable physicochemical and structural properties that influence pharmacokinetics (ADMET). To rapidly evaluate these characteristics in-silico, we utilized cheminformatics toolkits such as RDKit Bento et al., 2020, which provides robust functionalities for computing a broad array of molecular properties from SMILES strings. Our analysis

of the generated compounds included key properties such as Molecular Weight (MolWt), Topological Polar Surface Area (TPSA), Hydrogen Bond Donors and Acceptors (NumHDonors, NumHAcceptors), Number of Rotatable Bonds (NumRotatableBonds), Number of Aromatic Rings (NumAromaticRings), Heavy Atom Count (HeavyAtomCount), and Synthetic Accessibility Score (SAScore). Additionally, we assessed the Murcko Scaffold to gain insight into the fundamental core structures, evaluating the diversity and novelty of the generated molecular frameworks. For further theoretical information about these metrics refer to C.2.

3.8 Pipeline

Our pipeline involved three main stages: model training, molecule generation, and analysis of the generated compounds. This procedure was applied consistently across all experimental setups, including in-domain, out-domain, and specific case studies.

3.8.1 Model Training and Checkpoint Selection

All models were trained for 30 epochs, following to the architecture and training methodology detailed in Section 3.3. Following training, a specific checkpoint was selected for each model to be used for molecule generation. This selection criterion focused on identifying checkpoints where the training loss (Fig. 3.5) was below 0.2 but before it had flattened, typically observed between epoch 6 and 10, with a target loss value around 0.17. This approach was inspired by similar practices in literature, such as Creanza et al., 2025, which reported optimal performance at a training loss of 0.16. These selected checkpoints contained the optimized learned weights necessary for the molecular generation algorithm, as described in Section 3.4.

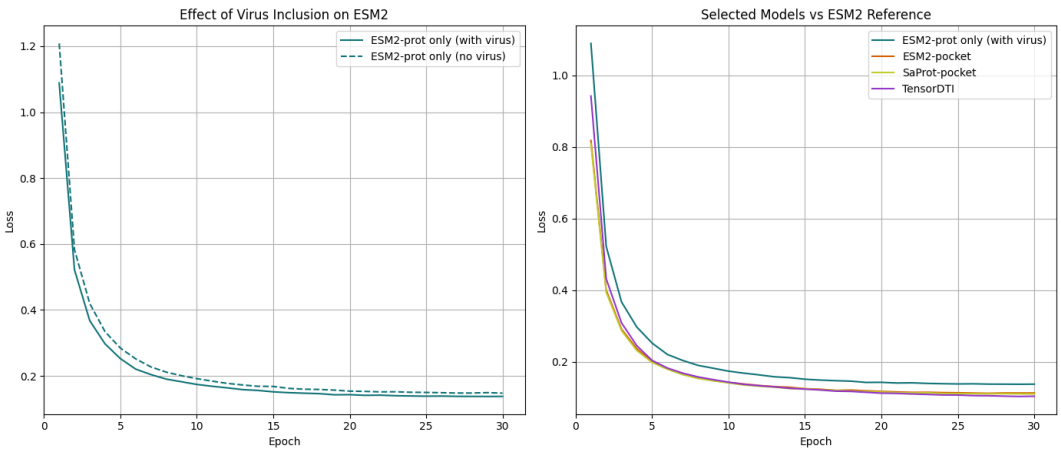


FIGURE 3.5: Loss curves for the selected models we are analyzing

3.8.2 Molecule Generation

Leveraging the trained models, 1000 molecules were generated for each given protein and pocket embedding. This process resulted in a diverse set of candidate compounds for each target.

3.8.3 Analysis of Generated Molecules

The generated molecules underwent a multi-step analysis to assess their quality and relevance. This process included:

- **Uniqueness Assessment:** To evaluate the novelty of the generated molecules, a nearest-neighbors Tanimoto similarity comparison was performed against the training set. Molecules with a Tanimoto similarity of less than 0.5 to any molecule in the training set were considered unique and selected for further analysis.
- **Internal Diversity:** A Tanimoto similarity "all-against-all" analysis was then conducted among the unique molecules. This ensured that the generated compounds were structurally diverse from one another, preventing redundancy within the set of promising candidates.
- **Physicochemical Property Assessment:** Finally, the physicochemical properties (e.g., molecular weight, Synthetic Accessibility score, etc.) of the selected molecules were assessed to ensure they adhered to desired drug-like criteria.
- **Predicted Interaction with Target:** For the unique and valid molecules, TensorDTI inference was performed. This step identified which of the generated compounds were predicted to interact with the target protein, indicating their potential as functional binders.

Chapter 4

Results

Our research involved training generative models and generating molecules across a wide array of input combinations and for both 'with virus' and 'without top 5 virus' training datasets. To present a focused and interpretable analysis, and given that similar conclusions emerged from the broader set of experiments, this section will delve into a selected subset of these models. This strategic selection allows us to directly address our primary research questions concerning the influence of dataset composition and input representation on model performance. The comprehensive details of all trained model combinations are provided in Appendix .

Specifically, this analysis aims to benchmark two critical aspects:

1. The impact of dataset composition, particularly the proportion of viral protein families, on the model's performance in molecule generation.
2. The efficacy of different input representation types for guiding the molecule generation process.

By focusing on these key comparisons, we seek to provide clear insights into the factors that most significantly contribute to the successful generation of novel molecular candidates.

Finally, beyond these benchmarking analyses, we also investigated the model's capacity to generate unique SMILES strings specifically conditioned on the characteristics of individual binding pockets. To this end, we selected two retrospective case studies involving proteins known to possess two different binding pockets. The primary objectives of this analysis were twofold: first, to determine if the model could produce structurally distinct molecules when provided with different pocket contexts, and second, to evaluate whether these generated candidates demonstrated a high predicted propensity for interaction with their intended target pockets.

4.1 Benchmarking Dataset

To benchmark the impact of a high proportion of viral protein families within the training set on molecule generation, we conducted a comparative analysis using the simplest model configuration: one trained solely on protein embeddings derived from ESM2, without any pocket-specific information. For this comparison, the ESM2 model trained on the 'Without Top 5 Virus' dataset was selected from epoch 10, exhibiting a loss of 0.18. Similarly, the model trained on the 'With Virus' dataset was chosen from epoch 8, also with a loss of 0.18. See model configuration in Table 4.1. Molecules were then generated from the in-domain set of each corresponding

dataset, and their novelty was assessed by performing a nearest-neighbor Tanimoto similarity analysis against their respective training sets.

Model	Final Dim.	# Parameters	Non-virus	Virus
ESM2	256	6,332,928	82,875	121,363

TABLE 4.1: Configuration and dataset sizes for the model trained with ESM2 protein only representation on 'Without Top 5 Virus' and 'With Virus' datasets.

Regarding the similarity results (Table 4.2), no substantial differences were observed between the models trained on the 'With Virus' and 'Without Top 5 Virus' datasets. Both models consistently generated molecules exhibiting high similarity to their respective training sets. This aspect will be investigated in greater detail in subsequent sections. Furthermore, both generated sets demonstrated high internal diversity, as evidenced by the all-to-all Tanimoto similarity analysis within each dataset.

Metric	Non-Virus Dataset	Virus Dataset
Novelty Assessment (vs. Training Set)		
Total Valid Molecules Analyzed	**37,114**	**38,265**
Tanimoto $<$ 0.5 (Count)	5.27% (1956)	4.78% (1830)
0.5 $\leq$ Tanimoto $<$ 1.0 (Count)	4.60% (1711)	4.43% (1696)
Tanimoto = 1.0 (Count)	90.12% (33447)	90.79% (34739)
Internal Diversity (All-to-All within Generated Set)		
Total Molecules for Diversity Analysis[a]	**3667**	**3526**
Tanimoto $<$ 0.5 (% of total pairs)	96.51%	99.12%
0.5 $\leq$ Tanimoto $<$ 1.0 (% of total pairs)	2.94%	0.81%
Tanimoto = 1.0 (% of total pairs)	0.56%	0.07%

TABLE 4.2: Comparison of Molecular Generation Performance for ESM2 Protein-Only Model on 'With Vrius' and 'Without Top 5 Virus' Datasets. [a]These numbers represent the subset of molecules (Tanimoto $\neq 1$ vs. training set) that were carried forward for diversity analysis.

The cross-dataset analysis (Table 4.3) further supported these observations, revealing a low similarity between molecules generated from the 'With Virus' and 'Without Top 5 Virus' datasets. This low cross-similarity indicates that the models are capable of generating molecules that are structurally distinct and specific to the types of protein targets it was trained on.

Metric	Value
Total Pairs Compared	**3667** $\times$ **3526**
Tanimoto $<$ 0.5 (% of total pairs)	98.54%
0.5 $\leq$ Tanimoto $<$ 1.0 (% of total pairs)	1.40%
Tanimoto = 1.0 (% of total pairs)	0.05%

TABLE 4.3: Cross-Dataset Diversity: Tanimoto Similarity between Generated Molecules from 'With Virus' and 'Without Top 5 Virus' (ESM2 Protein-Only Model).

This absence of significant bias is further corroborated by the family-level analysis of generated molecules, shown in Figures D.1 and D.2. These figures illustrate the distribution of validly generated molecules across protein families for each dataset. It can be observed that even though the 'With Virus' dataset inherently contained a higher proportion of viral families during training, the resulting proportion of valid generated molecules for those families was not disproportionate. This indicates that the model did not exhibit a bias towards highly represented families, maintaining a consistent proportion of valid generations across various protein families. This can alse been confirmed with the distribution of physicochemical properties, as all properties follow the train set distribution as seen in Figure D.3.

Based on our comparative analysis, we can conclude that the over-representation of viral families within the dataset does not introduce a significant bias in the generation of molecules. Consequently, to streamline the subsequent benchmarking analysis, we will proceed solely with the 'With Virus' dataset, as its inclusion does not appear to compromise the model's generalizability or the diversity of generated outputs.

4.2 Benchmarking Types of Input Representations

The configurations of the models employed for benchmarking the various input representations are detailed in Table 4.4. This table outlines their respective final embedding dimensions, total number of trainable parameters, the size of the final dataset used for training, and the epoch and validation loss of the specific checkpoint selected for molecule generation, which the training curves can be seen in Figure 3.5 b.

Model	Final Dim.	# Parameters	Dataset (Virus)	Epoch	Loss
ESM2 + PickPocket	512	16,851,968	120,556	7	0.17
SaProt + PickPocket	512	16,851,968	123,272	7	0.15
TensorDTIprot+ TensorDTIpocket	512	16,851,968	116,295	6	0.16

TABLE 4.4: Models configurations, final dataset sizes for the different input representations, and the selected checkpoint details.

4.2.1 Benchmarking Pocket Representation

Our first evaluation assessed whether the inclusion of pocket information substantially contributes to the model's generative capacity. As presented in Table 4.5, a cross-comparison was performed between molecules generated by the ESM2 protein-only model and the ESM2 + PickPocket model. This all-to-all Tanimoto similarity analysis revealed that the vast majority of generated molecules were distinct between the two models. This high degree of dissimilarity indicates that incorporating pocket information provides novel contextual cues to the generative model, leading to the production of diverse molecular sets when comparing models with and without this additional input. When oberving the distribution of the physicochemical properties (Fig. D.4) between each other and the training dataset we can observe that they share similar distributions.

Metric	ESM2 Protein-Only vs. ESM2 + PickPocket
Total Pairs Compared	**3526 × 4504**
Tanimoto < 0.5 (% of total pairs)	99.42%
$0.5 \leq$ Tanimoto < 1.0 (% of total pairs)	0.53%
Tanimoto = 1.0 (% of total pairs)	0.05%

TABLE 4.5: Cross-Model Diversity: Impact of Adding Pocket Information (ESM2 Protein-Only vs. ESM2 + PickPocket).

4.2.2 Benchmarking Types of Protein Representations

Deciding that adding pocket information can contribute greatly in the diversity of the generated molecules, for the benchmarking of types of input representations we have combined all with pocket information. As presented in Table 4.6, a nearest-neighbor Tanimoto similarity analysis against the corresponding training set revealed a substantial overlap with known molecules. Specifically, for all evaluated input representations, a high percentage of generated molecules (ranging from 88.70% to 91.07%) were exact matches (Tanimoto = 1.0) to compounds in the training set.

Subsequently, focusing on the subset of molecules that were not exact matches to the training set (i.e., those with Tanimoto similarity less than 1.0), an all-to-all Tanimoto similarity analysis was performed within each generated set to assess internal diversity. Across all models, we observed a consistently high level of diversity, with approximately 99% of the generated molecular pairs exhibiting Tanimoto similarity less than 0.5.

Metric	ESM2 + PickPocket	SaProt + PickPocket	TensorDTI Prot+Pocket
Novelty Assessment (vs. Training Set)			
Total Valid Molecules Analyzed	**39,846**	**39,789**	**39,533**
T < 0.5 (Count)	6.47%	6.22%	4.79%
$0.5 \leq$ T < 1.0 (Count)	4.84%	4.28%	4.14%
T = 1.0 (Count)	88.70%	89.49%	91.07%
Internal Diversity (All-to-All within Generated Set)			
Total Mols. for Diversity Analysis[a]	**4504**	**4180**	**3532**
T$<$ 0.5 (% of total pairs)	99.47%	99.58%	99.32%
$0.5 \leq$ T < 1.0 (% of total pairs)	0.45%	0.36%	0.62%
T = 1.0 (% of total pairs)	0.08%	0.06%	0.05%

TABLE 4.6: Comparison of Molecular Generation Performance for Different Input Representations (Virus Dataset). T is abbreviation for Tanimoto.[a]These numbers represent the subset of molecules (Tanimoto $\neq 1$ vs. training set) that were carried forward for diversity analysis.

Following the internal diversity analysis, we further investigated the extent to which molecules generated by different input representation models overlap, as shown in Table 4.7. This cross-model diversity assessment revealed that the sets of molecules generated by each distinct input modality (ESM2 + PickPocket, SaProt + PickPocket, and TensorDTI Prot+Pocket) are highly distinct from one another.

Specifically, pairwise Tanimoto similarity comparisons between the generated sets consistently showed an high percentage of molecules with Tanimoto similarity less than 0.5, ranging from 99.48% to 99.59%. On the other hand, the proportion of molecules exhibiting moderate similarity ($0.5 \leq$ Tanimoto < 1.0) or exact matches (Tanimoto $= 1.0$) between different model outputs was exceedingly low, consistently below 1%.

This significant dissimilarity across generated sets implies that the different input representations guide the generative model to explore and produce molecules from distinct regions of the chemical space. This finding is highly advantageous, as it suggests that leveraging diverse protein and pocket embedding modalities allows for the generation of a broader and more varied pool of potential drug candidates.

Metric	**ESM2 + PickPocket vs. SaProt + PickPocket**	**ESM2 + PickPocket vs. TensorDTI Prot+Pocket**	**SaProt + PickPocket vs. TensorDTI Prot+Pocket**
Total Pairs Compared	**4504 × 4180**	**4504 × 3532**	**4180 × 3532**
T $<$ 0.5 (% of total pairs)	99.59%	99.48%	99.55%
$0.5 \leq$ T $<$ 1.0 (% of total pairs)	0.36%	0.48%	0.42%
T = 1.0 (% of total pairs)	0.05%	0.04%	0.03%

TABLE 4.7: Cross-Model Diversity: Tanimoto Similarity Between Generated Molecule Sets from Different Input Representation Models. T is abbreviation for Tanimoto.

After filtering for valid molecules that were not exact matches to the training set, we subjected the selected candidates to Tensor-DTI inference to predict their interaction with the target proteins. Table 4.8 summarizes these results for different input representation models. We evaluated predicted interactions using three distinct Tensor-DTI models, each pre-trained on a different dataset: SMPBind, PLINDER, and BindingDB.

	ESM2-PickPocket	**SaProt-PickPocket**	**TensorDTI**
SMPBind	478	458	393
Plinder	405	342	314
BindingDB	953	794	681
All Three Positive	19	23	15
Any Positive	1529	1341	1152
Total Rows	4504	4180	3532
Plinder % of Total	8.99%	8.18%	8.89%
Any Positive % of Total	33.95%	32.08%	32.62%

TABLE 4.8: Comparison of DTI predictions across three different protein/pocket representations - **in-domain** test set

Across all input representation combinations (ESM2-PickPocket, SaProt-PickPocket, and TensorDTI), the results showed a consistent pattern in the number of molecules predicted to interact with the targets. The table presents the raw counts of molecules predicted as positive by each individual Tensor-DTI model, as well as the count for molecules predicted as positive by "All Three" models and "Any Positive" by at

least one. The percentages indicate the proportion of positively predicted molecules relative to the "Total Rows" (which represents the initial set of unique and valid molecules passed to Tensor-DTI). Overall, all input representation models yielded similar rates of predicted binders, suggesting robust performance regardless of the specific protein/pocket embedding used. Despite producing similar numbers of predicted binders, each model generates chemically distinct molecules, highlighting their complementary exploration of chemical space.

Upon computing the distribution of physicochemical properties (Fig. D.5) for the molecules generated by the three different input representation models, we observed highly consistent profiles. This indicates that despite the structural distinctness among molecules generated by different models (as previously noted in our cross-model diversity analyses), they collectively exhibit desirable drug-like properties. These consistent distributions suggest that the generative process effectively guides the creation of compounds within a chemically relevant and druggable space. Notice also in Table 4.9 the SA score for all models is below 4 also indicating that the generated drugs for all the models seem easy to synthesize.

Metric	**ESM2 + PickPocket**	**SaProt + PickPocket**	**TensorDTI Prot+Pocket**
Validity Proportion	0.797 ± 0.016	0.796 ± 0.015	0.791 ± 0.036
Synthetic Accessibility Score	3.538 ± 1.037	3.458 ± 1.060	3.737 ± 0.985

TABLE 4.9: Validity (Mean ± SD) and Synthetic Accessibility Scores (Mean ± SD) for Molecules Generated by Different Input Representation Models (In-Domain).

4.3 Out-Domain Generation

We used the out-of-domain test set to determine if our models showed a bias toward the training dataset during molecule generation. Table E.2 reveals that the models continued to frequently generate molecules that were exact matches to compounds already present in the training set.

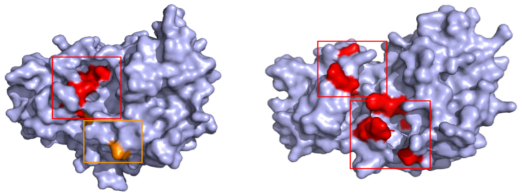
Crucially, when we compared the performance of molecules generated for out-of-domain targets with those from in-domain targets, we observed no major differences in their overall behavior. This includes their novelty and diversity profiles, suggesting a consistent generative capability. While the models consistently produce molecules seen in the training set, this behavior is not disproportionately worse for unseen targets, indicating they do not significantly overfit to the specific proteins from the training data. This consistency is further reflected in their validation scores, Synthetic Accessibility (SA) scores (Table E.4), and predicted TensorDTI performance (Table E.5), as well as the physicochemical distributions (Fig. E.1) all of which showed similar trends between in-domain and out-of-domain sets.

4.4 Assessing binding pocket specificity

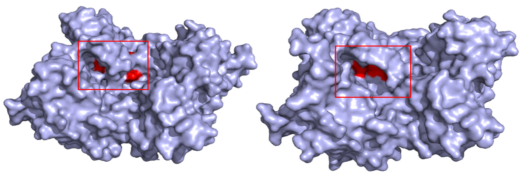
Beyond the quantitative benchmarking, we performed retrospective case studies to qualitatively assess the model's ability to generate structurally distinct molecules

specifically conditioned on different binding pockets, and to evaluate their predicted binding efficacy. For this analysis, we selected two well-characterized protein targets, CDK2 and RET, each with experimentally resolved structures featuring both canonical active sites and distinct cryptic or allosteric pockets.

- **CDK2:** A key cell cycle regulator and a prominent cancer target. We included the active site pocket from `3fwq` and the cryptic pocket from `5cu3` (Fig. 4.1a). The presence of both active and cryptic sites in CDK2 offers a robust test of the model's capacity to generate molecules for varying binding modes.
- **RET:** A receptor tyrosine kinase implicated in various cancers. This case study featured the active site pocket from `2ivs` and a cryptic site from `7ju5` (Fig. 4.1b). Similar to CDK2, RET allowed us to investigate the model's ability to differentiate between conventional orthosteric and less common allosteric/cryptic binding site chemistries.



(A) *Left.* CDK2 with ATP binding site (red) and closed cryptic site (orange). *Right.* Open cryptic cavity merging with ATP binding site (red).



(B) *Left.* RET with cryptic binding site in apo state (red). *Right.* Open cryptic cavity (red).

FIGURE 4.1: Structural arrangements from CDK2 and RET kinases in holo and apo states for their correspondent cryptic pockets.

The objective for these retrospective cases was to determine if, when conditioned on these distinct pocket types (active vs. cryptic), the model would generate chemically diverse molecules that are appropriate for their respective binding sites. Furthermore, we aimed to ascertain if the generated candidates showed a high predicted likelihood of interacting with the specific pocket they were conditioned on. This qualitative analysis provides insights into the model's precision in targeted molecule generation.

Metric	**ESM2 + PickPocket**	**SaProt + PickPocket**	**TensorDTI Prot+Pocket**
Novelty Assessment (vs. Training Set)			
Total Valid Molecules Analyzed	**74**	**113**	**549**
Tanimoto < 0.5 (Count)	90.54% (67)	24.78% (28)	15.66% (86)
$0.5 \leq$ Tanimoto < 1.0 (Count)	6.76% (5)	12.39% (14)	11.84% (65)
Tanimoto = 1.0 (Count)	2.70% (2)	62.83% (71)	72.50% (398)

TABLE 4.10: Novelty Assessment of Generated Molecules for **CDK2** (vs. Training Set).

Our analysis of 1000 generated molecules for a specific pocket reveals that the TensorDTI protein+Pocket model consistently generated more valid molecules exceeding the Tanimoto similarity thresholds for both target types. This is evident

Metric	ESM2 + PickPocket	SaProt + PickPocket	TensorDTI Prot+Pocket
Novelty Assessment (vs. Training Set)			
Total Valid Molecules Analyzed	**66**	**123**	**538**
Tanimoto < 0.5 (Count)	87.88% (58)	30.89% (38)	15.24% (82)
$0.5 \leq$ Tanimoto < 1.0 (Count)	10.61% (7)	13.82% (17)	16.54% (89)
Tanimoto = 1.0 (Count)	1.52% (1)	55.28% (68)	68.22% (367)

TABLE 4.11: Novelty Assessment of Generated Molecules for **RET** (vs. Training Set).

in Tables 4.10 and 4.11. Despite the differences in molecule generation, all models showed comparable performance when predicting the interaction of the *de novo* molecules against their respective targets.

When we examined the interaction of our generated molecules with CDK2 binding sites, only a small fraction (out of 1000) were predicted to bind. Interestingly, our models produced a similar number of interacting molecules for both the cryptic and active binding pockets (see Table 4.13). However, the molecules generated for the active site showed low Tanimoto similarity to those generated for the cryptic site (Table 4.12), suggesting they are distinct.

To investigate the pocket specificity of our generated molecules further, we conducted an additional Drug-Target Interaction prediction. This time, we swapped the pocket labels (i.e molecules generated for the cryptic site were paired with the active pocket, and vice versa). Our aim was to determine if the molecules were truly specific to the pocket they were designed for.

The results, presented in Table 4.14, showed that the DTI model predicted roughly the same number of interacting molecules across all three generation models, even with the swapped labels. This indicates that while our models can generate molecules that bind, they do not necessarily exhibit pocket-specific interactions. This lack of specificity might stem from the model's tendency to learn from highly binding motifs, despite efforts to avoid replicating exact training set molecules. Our PAINS analysis of the training set (3.1.2) revealed a significant presence of highly promiscuous drugs, which likely influenced the model to generate unique molecules that bind strongly to pockets in a general sense, rather than providing the desired pocket-specific binding. We leave the removal of PAINS-related interactions from the dataset and the subsequent generation of molecules under these conditions for future work.

Regarding the physicochemical property distributions, although we do not observe identical distributions to the training set, primarily due to differences in the number of interactions, the values of the generated and accepted molecules remain within ranges typically associated with drug-like compounds. See Figures F.2 and F.1 for the active and cryptic sites of CDK2, and Figures F.4 and F.3 for both binding sites of RET.

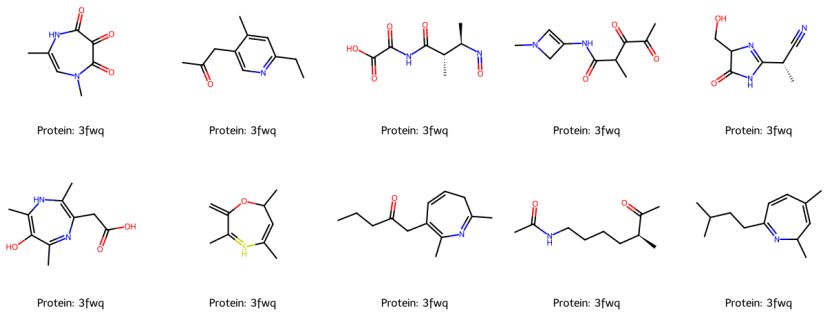
When performing the same types of analysis for RET target, we observe the same behaviours concluding the same, see F for DTI and physicochemical property analysis.

These figures present grid visualizations of generated SMILES compounds, identified by the ESM2+PickPocket model with high predicted drug-target interaction

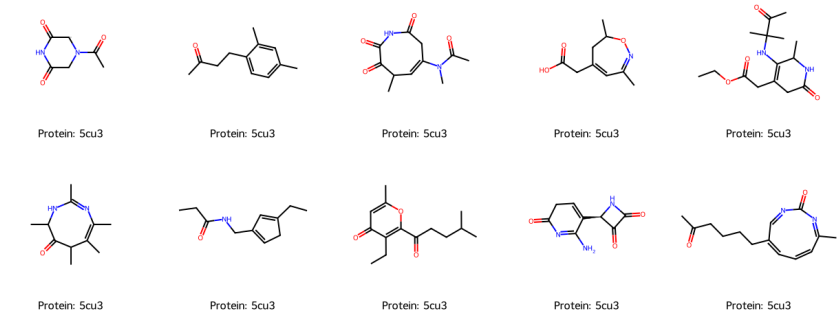
Tanimoto Similarity	ESM2 + PickPocket	SaProt + PickPocket	TensorDTI Prot + Pocket
CDK2 (3FW *vs* 5CU3)			
Tanimoto < 0.5	100.00%	92.00%	95.34%
$0.5 \leq$ Tanimoto < 1	0.00%	3.29%	4.53%
Tanimoto == 1	0.00%	4.71%	0.12%
RET (2IVS *vs* 7JU5)			
Tanimoto < 0.5	100.00%	96.55%	93.53%
$0.5 \leq$ Tanimoto < 1	0.00%	1.46%	6.38%
Tanimoto == 1	0.00%	1.99%	0.08%

TABLE 4.12: Similarity of Molecules Between Cryptic and Active Pockets.

scores for specific CDK2 and RET protein pockets (Figures F.5 and F.6).



(A) CDK2 Active Pockets



(B) CDK2 Cryptic Pockets

FIGURE 4.2: Top confident binding generated molecules for CDK2 protein, categorized by pocket type.

Pocket Type	Metric	ESM2-PickPocket	SaProt-PickPocket	TensorDTI
Cryptic Site (5cu3)	SMPBind	0	0	1
	Plinder	4	3	7
	BindingDB	13	6	17
	All Three Positive	0	0	0
	Any Positive	15	7	18
	Total Rows	34	17	67
	Plinder % of Total	11.76%	17.65%	10.45%
	Any Positive % of Total	44.12%	41.18%	26.87%
Active Site (3fwq)	SMPBind	0	0	0
	Plinder	2	3	5
	BindingDB	11	3	23
	All Three Positive	0	0	0
	Any Positive	11	4	25
	Total Rows	38	25	84
	Plinder % of Total	5.26%	12.00%	5.95%
	Any Positive % of Total	28.95%	16.00%	29.76%

TABLE 4.13: Interaction predictions for **CDK2** with correct (cryptic 5cu3 vs. active 3fwq) pocket assignment.

Pocket Type	Metric	ESM2-PickPocket	SaProt-PickPocket	TensorDTI
Active Site (5cu3)	SMPBind	0	0	0
	Plinder	4	4	6
	BindingDB	11	3	26
	All Three Positive	0	0	0
	Any Positive	12	5	26
	Total Rows	38	25	84
	Plinder % of Total	10.53%	16.00%	7.14%
	Any Positive % of Total	31.58%	20.00%	30.95%
Cryptic Site (3fwq)	SMPBind	0	0	0
	Plinder	3	2	5
	BindingDB	9	6	14
	All Three Positive	0	0	0
	Any Positive	11	6	15
	Total Rows	34	17	67
	Plinder % of Total	8.82%	11.76%	7.46%
	Any Positive % of Total	32.35%	35.29%	22.39%

TABLE 4.14: Interaction predictions for **CDK2** after swapping pocket labels (active 5cu3 vs. cryptic 3fwq).

Chapter 5

Conclusions and Future Work

This thesis successfully developed and evaluated a novel pocket-aware generative model designed for de novo molecule generation from pre-trained protein and pocket embeddings. Our findings provide crucial insights into the capabilities and limitations of such models in targeted drug discovery.

Firstly, our analysis revealed that the presence of a disproportionate amount of viral proteins in the training dataset does not introduce a significant bias in the molecule generation process. This suggests the model's ability to generalize across different protein families, which is vital for broad applicability in drug design.

However, a notable limitation observed across all evaluated models is their propensity to generate molecules identical to compounds within their training sets. This high fidelity to known structures, evidenced by significant Tanimoto = 1.0 similarities, indicates a potential issue of overfitting, thereby limiting the exploration of truly novel chemical space. Addressing this will be critical in future work, potentially through strategies like selecting earlier, less overfitted checkpoints, or implementing advanced generative techniques focused on promoting greater molecular diversity and novelty.

Despite this tendency towards known structures, within the subset of valid molecules that are dissimilar to the training set, our models demonstrated the ability to generate diverse compounds, as confirmed by all-to-all Tanimoto similarity analyses. Furthermore, these generated molecules consistently exhibited favorable physicochemical properties within acceptable ranges, suggesting their strong potential for practical application.

Yet, a deeper examination highlighted a challenge: while our model is sensitive to pocket information, it does not consistently produce truly pocket-specific molecules, as demonstrated in the two kinase case studies. This limitation stems primarily from the characteristics of the training dataset itself, specifically the high number of promiscuous molecules, as indicated by PAINS analysis. The model, learning from these promiscuous motifs, might inadvertently generate molecules with broad binding potential rather than highly selective pocket-specific binders.

Once these challenges related to novelty and pocket specificity are overcome, future work could naturally extend to more rigorous validation of the generated molecules. This could involve computational methods such as molecular docking to predict the precise binding pose and affinity of the generated SMILES with their target proteins. Ultimately, the most definitive step could be experimental assays to confirm the binding and functional modulation of these computationally designed molecules in a laboratory setting, thereby completing the drug discovery cycle.

Appendix A

Dataset

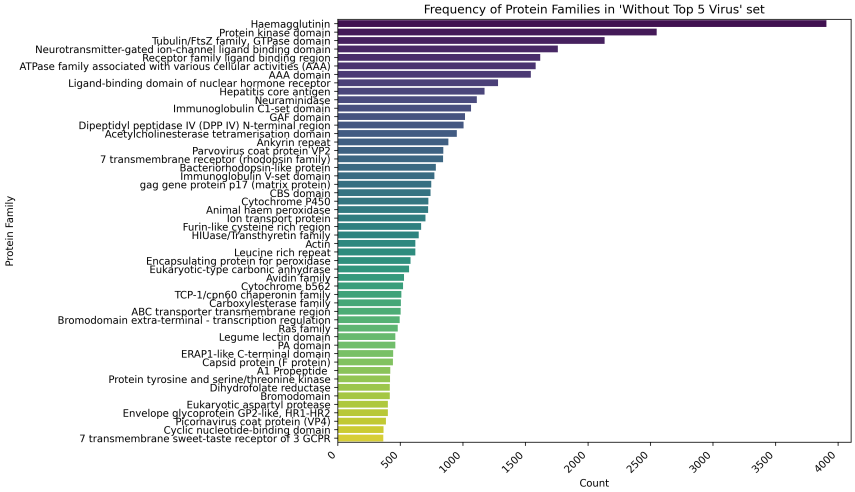


FIGURE A.1: Proportion of top 50 most frequent families in 'Without Top 5 Virus' set (In-domain)

Appendix B

Models Configurations

Table B.1 provides a comprehensive overview of the configurations for all models trained and used for molecule generation in this study. In addition to concatenation, we also explored weighted addition as a method for combining protein and pocket embeddings.

For weighted addition, both protein and pocket embeddings were resized to match dimensions (256). They were then combined using a weighted sum, expressed as $\alpha \times \text{src\_prot} + \beta \times \text{src\_pocket}$, where α and β are user-defined weights. This approach allowed the model to modulate the relative influence of global protein context versus localized pocket information. However, no significant difference in the performance of the generated molecules was observed when using weighted addition compared to concatenation. Consequently, models trained with weighted addition are not included in the main text or in further detailed analysis.

Model	Alpha/ Beta	Final Dim.	# Params	Non-virus	Virus	Fusion Type	Notes
PickPocket	0-1	256	6,332,928	84,806	123,272	-	Pocket
ESM2	1-0	256	6,332,928	82,875	121,363	-	Protein
SaProt	1-0	256	6,332,928	85,164	124,104	-	Protein
ESM2 + PickPocket	1-2	256	6,332,928	82,532	120,556	Weighted Add	Fusion
SaProt + PickPocket	1-2	256	6,332,928	84,806	123,272	Weighted Add	Fusion
ESM2 + PickPocket	-	512	16,851,968	82,532	120,556	Concat	Fusion
SaProt + PickPocket	-	512	16,851,968	84,806	123,272	Concat	Fusion
TensorDTI	0-1	256	6,332,928	79,085	116,295	-	Pocket
TensorDTI	1-0	256	6,332,928	78,073	117,860	-	Protein
TensorDTIprot + TensorDTIpocket	1-2	256	6,332,928	78,073	116,295	Weighted Add	Fusion
TensorDTIprot + TensorDTIpocket	-	512	16,851,968	78,073	116,295	Concat	Fusion

TABLE B.1: Configurations and dataset sizes for models trained on viral and non-viral DTI tasks.

Appendix C

Post-generation Analysis of Molecules

C.1 Drug Similarity Metrics

One of the most widely used and effective molecular similarity metrics is the Tanimoto similarity coefficient (Bajusz, Rácz, and Héberger, 2015), particularly when applied to molecular fingerprints. Molecular fingerprints (Muegge and Mukherjee, 2016) are bit string representations of molecules, where each bit corresponds to the presence or absence of a specific substructural feature or chemical property.

The Tanimoto coefficient (also known as the Jaccard index for binary data) quantifies the similarity between two sets of features. When applied to binary molecular fingerprints, it is calculated as the ratio of the number of common bits set in both fingerprints to the total number of bits set in either fingerprint. Mathematically, for two fingerprints A and B, the Tanimoto coefficient is given by:

$$Tanimoto(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{N_c}{N_A + N_B - N_c}$$

Where:

- N_c is the number of bits set in both fingerprint A and fingerprint B (common features).
- N_A is the number of bits set in fingerprint A.
- N_B is the number of bits set in fingerprint B.

A Tanimoto similarity value ranges from 0 to 1, where 0 indicates no common features (completely dissimilar) and 1 indicates identical fingerprints (identical molecules in terms of represented features). A higher Tanimoto score signifies greater structural and/or substructural similarity between the two molecules. In drug discovery, a Tanimoto similarity of 0.7 or higher between two compounds based on common fingerprints is often considered indicative of high structural similarity, frequently implying similar biological activity.

The application of Tanimoto similarity extends to evaluating the novelty and diversity of generated molecule sets. By comparing newly generated compounds against known active molecules or existing chemical libraries, researchers can assess

whether the generative model is merely reproducing known structures or exploring genuinely novel chemical space while maintaining structural resemblance to desired archetypes. This balance between novelty and retaining beneficial features is paramount for effective *de novo* drug design.

C.2 Physicochemical Properties

Beyond simply generating novel molecular structures, it is crucial that these generated compounds possess characteristics that make them "drug-like." This concept of drug-likeness refers to a set of physicochemical and structural properties commonly observed in successful drugs. Adhering to these properties increases the likelihood that a molecule will exhibit favorable pharmacokinetics, that is, how the body handles a drug in terms of absorption, distribution, metabolism, excretion and toxicity (ADMET), thereby improving its chances of progressing through the drug discovery pipeline.

The computation of these key physicochemical properties from a molecule's structure is essential for the rapid in-silico evaluation of generated compounds. This is frequently accomplished using cheminformatics toolkits, such as RDKit Bento et al., 2020. RDKit is an open-source cheminformatics software package widely used in drug discovery and computational chemistry for manipulating chemical structures, generating descriptors, and performing various cheminformatics tasks. It provides robust functionalities to parse molecular representations like SMILES strings and efficiently compute a broad array of molecular properties.

In order to analyse our generated molecules, we have assessed some selected properties:

- **Molecular Weight (MolWt):** This indicates the overall size of a molecule. Most oral drugs fall within a specific molecular weight range (e.g., generally less than 500 Daltons, as per Lipinski's Rule of Five), as excessively large molecules often struggle with absorption.
- **Topological Polar Surface Area (TPSA):** TPSA measures the sum of surfaces of all polar atoms (oxygen, nitrogen, and attached hydrogens). It's a good indicator of a molecule's ability to permeate cell membranes and is correlated with drug absorption and brain penetration.
- **Hydrogen Bond Donors (NumHDonors) and Acceptors (NumHAcceptors):** These counts reflect a molecule's capacity to form hydrogen bonds, which are vital for interacting with biological targets and influencing solubility and membrane permeability.
- **Number of Rotatable Bonds (NumRotatableBonds):** This indicates molecular flexibility. While some flexibility is beneficial for target binding (e.g., induced fit), excessive flexibility can lead to promiscuous binding and make a molecule difficult to optimize.
- **Number of Aromatic Rings (NumAromaticRings):** Aromaticity is common in drug molecules, influencing their stability, rigidity, and interaction with aromatic residues in proteins.

- **Heavy Atom Count (HeavyAtomCount):** The number of all non-hydrogen atoms in the molecule, providing another measure of molecular size and complexity.
- **Murcko Scaffold:** Beyond individual properties, analyzing the Murcko scaffold provides insight into the fundamental core structure of a molecule. This scaffold is the common ring and linker framework of a molecule, stripping away side chains. Generating molecules with desirable or novel scaffolds is often a goal in drug design, as scaffolds largely dictate the overall shape and binding potential. By examining the scaffold of generated molecules, researchers can assess their structural diversity and novelty relative to known drugs.
- **Synthetic Accessibility Score (SAScore):** A computational metric used to estimate how easy or difficult it is to synthesize a molecule. Lower scores generally indicate easier synthesizability, which is a critical factor for the practical development of a drug candidate (Ertl and Schuffenhauer, 2009).

Appendix D

Results on Different Benchmarkings

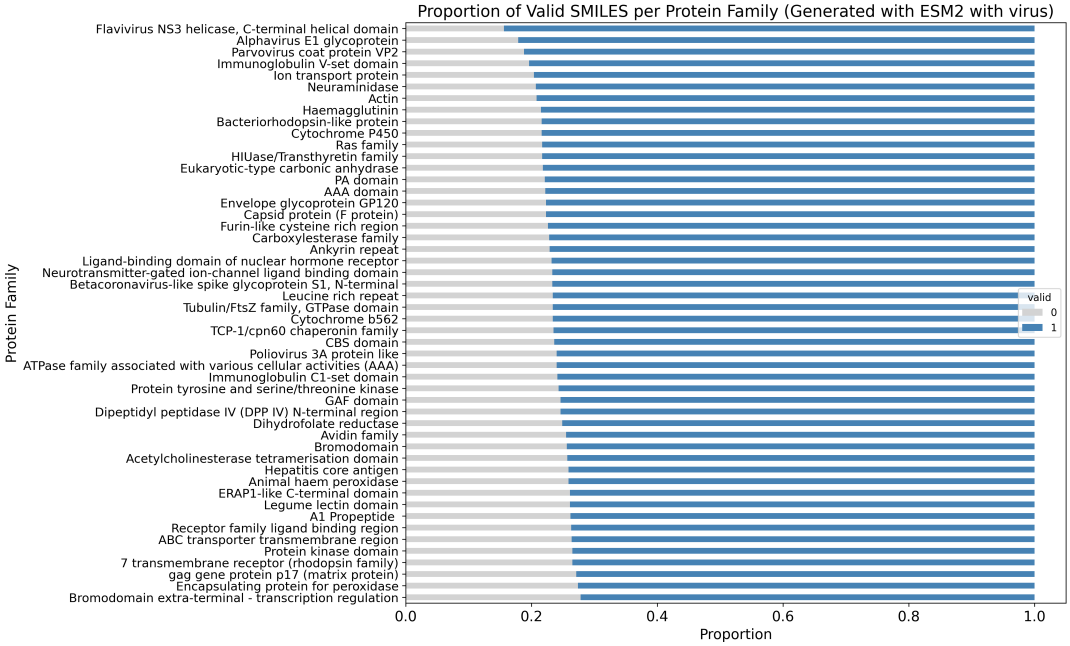


FIGURE D.1: Validity Proportion of molecules generated from In-domain 'With Virus' set

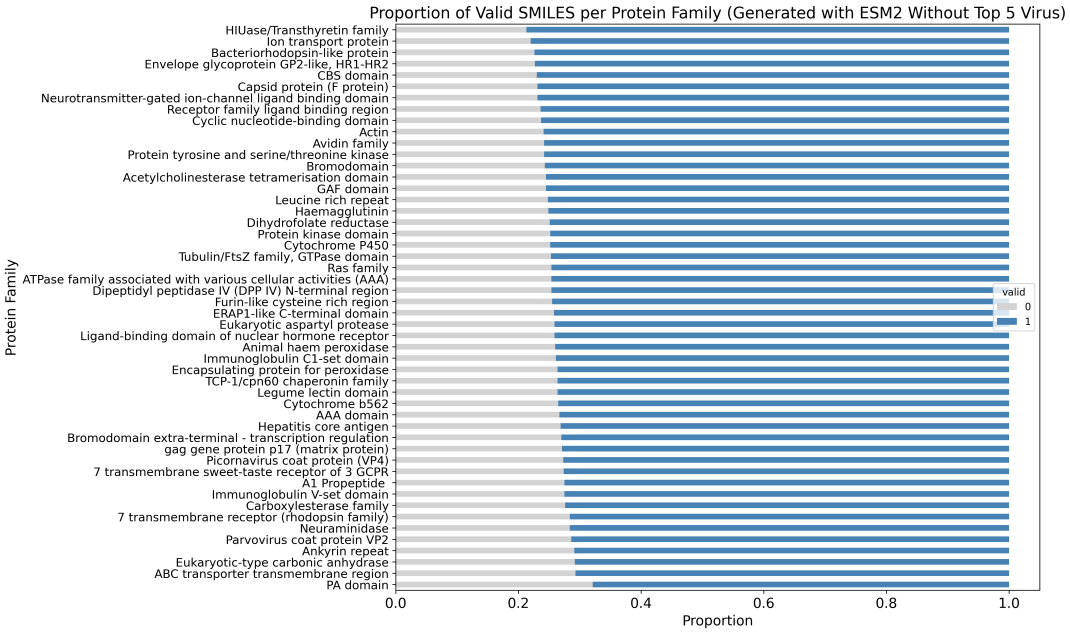


FIGURE D.2: Validity Proportion of molecules generated from In-domain 'Without top 5 Virus' set

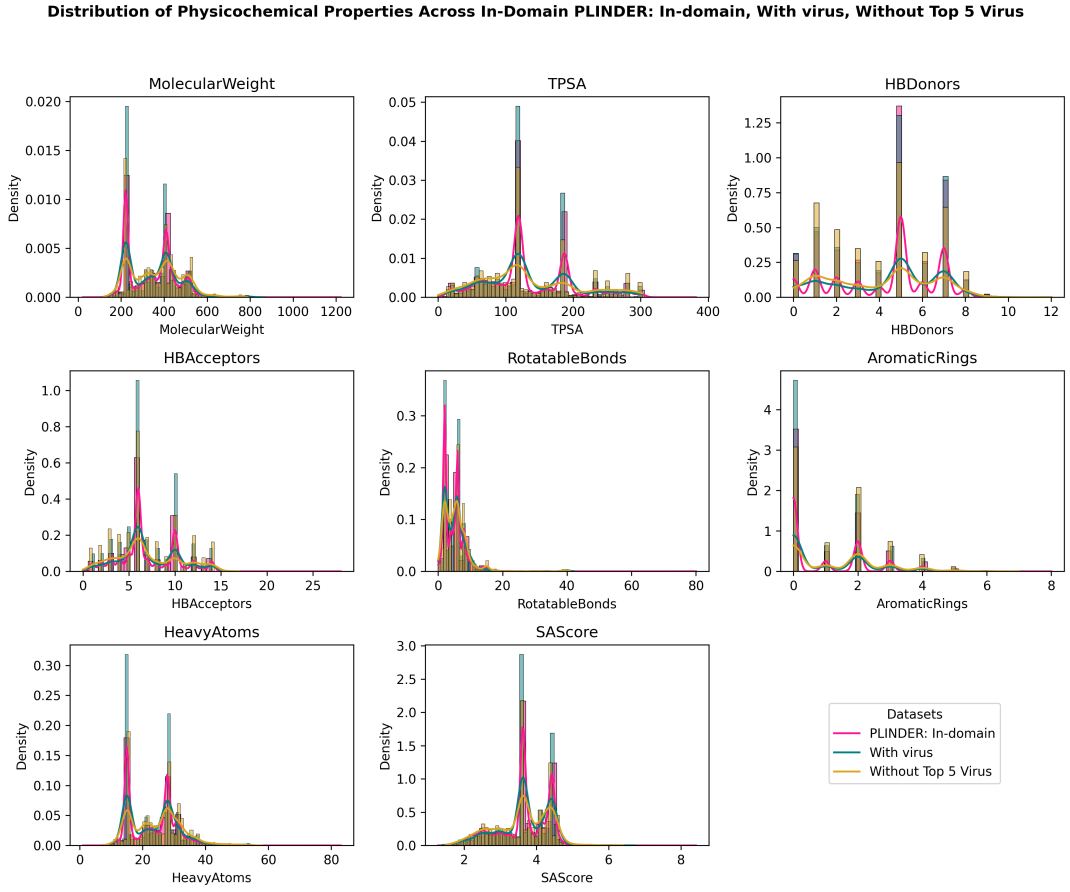


FIGURE D.3: Physicochemical Properties distribution of molecules generated from In-domain With Virus set and In-domain Without Top 5 Virus set compared to original train+val PLINDER split.

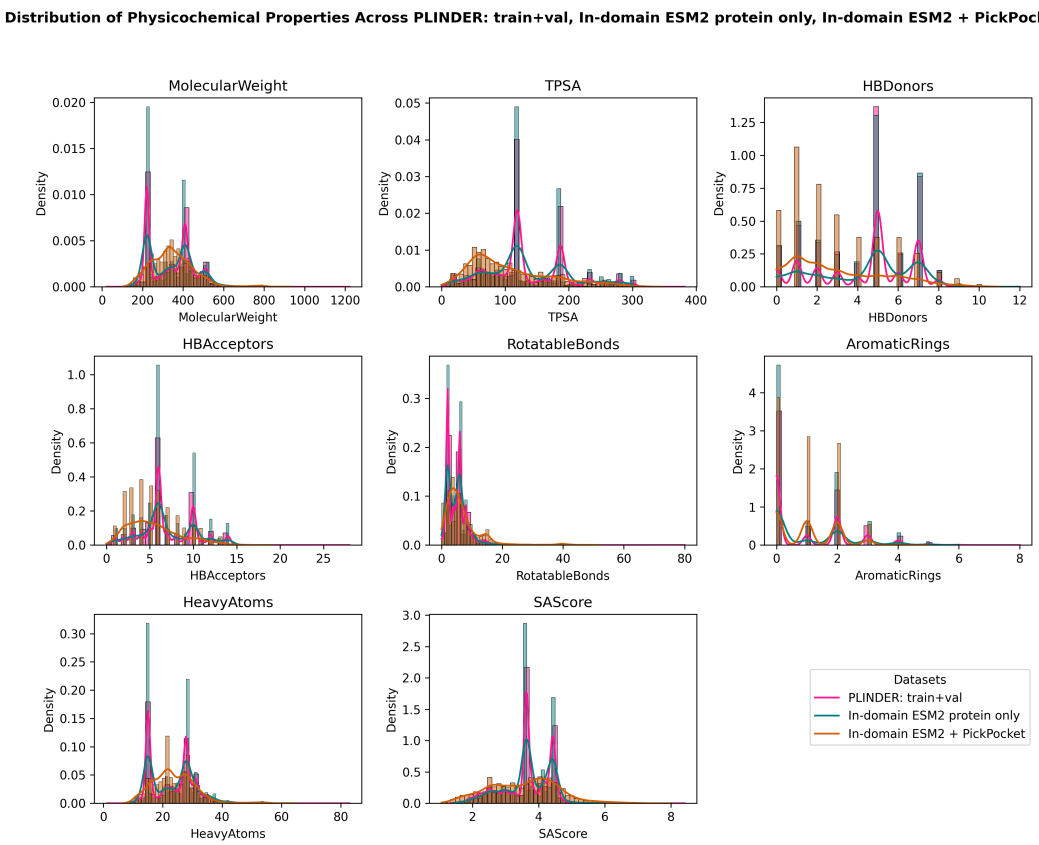


FIGURE D.4: Physicochemical Properties distribution of molecules generated from In-domain With Virus set from models ESM2 protein only and ESM2 + PickPocket compared to original train+val PLINDER split.

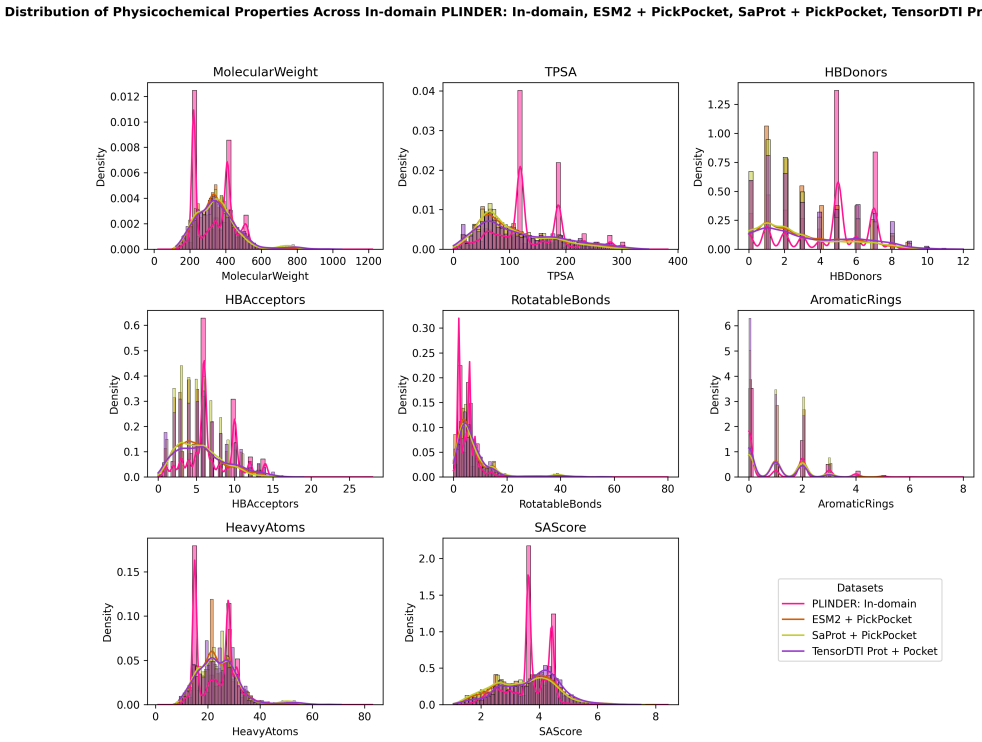


FIGURE D.5: Physicochemical Properties distribution of molecules generated from In-domain With Virus set from models ESM2 + PickPocket, SaProt+PickPocket and TensorDTI Prot+Pocket compared to original train+val PLINDER split.

Appendix E

Out-of-Domain Dataset

Metric	ESM2 Protein-Only vs. ESM2 + PickPocket
Tanimoto Similarity Between Generated Molecule Sets	
Total Pairs Compared	**3545 × 4187**
Tanimoto < 0.5 (% of total pairs)	99.35%
$0.5 \leq$ Tanimoto < 1.0 (% of total pairs)	0.61%
Tanimoto = 1.0 (% of total pairs)	0.04%

TABLE E.1: Cross-Model Diversity (Out-of-Domain): Impact of Adding Pocket Information (ESM2 Protein-Only vs. ESM2 + PickPocket).

Metric	ESM2 + PickPocket	SaProt + PickPocket	TensorDTI Prot+Pocket
Novelty Assessment (vs. Training Set)			
Total Valid Molecules Analyzed	**39,276**	**39,564**	**38,897**
Tanimoto < 0.5 (Count)	5.73% (2249)	6.52% (2581)	5.59% (2174)
$0.5 \leq$ Tanimoto < 1.0 (Count)	4.93% (1938)	4.67% (1846)	4.48% (1741)
Tanimoto = 1.0 (Count)	89.34% (35089)	88.81% (35137)	89.93% (34982)
Internal Diversity (All-to-All within Generated Set)			
Total Molecules for Diversity Analysis[a]	**4187**	**4427**	**3915**
Tanimoto < 0.5 (% of total pairs)	99.31%	99.58%	99.41%
$0.5 \leq$ Tanimoto < 1.0 (% of total pairs)	0.64%	0.35%	0.55%
Tanimoto = 1.0 (% of total pairs)	0.05%	0.08%	0.04%

TABLE E.2: Comparison of Molecular Generation Performance for Different Input Representations (Out-of-Domain Dataset).[a]These numbers represent the subset of molecules (Tanimoto < 1.0 vs. training set) that were carried forward for diversity analysis.

Metric	**ESM2 + PickPocket vs. SaProt + PickPocket**	**ESM2 + PickPocket vs. TensorDTI Prot+Pocket**	**SaProt + PickPocket vs. TensorDTI Prot+Pocket**
Tanimoto Similarity Between Generated Molecule Sets			
Total Pairs Compared	**4187 × 4427**	**4187 × 3915**	**4427 × 3915**
Tanimoto < 0.5 (% of total pairs)	99.53%	99.43%	99.60%
$0.5 \leq$ Tanimoto < 1.0 (% of total pairs)	0.43%	0.54%	0.38%
Tanimoto = 1.0 (% of total pairs)	0.04%	0.03%	0.03%

TABLE E.3: Cross-Model Diversity (Out-of-Domain): Tanimoto Similarity Between Generated Molecule Sets from Different Input Representation Models.

Metric	**ESM2 + PickPocket**	**SaProt + PickPocket**	**TensorDTI Prot+Pocket**
Validity Proportion	0.786 ± 0.017	0.791 ± 0.017	0.778 ± 0.035
Synthetic Accessibility Score	3.733 ± 1.012	3.439 ± 1.058	3.786 ± 0.999

TABLE E.4: Validity (Mean ± SD) and Synthetic Accessibility Scores (Mean ± SD) for Molecules Generated by Different Input Representation Models (Out-of-Domain).

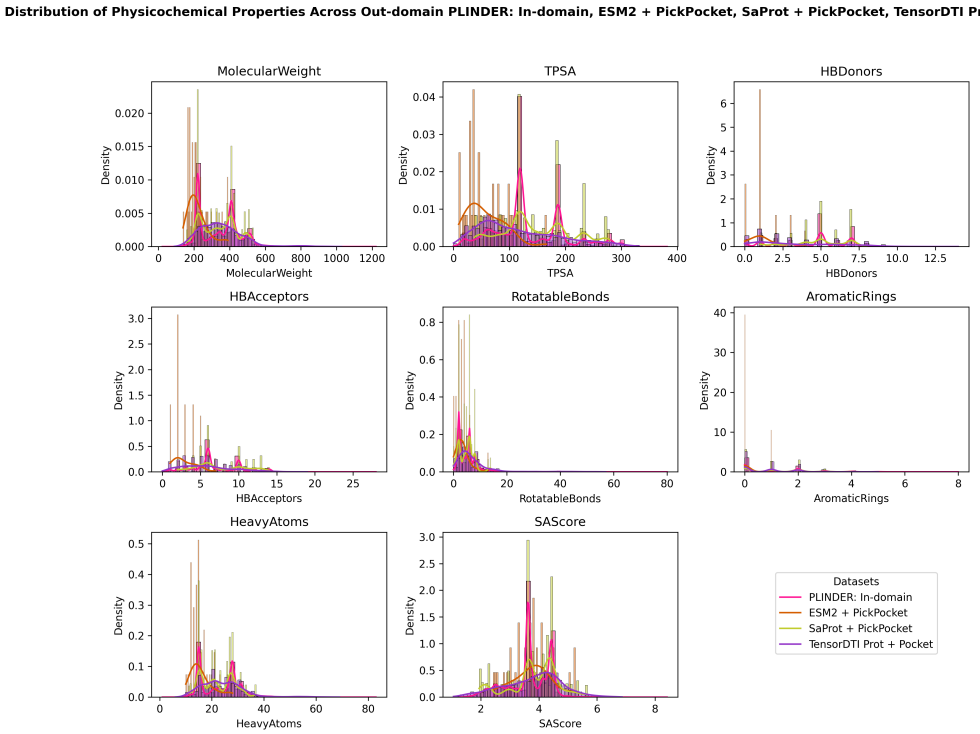


FIGURE E.1: Physicochemical Properties distribution of molecules generated from Out-domain With Virus set from models ESM2 + PickPocket, SaProt+PickPocket and TensorDTI Prot+Pocket compared to original train+val PLINDER split.

	ESM2-PickPocket	**SaProt-PickPocket**	**TensorDTI**
SMPBind	314	284	365
Plinder	543	459	482
BindingDB	728	730	735
All Three Positive	9	14	6
Any Positive	1403	1293	1381
Total Rows	4187	4427	3915
Plinder % of Total	12.97%	10.37%	12.31%
Any Positive % of Total	33.51%	29.21%	35.27%

TABLE E.5: Comparison of DTI predictions across three different protein/pocket representations - **Out-of-Domain** test set

Appendix F

Retrospective Extra Figures

In this section we can find the Tables and Figures regarding the same analysis done with the retrospective case of CDK2 but with RET target. The Physicochemical property distributions are also found here.

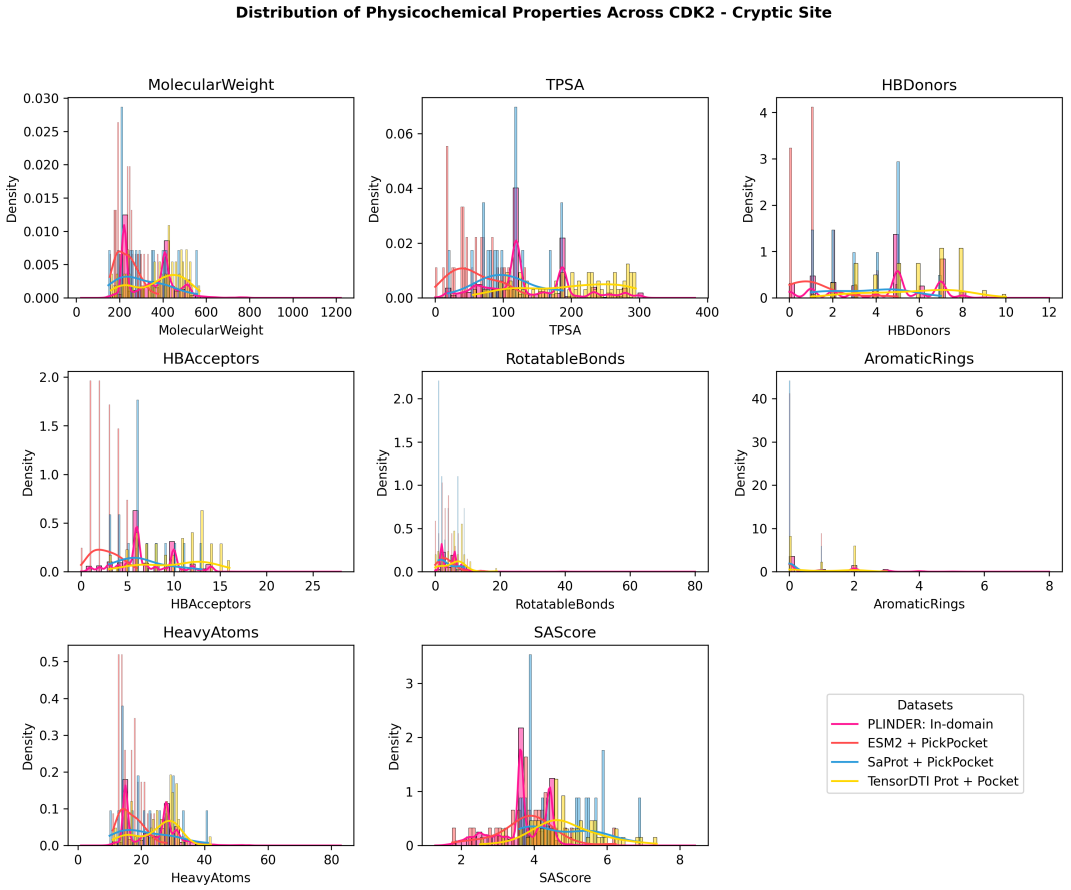


FIGURE F.1: Physicochemical properties distribution of generated and accepted molecules targeting the **cryptic site** of CDK2. Although absolute distributions differ from the training set due to fewer interactions, values remain within drug-like ranges.

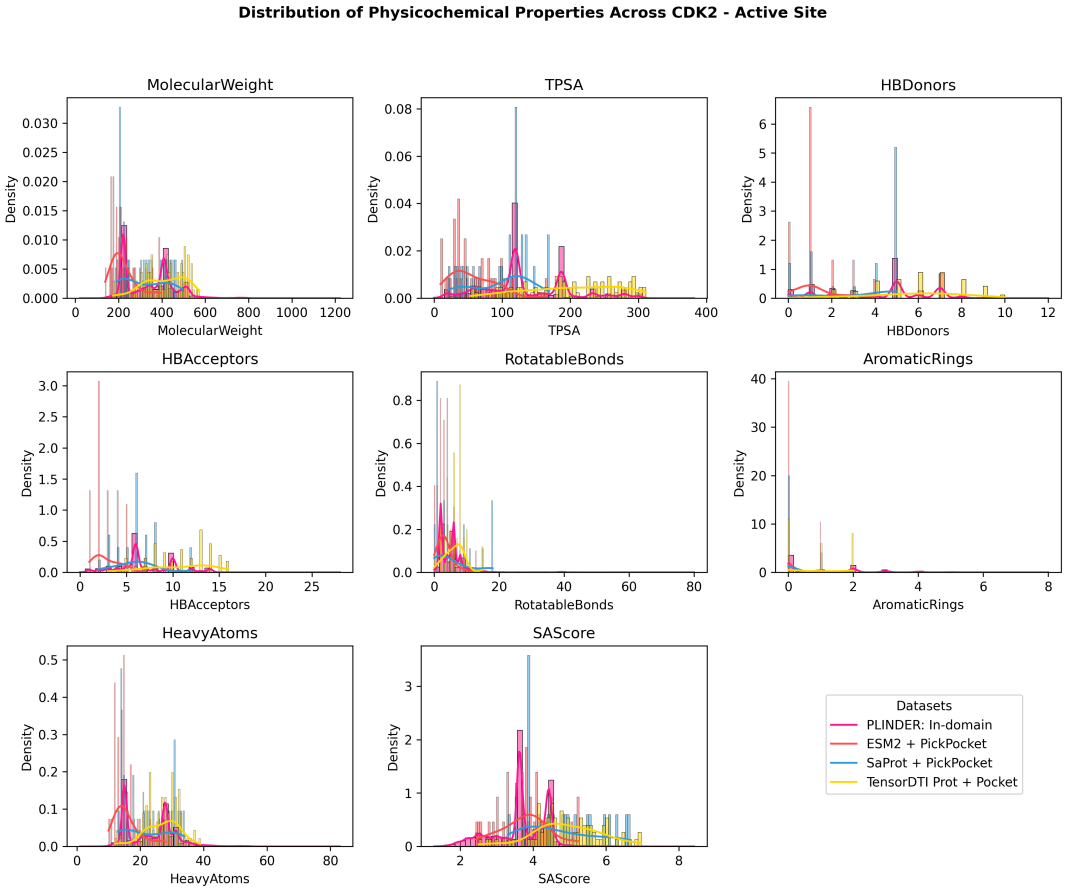


FIGURE F.2: Physicochemical properties distribution of generated and accepted molecules targeting the **active site** of CDK2. Distributions align with druggable molecule profiles.

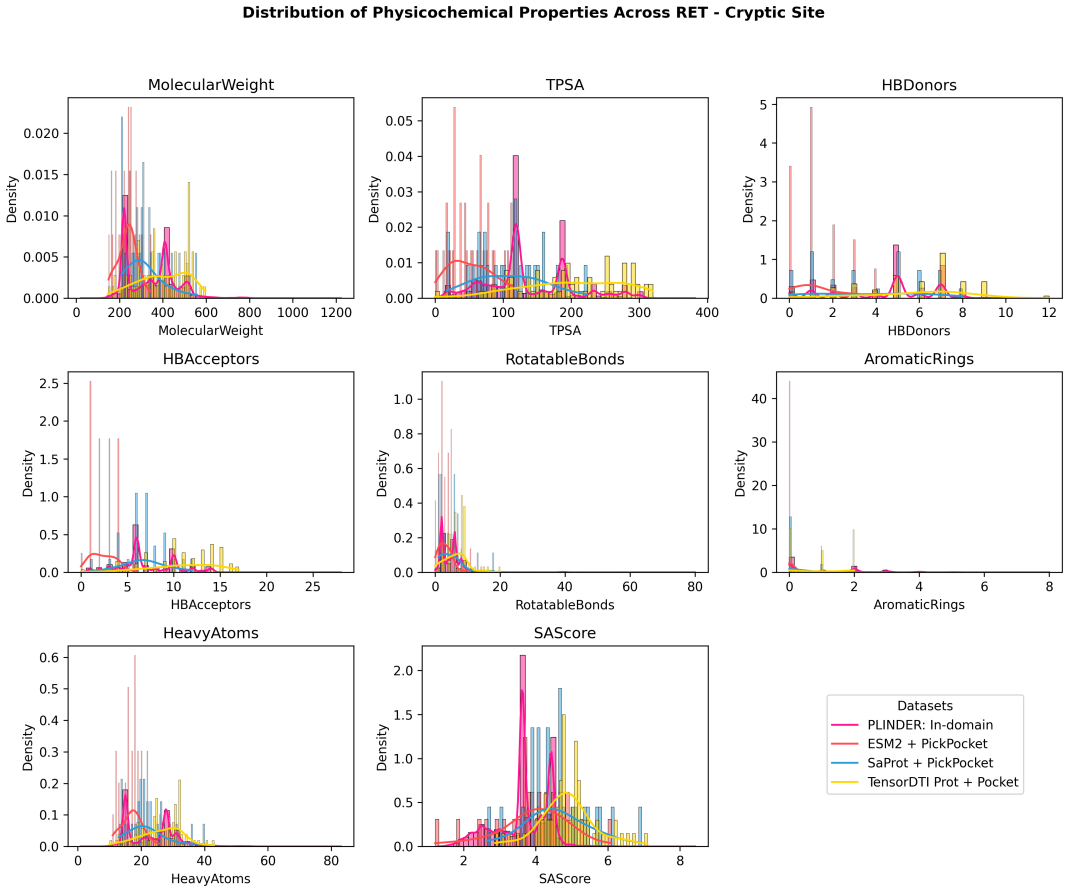


FIGURE F.3: Physicochemical properties distribution of generated molecules for the **cryptic site** of RET. Despite fewer counts, generated molecules fall within expected drug-like ranges.

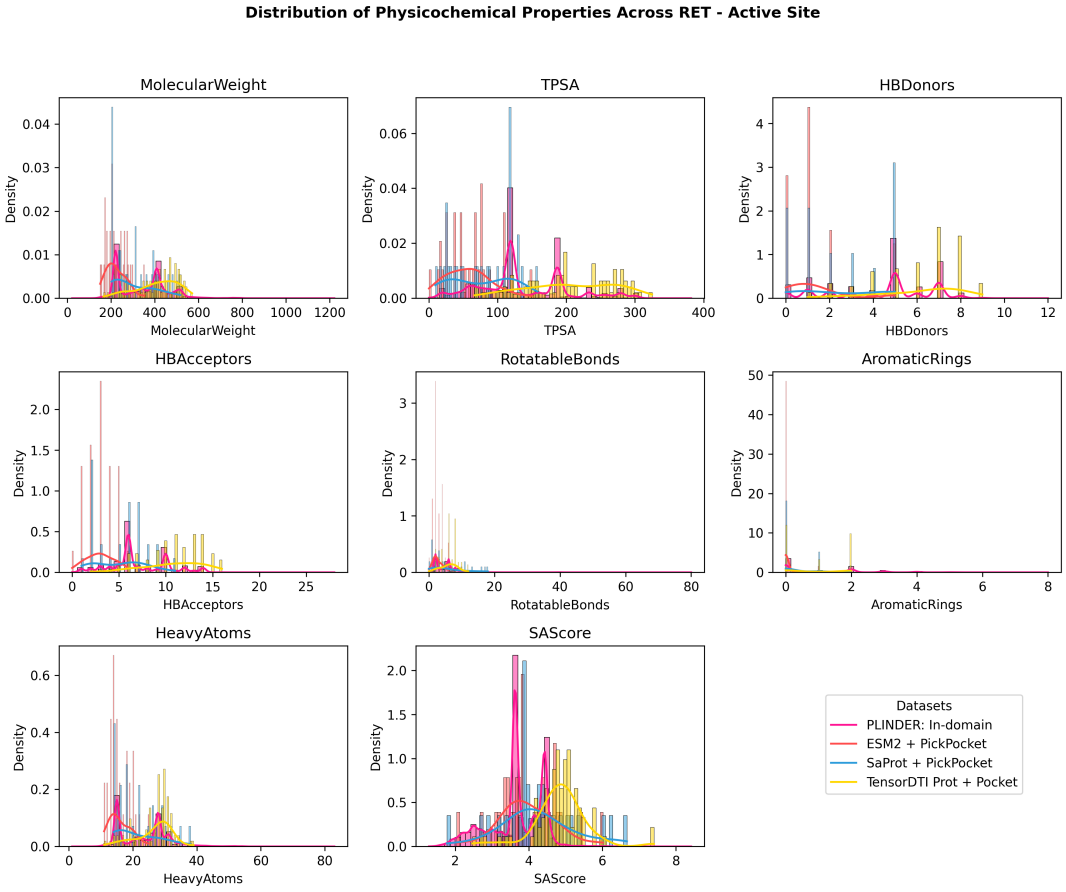


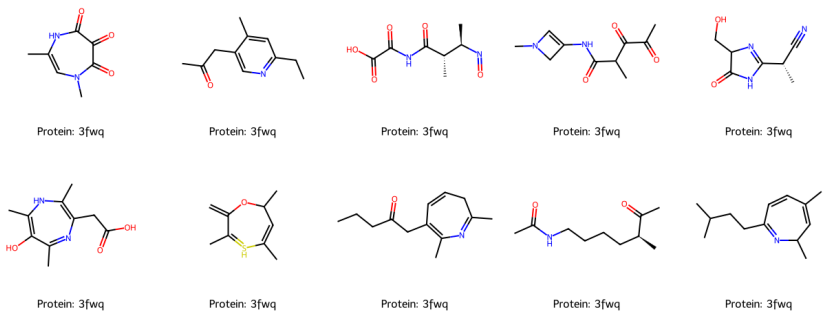
FIGURE F.4: Physicochemical properties distribution of generated molecules for the **active site** of RET. Similarity to training distribution reflects alignment with known drug-like space.

Pocket Type	Metric	ESM2-PickPocket	SaProt-PickPocket	TensorDTI
Cryptic Site (2IVS)	SMPBind	0	0	0
	Plinder	0	2	6
	BindingDB	26	22	58
	All Three Positive	0	0	0
	Any Positive	26	22	59
	Total Rows	33	26	79
	Plinder % of Total	0.00%	7.69%	7.59%
	Any Positive % of Total	78.79%	84.62%	74.68%
Active Site (7JU5)	SMPBind	0	0	0
	Plinder	0	1	2
	BindingDB	23	19	58
	All Three Positive	0	0	0
	Any Positive	23	19	58
	Total Rows	32	29	92
	Plinder % of Total	0.00%	3.45%	2.17%
	Any Positive % of Total	71.88%	65.52%	63.04%

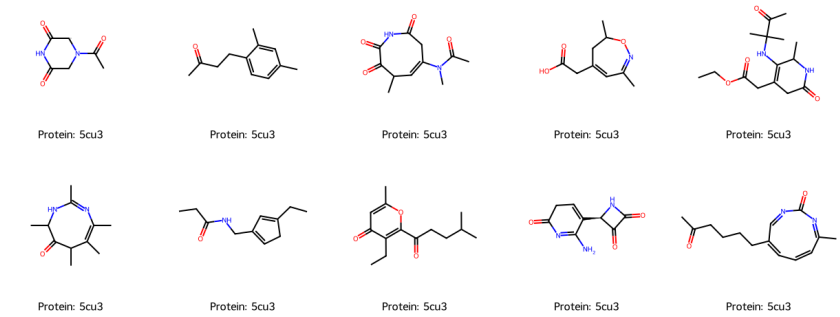
TABLE F.1: Interaction predictions for **RET** with correct (cryptic 2IVS vs. active 7JU5) pocket assignment.

Pocket Type	Metric	ESM2-PickPocket	SaProt-PickPocket	TensorDTI
Active Site (2IVS)	SMPBind	0	0	0
	Plinder	0	1	2
	BindingDB	25	22	53
	All Three Positive	0	0	0
	Any Positive	25	22	53
	Total Rows	32	29	92
	Plinder % of Total	0.00%	3.45%	2.17%
	Any Positive % of Total	78.13%	75.86%	57.61%
Cryptic Site (7JU5)	SMPBind	0	0	0
	Plinder	0	3	7
	BindingDB	26	19	64
	All Three Positive	0	0	0
	Any Positive	26	19	65
	Total Rows	33	26	79
	Plinder % of Total	0.00%	11.54%	8.86%
	Any Positive % of Total	78.79%	73.08%	82.28%

TABLE F.2: Interaction predictions for **RET** after swapping pocket labels (active 2IVS vs. cryptic 7JU5).

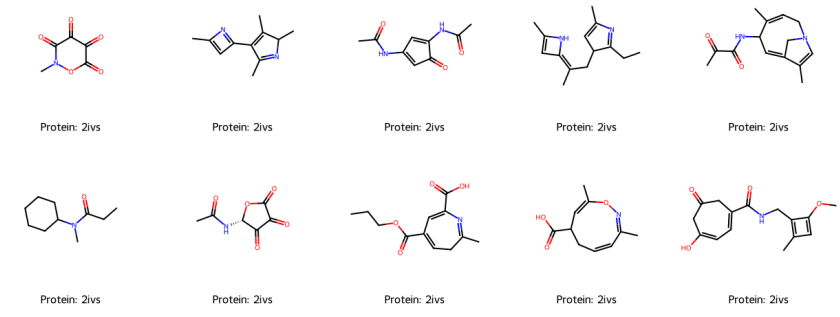


(A) CDK2 Active Pockets

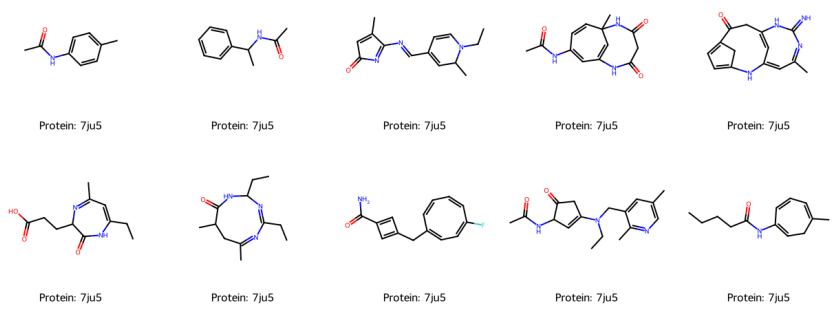


(B) CDK2 Cryptic Pockets

FIGURE F.5: Top confident binding generated molecules for CDK2 protein, categorized by pocket type.



(A) RET Active Pockets



(B) RET Cryptic Pockets

FIGURE F.6: Top confident binding generated molecules for RET protein, categorized by pocket type.

Bibliography

Alakhdar, Amira, Barnabas Poczos, and Newell Washburn (2024). "Diffusion models in de novo drug design". In: *Journal of Chemical Information and Modeling* 64.19, pp. 7238–7256.

Bajusz, Dávid, Anita Rácz, and Károly Héberger (2015). "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?" In: *Journal of cheminformatics* 7, pp. 1–13.

Bento, A Patrícia et al. (2020). "An open source chemical structure curation pipeline using RDKit". In: *Journal of Cheminformatics* 12, pp. 1–16.

Berman, Helen M et al. (2000). "The protein data bank". In: *Nucleic acids research* 28.1, pp. 235–242.

Bjerrum, Esben Jannik and Richard Threlfall (2017). "Molecular generation with recurrent neural networks (RNNs)". In: *arXiv preprint arXiv:1705.04612*.

Chakraborty, Sabyasachi, Prakriti Kayastha, and Raghunathan Ramakrishnan (2019). "The chemical space of B, N-substituted polycyclic aromatic hydrocarbons: Combinatorial enumeration and high-throughput first-principles modeling". In: *The Journal of chemical physics* 150.11.

Chen, Yangyang et al. (2023). "Deep generative model for drug design from protein target sequence". In: *Journal of Cheminformatics* 15.1, p. 38.

Creanza, Teresa Maria et al. (2025). "Transformer Decoder Learns from a Pretrained Protein Language Model to Generate Ligands with High Affinity". In: *Journal of Chemical Information and Modeling*.

David, Laurianne et al. (2020). "Molecular representations in AI-driven drug discovery: a review and practical guide". In: *Journal of cheminformatics* 12.1, p. 56.

Durairaj, Janani et al. (2024). "PLINDER: The protein-ligand interactions dataset and evaluation resource". In: *bioRxiv*, pp. 2024–07.

Elnaggar, Ahmed et al. (2021). "Prottrans: Toward understanding the language of life through self-supervised learning". In: *IEEE transactions on pattern analysis and machine intelligence* 44.10, pp. 7112–7127.

Ertl, Peter and Ansgar Schuffenhauer (2009). "Estimation of synthetic accessibility score of drug-like molecules based on molecular complexity and fragment contributions". In: *Journal of cheminformatics* 1, pp. 1–11.

Ferruz, Noelia, Steffen Schmidt, and Birte Höcker (2022). "ProtGPT2 is a deep unsupervised language model for protein design". In: *Nature communications* 13.1, p. 4348.

Gil-Sorribes, Manel, Álvaro Ciudad Serrano, and Alexis Molina (2025). "TensorDTI: Enhancing Biomolecular Interaction Prediction with Contrastive Embedding Learning". In: *ICLR 2025 Workshop on Learning Meaningful Representations in Life Sciences (LMRL)*.

Gilson, Michael K et al. (2016). "BindingDB in 2015: a public database for medicinal chemistry, computational chemistry and systems pharmacology". In: *Nucleic acids research* 44.D1, pp. D1045–D1053.

Greives, Nicholas and Huan-Xiang Zhou (2014). "Both protein dynamics and ligand concentration can shift the binding mechanism between conformational selection and induced fit". In: *Proceedings of the National Academy of Sciences* 111.28, pp. 10197–10202.

Hughes, James P et al. (2011). "Principles of early drug discovery". In: *British journal of pharmacology* 162.6, pp. 1239–1249.

Jumper, John et al. (2021). "Highly accurate protein structure prediction with AlphaFold". In: *nature* 596.7873, pp. 583–589.

Kant, Shashi, Saheli Roy, et al. (2025). "Artificial intelligence in drug discovery and development: transforming challenges into opportunities". In: *Discover Pharmaceutical Sciences* 1.1, pp. 1–14.

Li, Qingxin and CongBao Kang (2020). "Mechanisms of action for small molecules revealed by structural biology in drug discovery". In: *International journal of molecular sciences* 21.15, p. 5262.

Lin, Eugene, Chieh-Hsin Lin, and Hsien-Yuan Lane (2020). "Relevant applications of generative adversarial networks in drug design and discovery: molecular de novo design, dimensionality reduction, and de novo peptide and protein design". In: *Molecules* 25.14, p. 3250.

Lin, Zeming et al. (2022). "Language models of protein sequences at the scale of evolution enable accurate structure prediction". In: *bioRxiv*.

Lin, Zeming et al. (2023). "Evolutionary-scale prediction of atomic-level protein structure with a language model". In: *Science* 379.6637, pp. 1123–1130.

Lipinski, Christopher A et al. (1997). "Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings". In: *Advanced drug delivery reviews* 23.1-3, pp. 3–25.

Liu, Qi et al. (2018). "Constrained graph variational autoencoders for molecule design". In: *Advances in neural information processing systems* 31.

Lopez, Michael J and Shamim S Mohiuddin (2024). "Biochemistry, essential amino acids". In: *StatPearls [Internet]*. StatPearls Publishing.

Luong, Kha-Dinh and Ambuj Singh (2024). "Application of transformers in cheminformatics". In: *Journal of Chemical Information and Modeling* 64.11, pp. 4392–4409.

Mak, Kit-Kay, Yi-Hang Wong, and Mallikarjuna Rao Pichika (2024). "Artificial intelligence in drug discovery and development". In: *Drug discovery and evaluation: safety and pharmacokinetic assays*, pp. 1461–1498.

Miller, Mitchell D and George N Phillips (2021). "Moving beyond static snapshots: Protein dynamics and the Protein Data Bank". In: *Journal of Biological Chemistry* 296.

Mouchlis, Varnavas D et al. (2021). "Advances in de novo drug design: from conventional to machine learning methods". In: *International journal of molecular sciences* 22.4, p. 1676.

Mswahili, Medard Edmund and Young-Seob Jeong (2024). "Transformer-based models for chemical SMILES representation: A comprehensive literature review". In: *Heliyon*.

Muegge, Ingo and Prasenjit Mukherjee (2016). "An overview of molecular fingerprint similarity search in virtual screening". In: *Expert opinion on drug discovery* 11.2, pp. 137–148.

Rasul, Hezha O et al. (2024). "Decoding Drug Discovery: Exploring A-to-Z In Silico Methods for Beginners". In: *Applied Biochemistry and Biotechnology*, pp. 1–51.

Sadybekov, Anastasiia V and Vsevolod Katritch (2023). "Computational approaches streamlining drug discovery". In: *Nature* 616.7958, pp. 673–685.

Singh, Rohit et al. (2023). "Contrastive learning in protein language space predicts interactions between drugs and protein targets". In: *Proceedings of the National Academy of Sciences* 120.24, e2220778120.

Stevenson, Garrett A et al. (2023). "Clustering protein binding pockets and identifying potential drug interactions: a novel ligand-based featurization method". In: *Journal of Chemical Information and Modeling* 63.21, pp. 6655–6666.

Su, Jin et al. (2023). "Saprot: Protein language modeling with structure-aware vocabulary". In: *bioRxiv*, pp. 2023–10.

Tarasi, Stelina, Laura Malo, and Alexis Molina (2025). "PickPocket Enables Binding Site Prediction at the Proteome Scale". In: *ICLR 2025 Workshop on Generative and Experimental Perspectives for Biomolecular Design*.

Thangudu, Ratna Rajesh et al. (2012). "Modulating protein–protein interactions with small molecules: the importance of binding hotspots". In: *Journal of molecular biology* 415.2, pp. 443–453.

"UniProt: the Universal protein knowledgebase in 2025" (2025). In: *Nucleic Acids Research* 53.D1, pp. D609–D617.

Van Kempen, Michel et al. (2024). "Fast and accurate protein structure search with Foldseek". In: *Nature biotechnology* 42.2, pp. 243–246.

Weininger, David (1988). "SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules". In: *Journal of chemical information and computer sciences* 28.1, pp. 31–36.

Wu, Lirong et al. (2022). "A survey on protein representation learning: Retrospect and prospect". In: *arXiv preprint arXiv:2301.00813*.

Yi, Jia-Cai et al. (2024). "ChemMORT: an automatic ADMET optimization platform using deep learning and multi-objective particle swarm optimization". In: *Briefings in Bioinformatics* 25.2, bbae008.

Zdrazil, Barbara et al. (2024). "The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods". In: *Nucleic acids research* 52.D1, pp. D1180–D1192.

Zhang, Zuobai et al. (2022). "Protein representation learning by geometric structure pretraining". In: *arXiv preprint arXiv:2203.06125*.

Zhang, Zuobai et al. (2023). "A Systematic Study of Joint Representation Learning on Protein Sequences and Structures". In: *arXiv preprint arXiv:2303.06275*.

Zhou, Jie et al. (2020). "Graph neural networks: A review of methods and applications". In: *AI open* 1, pp. 57–81.